UNIVERSITÉ PIERRE ET MARIE CURIE

École Doctorale 386
Sciences mathématiques de Paris Centre

Laboratoire de Statistique Théorique et Appliquée

Thèse de doctorat

en vue de l'obtention du grade de
Docteur en Sciences de l'Université Pierre et Marie Curie

Discipline : Mathématiques

Spécialité : Statistiques

présentée par

Patricia CONDE CESPEDES

Modélisations et extensions du formalisme de l'Analyse Relationnelle Mathématique à la modularisation des grands graphes

---

dirigée par
M. Jean-François MARCOTORCHINO

Soutenue le 18 décembre 2013 devant le jury composé de :

M. Paul DEHEUVELS	Professeur, Université Paris VI	Président
M. Jean-Loup GUILLAUME	MdC, Université Paris VI	Examinateur
M. Jean-François MARCOTORCHINO	Directeur Scientifique, Thales	Directeur
M. Gilbert SAPORTA	Professeur, CNAM	Examinateur
M. Michalis VAZIRGIANNIS	Professeur, École Polytechnique	Examinateur
M. Emmanuel VIENNET	Professeur, Université Paris XIII	Rapporteur

Rapporteurs :

M. Renaud LAMBIOTTE	Professeur, Université de Namur
M. Emmanuel VIENNET	Professeur, Université Paris XIII

Laboratoire de Statistique Théorique et Appliquée
4, place Jussieu
75 252 Paris

École doctorale Paris centre Case 386
4 place Jussieu
75 252 Paris cedex 05

Remerciements

Tout ce qui est écrit dans cette section n'est qu'une traduction de mes sentiments vis-à-vis des sensations et de mes expériences vécues ces trois dernières années.

Je tiens tout d'abord à exprimer ma profonde gratitude à mon directeur de thèse, le Professeur Jean-François Marcotorchino, que j'admire et respecte. Je le remercie d'abord de m'avoir donné l'opportunité d'élaborer un travail de recherche sur un domaine d'actualité et très passionnant. J'ai apprécié particulièrement ses remarques très constructives, nos échanges et le temps consacré à la lecture de mes rapports. Je le remercie également de m'avoir appris que la recherche est une passion, beaucoup plus qu'un métier.

Je tiens à remercier également le Professeur Paul Deheuvels pour avoir accepté d'être membre de mon jury et pour m'avoir permis d'élaborer ce travail au sein du Laboratoire de Statistique Théorique et Appliquée (LSTA).

Je remercie les Professeurs Emmanuel Viennet et Renaud Lambiotte pour avoir eu la gentillesse d'accepter d'être rapporteurs de ma thèse. Leurs remarques, très constructives et enrichissantes, m'ont permis d'améliorer mon travail et m'ont donné des pistes pour de futurs travaux.

Je voudrais remercier aussi les Professeurs Gilbert Saporta, Michalis Vazirgiannis et Jean-Loup Guillaume de me faire l'honneur de faire partie de mon jury de thèse et pour l'intérêt porté à mon travail.

J'adresse un remerciement tout particulier à Romain et Jean-Loup du LIP6 pour m'avoir prêté main forte pour la partie pratique de mon travail. Cela a eu une influence importante sur les résultats finaux de mes travaux. Je me sens également redevable envers eux pour la lecture du dernier chapitre de ma thèse et pour leurs remarques très pertinentes qui ont permis d'améliorer la rédaction de mon travail.

Je remercie aussi le Professeur Gérard Biau, directeur du LSTA, pour tous ses conseils et orientations au début de ma thèse. Je remercie également le Professeur Michel Broniatowski pour m'avoir assisté vers l'étape finale de ce long chemin et pour l'intérêt porté à mon travail. Je profite aussi pour remercier tous les Professeurs et Maîtres de Conférences du LSTA que j'ai pu côtoyer pendant ces trois dernières années. Je m'adresse aussi à Louise et Corinne pour leur gentillesse et aide quotidienne.

J'adresse aussi mes sincères remerciements à mes collègues et amis du LSTA. Je remercie particulièrement mes camarades de salle Svetlana, Soumeya, Alexis, Salim, Petan et

Mamadou qui ont rendu ce séjour de trois ans très agréable avec des échanges et des rires au quotidien. Plus particulièrement, je remercie Svetlana et Alexis pour l'aide précieuse et opportune apportée à l'avancement de mon travail ainsi que Salim, Tabéa et Fanny pour leurs bons conseils. Je remercie également l'ensemble des doctorants et collègues du LSTA que j'ai eu la chance de rencontrer pendant les pauses café et déjeuners : Assia, Abdoullah, Amadou, Baptiste, Benjamin, Boris, Cécile, Emmanuel, Layal, Matthieu, Moïse, Mokhtar, Sarah, Tarn et Zhansheng. J'adresse également mes remerciements à Claire David pour toute la confiance qu'elle m'a accordée dans le cadre de ma mission enseignement.

El ultimo párrafo de esta etapa de "agradecimientos" le corresponde sin duda alguna a mi familia. Especialmente a mis papitos René y Dora, que durante toda mi vida fueron incondicionales y confiaron en mi, estuvieron siempre ahí para ayudarme a vencer todas las etapas necesarias para llegar hasta aquí. Como poder en unas cuantas líneas expresar mi gratificación a dos seres a quienes debo todo lo que soy ? Imposible ! Entonces no hay nada más que decir. Y no podía faltar mi hermano Iván por su colaboración desinteresada y confianza y por supuesto mi hermana Karen, ejemplo del esfuerzo, la perseverancia y la originalidad.

Enfin, je remercie les autres membres de ma famille, mes amis en dehors du cadre universitaire et j'espère que les gens que j'ai côtoyés mais que je n'ai pas mentionnés me pardonneront. Je voudrais clore cette section qui marque le début d'une aventure...

Patricia

Résumé

Un graphe étant un ensemble d'objets liés par une certaine relation typée, le problème de "modularisation" des grands graphes (qui revient à leur partitionnement en classes) peut, alors, être modélisé mathématiquement en utilisant l'Analyse Relationnelle. Cette modélisation permet de comparer sur les mêmes bases un certain nombre de critères de découpage de graphe c'est-à-dire de modularisation. Nous proposons une réécriture Relationnelle des critères de modularisation connus tels le critère de Newman-Girvan, Zahn-Condorcet, Owsiński-Zadrożny, Condorcet pondéré, Demaine-Immorlica, Wei-Cheng, la Différence de profils et Michalski-Goldberg. Nous introduisons trois critères : la Modularité Équilibrée, l'Écart à l'Indétermination et l'Écart à l'Uniformité. Nous identifions les propriétés vérifiées par ces critères et pour certains critères, notamment les critères linéaires, nous caractérisons les partitions obtenues via leur optimisation dans le but de faciliter leur compréhension et d'interpréter plus clairement leurs finalités en y associant la preuve de leur utilité dans certains contextes pratiques. Les résultats trouvés sont testés sur des graphes réels de tailles différentes avec l'algorithme de Louvain générique.

Mots-clefs

Modularisation, critères de modularisation, modularité, classification, communautés, Analyse Relationnelle Mathématique, graphes, réseaux, algorithme de Louvain

---

Modelling and extensions of mathematical Relational Analysis to complex networks clustering (graphs)

Abstract

Graphs are the mathematical representation of networks. Since a graph is a special type of binary relation, graph clustering (or modularization), can be mathematically modelled using the Mathematical Relational analysis. This modelling allows to compare numerous graph clustering criteria on the same type of formal representation. We give through a relational coding, the way of comparing different modularization criteria such as: Newman-Girvan, Zahn-Condorcet, Owsiński-Zadrożny, Demaine-Immorlica, Wei-Cheng, Profile Difference et Michalski-Goldberg. We introduce three modularization criteria: the Balanced Modularity, the deviation to Indetermination and the deviation to Uniformity. We identify the properties verified by those criteria and for some of those criteria, specially linear criteria, we characterize the partitions obtained by the optimization of these criteria. The final goal is to facilitate their understanding and their usefulness in some practical contexts, where their purposes become easily interpretable and understandable. Our results are tested by modularizing real networks of different sizes with the generalized Louvain algorithm.

Keywords

Modularization, modularization function, modularity, clustering, communities, Mathematical Relational Analysis, graphs, networks, Louvain algorithm

Table des matières

Introduction générale

Il existe aujourd'hui un renouveau incontesté de l'analyse des graphes. Ce renouveau s'explique essentiellement par l'engouement très actuel, lié à l'impact des "Réseaux Sociaux" dans les comportements *relationnels* communicants de nos contemporains, utilisateurs de média digitaux (portables, smart phones :'i-Phone', smart tablets : i-Pad, PC books etc..). En effet le développement des réseaux sociaux comme : FaceBook, Twitter, Flickr, Plaxo, Linkedln.. etc, et plus généralement des réseaux P2P (pairs à pairs) a rajouté un niveau supplémentaire de complexité à celles déjà créées par les approches exploratoires du Web en mode recherche d'information.

Or qui dit "graphe" sous- entend "relations" ou liens entre nœuds ou sommets du graphe qu'on appelle aussi "arêtes" (en cas de non orientation du lien, ou "arcs" dans le cas contraire), et qui dit "relations" préfigure leur analyse au moyen d'outils tels que l'Analyse Relationnelle, puisque c'est justement la vocation de cette discipline. Ceci justifie d'autant, (on le verra plus avant), que par le biais de l'analyse Relationnelle l'on s'intéresse en profondeur à ce sujet et que, dès lors, soient identifiées au sein de l'analyse des grands graphes et des grands réseaux, des thématiques qui sont en ligne directe avec les méthodologies, les approches et les critères qui seront évoqués dans les chapitres suivants.

L'une des difficultés de l'étude de ces réseaux et graphes du "Net" réside dans l'effet taille associé à des problématiques bien définies mais complexes, tels les processus de "navigation", de "cheminement", de "modularisation" au sein de ces grands graphes (on qualifie aujourd'hui ces difficultés liées à la taille des problèmes précédents d'effet "Big Data"). Cependant sans minimiser nullement les problèmes d'exploration des pages URL du Web à des fins de recherche d'information, par des algorithmes du type "PageRank" (Google), une grande difficulté résiduelle persiste néanmoins, celle relative à la "modularisation" de grands graphes (on parle également à ce sujet de recherche de "communautés" dans les réseaux).

En effet dans les grands réseaux, la détection de sous-ensembles de sommets (ou nœuds) plus systématiquement et densément connectés entre eux qu'avec les autres, appelés dans le jargon "Social Média" des "communautés" (c'est-à-dire des ensembles d'individus qui échangent plus particulièrement entre eux), est un problème rémanent que l'on retrouve également dans différents types de domaines, n'ayant pas grand-chose à voir avec les réseaux sociaux, tout au moins au niveau des disciplines impactées.

Ainsi en Biologie Moléculaire (l'interaction entre protéines), en Informatique (la recherche d'informations dans des réseaux sémantiques ou sur le Web d'une façon générale), dans les réseaux d'infrastructure IT (recherche des zones de connexion les plus denses en

vue de les isoler en Cyber Sécurité, appelée également analyse topologique des vulnérabilités d'un réseau), la recherche des zonages de tarification dans les réseaux de transports ou les réseaux de communications téléphoniques, la programmation en mode "clusters modulaires", appelée "Cluster Programming ou Cluster Computing" utilisant les partitionnements modulaires en clusters des codes développés, etc.. relèvent toutes de la même problématique qu'on peut traduire en langage de la Théorie des Graphes sous des intitulés variés comme : "recherche de classes dans un graphe", "partitionnement optimal de graphe", " modularisation optimale de graphes", "recherche de communautés", cette dernière expression caractérisant plus particulièrement le domaine des réseaux sociaux.

C'est l'intitulé : "modularisation optimale de graphe" que nous avons choisi, comme représentatif de l'ensemble car il évoque de façon sous-jacente la notion de "module" qui correspond assez bien au but que l'on s'est fixé, à savoir décomposer en sous-éléments ou sous ensembles, gérables et analysables, les grandes structures de réseaux, qui prises dans leur forme brute d'origine, sont particulièrement difficiles voire impossibles à appréhender dans leur entièreté.

Le fait qu'il s'agisse de trouver des classes homogènes (à fort partage de liens) dans un graphe, nous renvoie, bien entendu, à la notion de "partitionnement" de structures en classes disjointes (partition vraie) ou en classes chevauchantes (un individu, en l'occurrence ici un nœud ou un sommet [1], pouvant appartenir à deux classes différentes, voire plus, on parle alors de "recouvrement"). L'obtention de ce résultat nécessite que l'on dispose d'un "bon" critère, qualifiant le processus de partitionnement, afin de l'optimiser en fonction de considérations à la fois globales et contextuelles.

C'est le choix de critères potentiellement compatibles avec la difficulté de cette problématique qu'il importe de justifier et c'est l'ensemble des problématiques associées et leurs propriétés au sens axiomatique que nous aborderons et chercherons à comprendre plus en détail.

Le présent document est structuré de la façon suivante :

Pour aborder cette problématique de la "modularisation de graphes" ou dit autrement de la "recherche de modularité dans les graphes", il faut se replonger dans l'univers de la Théorie des Graphes. Le chapitre 1 présente, donc, une introduction à la théorie des graphes et la définition des principales propriétés que nous utiliserons dans la suite de ce document.

Le chapitre 2 présente l'outil principal que nous allons utiliser pour la comparaison des critères de modularisation : l'Analyse Relationnelle.

Le chapitre 3 met en exergue l'importance de la définition de la notion de *communauté* dans la formulation d'un critère de modularisation.

Le chapitre 4 présente un descriptif des principales propriétés qui peuvent être vérifiées par certains critères de modularisation. Nous nous focalisons sur trois propriétés : linéarité, séparabilité et équilibre. La propriété d'équilibre pour les critères linéaires est discutée plus

1. Dans ce document on utilisera les mots *nœud* et *sommet* de façon indistincte, bien que le premier fasse référence à un réseau et le deuxième à un graphe.

en détail et étendue. Les conséquences de la vérification ou non vérification des propriétés énoncées sont décrites dans ce chapitre.

Le chapitre 5 présente une liste non exhaustive des critères de modularisation trouvés dans la littérature en notations relationnelles. Les propriétés vérifiées par ces critères sont étudiées dans ce chapitre. Dans ce chapitre nous introduisons aussi trois critères de modularisation : la Modularité Équilibrée, l'Écart à l'Indétermination et l'Écart à l'Uniformité. Nous étudions plus en détail la dualité indépendance-indétermination qui nous permet d'élaborer un de ces nouveaux critères.

Le chapitre 6 présente une comparaison des partitions trouvées avec les critères énoncés au chapitre 5. Les caractéristiques des partitions trouvées via l'optimisation de chaque critère sont énoncées à la fin du chapitre.

Le chapitre 7 présente des applications pratiques. Nous approchons la partition optimale sur des graphes réels de tailles différentes avec l'algorithme de Louvain générique pour les critères étudiés au chapitre 5.

Finalement le chapitre 8 présente les conclusions et futurs travaux pour la suite du présent travail de recherche.

Chapitre 1

Introduction à la théorie des graphes

1.1 La théorie des graphes

Un graphe est une structure mathématique permettant de modéliser tout système consistant en un ensemble d'objets similaires susceptibles d'être liés par une certaine "relation typée".

Les grands graphes permettent de décrire des systèmes complexes issus de différents domaines. Exemples : en biologie (les réseaux d'interactions entre protéines), en informatique (les sites web et les hyperliens entre eux), en recherche opérationnelle (les entrepôts ou sites connectés en réseau et les échanges inter-sites), ou les structures de transports entre villes, représentables sous forme de réseaux d'infrastructure (routes, lignes de chemin de fer, lignes de tramways etc..), en sociologie (les réseaux sociaux et leurs interactions comme : les relations d'amitié, les relations de travail, les relations d'affinité en général etc.).

Le premier article publié sur la théorie des graphes date en fait de 1741. Ce document, dont l'auteur est le célèbre mathématicien suisse Leonhard Euler, traite du problème dit *des sept ponts de Königsberg*. Le problème consistait en fait à trouver un cheminement à partir d'un point donné qui fasse revenir à ce point de départ en passant une fois et une seule par chacun des sept ponts de la ville. Euler a considéré chaque ile de la ville de Königsberg comme un sommet et les sept ponts comme des arêtes.

Au milieu du XIXe siècle, le mathématicien britannique Arthur Cayley s'intéressa aux arbres, qui sont des graphes n'ayant pas de cycle, i.e. dans lequel il est impossible de revenir à un point de départ sans faire le chemin inverse. En particulier, il étudia le nombre d'arbres à N sommets et montra qu'il en existe n^{n-2}. Ceci constitua une des plus importantes formules en combinatoire énumérative. Le mathématicien anglais James Joseph Sylvester fut, quant à lui, le premier à utiliser le terme "graphe" en 1878, en relation avec certaines considérations sur la formation des molécules en chimie.

Une application très connue de la théorie des graphes est le problème dit des *quatre couleurs* énoncé en 1852 par le mathématicien sud-africain Francis Guthrie. Ce problème

consiste à déterminer combien de couleurs différentes il faut utiliser pour colorier une carte (donc le graphe associé) de façon telle que deux pays adjacents ne soient pas coloriés avec la même couleur (ce qui se traduit par l'expression suivante en Théorie des Graphes : comment colorier tous les sommets d'un graphe de façon à ce qu'aucun sommet n'ait la même couleur que les sommets voisins qui lui sont directement adjacents).

L'étude de ce problème entraîna de nombreux développements en théorie des graphes, entre autre par Peter Guthrie Tait, Percy John Heawood, Frank Ramsey et Hugo Hadwiger.

Cependant, et de façon plus classiquement attestée, les principaux développements sur la théorie des graphes et surtout leurs impacts dans l'univers mathématique, débutent vers la fin des années 50, avec, entre autres, le très utile Théorème de Ford et Fulkerson publié en 1956 sur l'équivalence "Flot Max/Coupe Min" (voir Ford and Fulkerson [1956], et la forte contribution au domaine de Claude Berge au travers de son ouvrage "Théorie des Graphes et Applications" publié en 1958 (Berge [1958]). Cet ouvrage fut assez fondamental au regard de la diffusion de cette discipline en France et en Europe. Claude Berge a d'ailleurs publié en 1970, un autre ouvrage chez le même éditeur Berge [1970], beaucoup plus complet et complexe que le précédent, intitulé "Graphes et Hypergraphes", qui est devenu de facto, un standard international sur le sujet, malgré le handicap d'avoir été dédié à un lectorat francophone. Il marquera également la recherche française en ce domaine, par la création conjointe avec Marcel-Paul Schützenberger d'un séminaire hebdomadaire à l'Institut Henri Poincaré, des réunions le lundi à la Maison des Sciences de l'Homme, et la direction de l'équipe Combinatoire de Paris.

Postérieurement aux ouvrages de C. Berge, le livre très didactique et complet de Gondran and Minoux [1995] plus orienté que ceux cités précédemment sur les algorithmes utiles en théorie des graphes, s'est avéré également comme un standard de facto. Côté américain, l'ouvrage de référence, dédié entièrement à la Théorie des Graphes, quoique postérieur à ceux de Claude Berge, reste le livre de Harary [1969], "Graph Theory" de 1969, qui s'est imposé comme un ouvrage didactique de base très cité et fortement diffusé dans le monde scientifique anglo-saxon.

Aujourd'hui les articles concernant les graphes et leurs usages concernent beaucoup moins qu'auparavant les aspects combinatoires (dénombrements de structures particulières) ou les implications de leurs propriétés intrinsèques en termes de descriptions et d'existence. Les articles récents sont en effet consacrés presque exclusivement (en se basant sur le nombre de publications constatées), via la dualité simultanée "Données massives" / "Réseaux Complexes", à des développements algorithmiques pratiques plus que de mathématiques pures, ou à la mise en exergue de principes de faisabilité plus qu'à des justifications axiomatiques profondes. Sans faire abstraction des résultats à mettre au crédit de la Théorie des Graphes par elle-même, et dont certains nous serons utiles, le sujet lié à la "modularisation", peut, d'une certaine façon, être considéré comme un sujet nouveau et récent, même si on peut néanmoins le faire remonter au début des années 2000.

1.2 Définition d'un graphe et principales propriétés utiles pour notre propos

Par définition un graphe (en se basant sur la théorie des ensembles) peut être décrit de la façon suivante

Définition 1.1 (Graphe)**.** *Un graphe est un couple $G = (V, E)$ d'ensembles finis. L'ensemble V est appelé ensemble des sommets (ou nœuds). E est un ensemble de paires (si graphe non orienté) ou de couples (si graphe orienté) d'éléments de V.*

Quelques remarques sur cette définition :

- L'ensemble V est supposé non vide. Son cardinal, le nombre N de sommets, noté $|V|$ est appelé l'ordre du graphe G.
- Le cardinal de l'ensemble $|E|$ sera noté M.
- Les éléments de l'ensemble E sont appelés *arêtes* si le graphe est non-orienté et *arcs* si le graphe est orienté.
- Une arête qui part d'un sommet et revient sur ce dernier s'appelle une *boucle*.

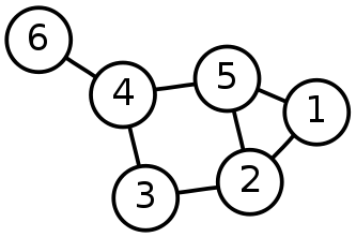

Figure 1.1 – Graphe non orienté à 6 sommets ($N = 6$) et 7 arêtes ($M = 7$)

1.2.1 Propriétés basiques à connaitre autour de la notion de graphe

- **Degré d'un sommet** : étant donné un graphe $G = (V, E)$ non-pondéré, le degré $d(v)$ du sommet $v \in V$ est le nombre d'arêtes aboutissant ou partant du sommet v. Dans le cas d'un graphe orienté on distingue entre degré entrant $d^+(v)$, le nombre d'arcs vers v et le degré sortant $d^-(v)$, le nombre d'arcs sortant de v. Par exemple, les sommets 2 et 6 de la figure 1.1 ont pour degrés 3 et 1 respectivement.

- **Matrice d'adjacence A** : étant donné un graphe non-orienté (G, V) à $|V| = N$ sommets et M arêtes. On note $v_1, v_2, ..., v_n$ les sommets de V. La matrice d'adjacence $\mathbf{A}$ de V est une matrice carrée d'ordre N dont les éléments sont définis de la façon suivante

$$a_{ii'} = \begin{cases} 1 & \text{s'il existe une arête entre } i \text{ et } i', \\ 0 & \text{sinon.} \end{cases} \tag{1.1}$$

  Toute l'information concernant la topologie du graphe est contenue dans la matrice d'adjacence. À titre d'exemple, la matrice d'adjacence du graphe de la figure 1.1 est donnée par la matrice

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

La matrice d'adjacence possède des propriétés à la fois utiles pour notre propos et remarquables par essence :

1. Elle est symétrique.
2. Son terme général est binaire, soit $a_{ii'} \in \{0, 1\} \forall i, i'$.
3. La somme de ses éléments est égal à deux fois le nombre total d'arêtes, soit $\sum_{i=1}^{N} \sum_{i'=1}^{N} a_{ii'} = 2M$.
4. La somme des éléments de la ligne (ou colonne) i est égal au degré du sommet i, soit $\sum_{i=1}^{N} a_{ii'} = \sum_{i'=1}^{N} a_{ii'} = d(i)$.

– **Matrice des degrés D** : il s'agit d'une matrice diagonale dont l'élément d_{ii} correspond au nombre de connexions du sommet i, c'est-à-dire à son degré.

A titre d'exemple, la matrice des degrés du graphe de la figure 1.1 est donnée par la matrice suivante

$$\mathbf{D} = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

– **Matrice Laplacienne L**(Appelée aussi *Matrice d'admittance*) : la matrice laplacienne d'un graphe G non-orienté et non-réflexif est définie par

$$L = D - A.$$

.
C'est-à-dire la différence entre la matrice des degrés D et la matrice d'adjacence A. La matrice laplacienne du graphe de la figure 1.1 est donnée par

$$\mathbf{L} = \begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix}$$

La matrice laplacienne est utilisée dans le cadre du partitionnement de graphe par les méthodes spectrales comme le nous verrons par la suite.

1.3 Différents types de graphes

- **Graphe simple** : un graphe qui contient ni boucles ni arêtes multiples.
- **Graphe orienté** et **graphe non-orienté** : dans le cas d'un graphe orienté $G = (V, E)$ l'ensemble E est un ensemble des *couples* de V ; dans le cas d'un graphe non-orienté, E est un ensemble des *paires* de V. Un *couple* est une *paire ordonnée*, ainsi l'ordre des objets composant un couple est important. Par exemple dans la figure 1.2 le lien existant entre les sommets 4 et 5 part du sommet 4 vers le sommet 5 et il est noté $(4, 5)$ et non pas $(5, 4)$ qui signifierait que l'arc va du sommet 5 vers le sommet 4. La flèche indique la direction.

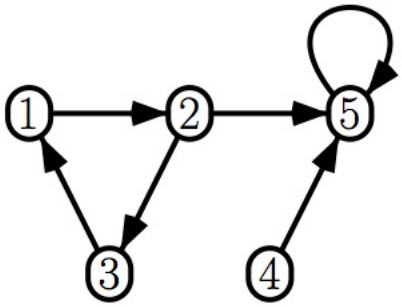

Figure 1.2 – Graphe orienté à 5 sommets et 6 arcs.

- **Graphe complet** : graphe non-orienté dont toutes les paires de sommets sont connectées par une arête. Un graphe complet à N sommets est noté K_N et le nombre total d'arêtes qu'il possède est égal à $\frac{N(N-1)}{2}$.

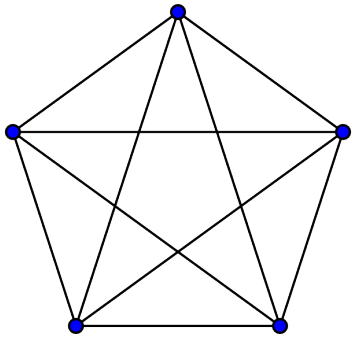

Figure 1.3 – Graphe complet à 5 sommets : K_5.

La figure 1.3 montre un graphe complet K_5, il contient, donc 10 arêtes. Étant donné un graphe G, une *clique* est un sous-graphe complet de G. Par exemple, dans la figure 1.1 le sous-graphe induit par les sommets $1, 2$ et 5 est une clique K_3.

- **Graphe pondéré** : un graphe peut être pondéré par rapport à ses sommets ou par rapport à ses arêtes. Dans chaque cas il existe une fonction d'évaluation appelé *poids* attribuée à chacun de ses sommets ou à chacune de ses arêtes. Dans la plupart des ouvrages un graphe pondéré correspond au deuxième cas et le poids peut être in-

terprété comme une mesure de l'importance que l'on attribue à chaque arête. Dans la suite, on utilisera le terme *graphe pondéré* en se référant à un graphe dont les arêtes sont pondérées.

La matrice d'adjacence d'un graphe pondéré, notée **W**, est appelée *Matrice des poids*. Elle se différencie de la matrice d'adjacence d'un graphe non pondéré par le fait de posséder des valeurs réelles au lieu de binaires. Son terme général $w_{ii'}$ peut être interprété comme le nombre total d'arêtes reliant le sommet i au sommet i' ou vice-versa.

– **Graphe connexe** : un graphe non orienté $G = (V, E)$ est dit connexe si quels que soient les sommets u et v de V, il existe une suite d'arêtes permettant d'atteindre v à partir de u ou l'inverse.

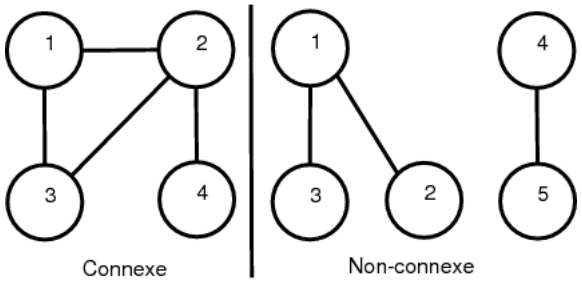

FIGURE 1.4 – Graphe connexe

– **Arbre** : c'est un graphe non-orienté où chaque couple de sommets est connecté par un chemin unique. C'est-à- dire qu'il n'y a pas de *cycle* dans le graphe. Par conséquent, un arbre d'ordre N possède $N - 1$ arêtes. Si une seule arête est enlevée le graphe devient déconnecté. Si l'on ajoute une seule arête sans rajouter un nœud il y aura au moins un cycle.

– **Graphe biparti** : un graphe est dit biparti s'il existe une partition de son ensemble de sommets en deux sous-ensembles U et V telle que chaque arête ait une extrémité dans U et l'autre dans V. Un graphe est dit **biparti complet** (ou encore est appelé une biclique) s'il est biparti et en plus il contient le nombre maximal d'arêtes. En d'autres termes, chaque sommet de U est relié à chaque sommet de V. Si U est de cardinal N_1 et V est de cardinal N_2 le graphe biparti complet est noté $K_{N_1,N-2}$.

– **Graphe étoile** : c'est un graphe biparti complet $K_{1,k}$. On peut aussi le voir comme un arbre avec un sommet central et k feuilles, du moins lorsque $k > 1$. Par exemple, la figure suivante montre un graphe étoile $K_{1,8}$, le degré du sommet central est égal au nombre total d'arêtes, 7 sommet bleu.
– **Graphe aléatoire (random graph)** : c'est un graphe généré par un processus aléatoire. Le modèle le plus connu servant à générer des graphes aléatoires est celui d'Erdős and Rényi [1959] [1]. Dans un graphe aléatoire d'Erdös-Renyi la probabilité qu'il existe une arête entre toute paire de sommets est la même pour toutes les paires et elle est indépendante des autres paires de sommets.

1. Le concept de graphe aléatoire a servi de base pour la conception du critère de modularisation le plus connu, à savoir, la modularité de Newman-Girvan.

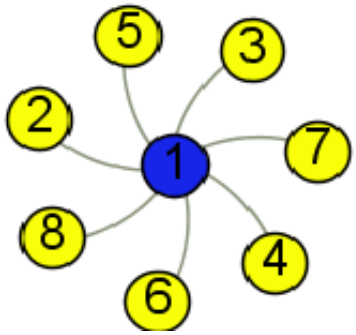

Figure 1.5 – Graphe étoile.

1.4 Indicateurs structurels des graphes

1.4.1 Coefficient de classification

Le coefficient de classification d'une structure réticulaire (graphes ou réseaux) mesure la propension qu'ont les sommets d'un graphe à former des communautés auto organisées ou à se séparer en classes. Dans la plupart des réseaux réels, en particulier les réseaux sociaux, la probabilité de voir se former des communautés tend à être plus importante que la probabilité moyenne qu'il existe un lien entre deux sommets et que ce dernier ait été obtenu par hasard pur.

Coefficient de classification Global

Le coefficient de classification global est obtenu à partir du comptage des *triplets* de sommets. Un triplet de sommets est un groupe de trois sommets soit connectés par deux arêtes (triplet ouvert), soit connecté par trois arêtes (triplet fermé). Étant donné un graphe $G = (V, E)$ le coefficient de classification global de G, $C(G)$, est le quotient entre le nombre total de triplets fermés et le nombre total de triplets (ouverts et fermés)

$$C(G) = \frac{\text{nombre de triplets fermés}}{\text{nombre de triplets connectés}}. \tag{1.2}$$

Coefficient de classification local

Le coefficient de classification local est une mesure locale car il est calculé pour chaque sommet. Il mesure à quel niveau un sommet et ses *voisins* sont susceptibles de former un graphe complet [2].

Définition 1.2 (Sommet voisin)**.** *Étant donné un graphe $G = (V, E)$ le voisinage η_i pour un sommet $v_i \in V$ est défini comme l'ensemble de sommets adjacents à v_i appelés voisins :*

$$\eta_i = \{v_j : e_{ij} \in E \text{ ou } e_{ji} \in E\}.$$

Pour un graphe orienté le coefficient de classification local du sommet v_i, $C(i)$ est la proportion entre le nombre d'arcs à l'intérieur du voisinage de v_i divisé par le nombre maximal des arcs qui pourraient exister dans η_i, soit

2. Cette mesure fut introduite par Duncan J. Watts and Steven Strogatz en 1998 pour déterminer la proximité d'un *graphe petit monde.*

$$C(i) = \frac{|e_{jk}|}{|\eta_i|[(|\eta_i| - 1)}; v_j, v_k \in \eta_i; e_{ek} \in E. \tag{1.3}$$

Le numérateur représente le nombre d'arcs du sous-graphe induit par le voisinage du sommet v_i. Comme il s'agit d'un graphe orienté, la direction de chaque sommet est prise en compte, le dénominateur est calculé alors comme le nombre d'*arrangements* de 2 voisins parmi les $|\eta_i|$ existants.

Si le graphe est non-orienté l'arête n'a pas de direction, le coefficient de classification local se calcule alors de façon analogue au graphe orienté :

$$C(i) = \frac{2|e_{jk}|}{|\eta_i|[(|\eta_i| - 1)}; v_j, v_k \in \eta_i; e_{ek} \in E. \tag{1.4}$$

Cette fois-ci le dénominateur est calculé comme le nombre de combinaisons de 2 sommets parmi les $|\eta_i|$ possibles.

Le coefficient de classification peut prendre des valeurs entre 0 et 1.

À titre d'exemple la figure 1.6 montre le calcul du coefficient de classification pour un sommet avec 3 voisins dans un graphe non orienté.

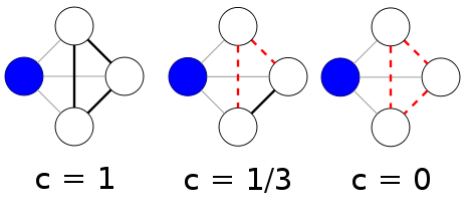

FIGURE 1.6 – Coefficient de classification local du sommet bleu. Les segments noirs montrent les connexions existantes, les segments rouges en pointillé montrent les connexions non-existantes entre les voisins.

La figure 1.6 montre le coefficient de classification pour le sommet bleu dans un graphe non orienté. Ce sommet possède 3 voisins, si ces derniers forment un graphe complet il y a 3 connexions entre eux, ils forment alors un graphe complet et $C = 1$ (gauche). En revanche, s'il n'y a aucune connexion entre voisins, on a naturellement $C = 0$ (cas du graphe de droite).

1.4.2 Mesures de centralité *(centrality)*

Les mesures de centralité d'un sommet déterminent l'importance relative du sommet dans le graphe. Cela peut être utile dans plusieurs circonstances : dans un réseau social on peut s'intéresser, par exemple, à l'importance qu'une personne a par rapport aux autres membres du réseau ou de sa communauté.

Historiquement les mesures de centralité furent introduites autour des années 1960 de façon assez théorique par Moon and Moser [1965] (pour la théorie des graphes) et par les spécialistes du "clustering" (cette notion étant associée dans ce cas au concept de "meilleur

représentant" d'une classe ou d'un "cluster"), et de façon plus pragmatique dans l'intention de détecter les composantes essentielles dans les réseaux informatiques, ce concept fut repris par Alex Bavelas et son groupe de travail "Group Networks Laboratory" au(MIT) Bavelas [1948]. Puis à partir des années 70 d'autres auteurs ont employé ces mesures pour faire référence aux flux d'informations transportées par les réseaux informatiques, ainsi qu'aux flux générés par les mouvements et transferts financiers dans les réseaux bancaires, mais cela sert aussi à l'étude de la propagation des rumeurs ou des fausses informations dans le domaine de la e-Réputation ou encore à l'étude de la propagation d'infections en épidémiologie.

La centralité d'un sommet ou d'un nœud est une propriété structurelle d'un sommet relativement au graphe qui le contient, et non pas intrinsèque au sommet lui-même. Il s'agit d'une qualification que le sommet possède en vertu de sa position relative par rapport aux autres sommets. La centralité est une sorte de mesure de l'influence qu'un sommet peut avoir sur les autres ou de la pertinence qu'il peut avoir à représenter les autres. Dans un graphe en forme d'étoile, par exemple, le sommet central possède une valeur de centralité élevée tandis que les autres sommets situés en périphérie ont une valeur de "centralité" beaucoup plus faible.

Les mesures de centralité les plus utilisées dans la littérature sont au nombre de 4 : centralité de degré, centralité de proximité, centralité d'intermédiarité et centralité spectrale (vecteur propre).

Centralité de degré *(degree centrality)*

Dans la section 1.2.1 nous avons défini le degré d'un sommet. Il s'agit du nombre de liens qu'un sommet possède avec les autres sommets. Dans un réseau social le degré peut être interprété comme le nombre de connexions qu'une personne possède avec les autres, c'est une mesure de la popularité de cette personne ; dans le cas d'un phénomène de propagation d'infections il s'agit d'une mesure du risque de contamination ; dans le cas d'une rumeur qui circule dans le réseau cette mesure peut être le moyen de caractériser le risque de propagation d'informations délictueuses ou fausses.

Mathématiquement étant donné un graphe à N sommets le degré du sommet i est la somme des éléments de la ligne (ou colonne) i de la matrice d'adjacence $\mathbf{A}$, soit

$$d_i = \sum_{i'=1}^{N} a_{ii'} \qquad \forall i. \tag{1.5}$$

Une façon de normaliser cette mesure est de la diviser par la valeur maximale qu'elle peut prendre, soit $N-1$ voisins pour un graphe non-orienté :

$$\hat{d}_i = \frac{d_i}{N-1}. \tag{1.6}$$

Si le graphe est orienté et selon que les arcs entrent ou sortent du somment i il y a deux mesures différentes de "degré de centralité". Dans le premier cas on compte le nombre d'arcs entrant vers i (*in-degree*) et dans le deuxième cas on compte le nombre d'arcs sortant du sommet i (*out-degree*). Dans un réseau social le premier permet de connaître la popularité du sommet concerné et le deuxième est plutôt un indicateur de la grégarité associée.

Centralité de Proximité *(Closeness centrality)*

Dans Bavelas [1948], l'auteur avait déjà suggéré d'employer la proximité comme mesure de centralité. Il avait remarqué qu'un message originaire du sommet le plus central se propagerait dans le réseau en temps minimal. Plus tard, en 1965 d'autres auteurs comme Beauchamp, Hakimi et Sabidussi ont donné une définition plus précise au "*sommet le plus au centre d'un réseau*" comme : "le point à partir duquel les autres points sont atteints à coût et temps minimaux".

En 1966 Sabidussi [3] Sabidussi [1966] a proposé de mesurer la centralité d'un sommet au moyen de la somme des distances géodésiques du sommet considéré aux autres sommets dans le graphe, en fait cette idée remonte bien plus avant historiquement . En effet dans le cas général, elle fait allusion de façon sous-jacente au concept de "point médian" au sens du problème de Fermat Weber, ou aux "médianes" au sens de Fréchet. Cette mesure est inverse à la centralité car plus loin le sommet est situé par rapport aux autres sommets plus importante est la valeur de cette mesure. Dans Freeman [1979] l'auteur définit la distance géodésique comme [4] :

$d(i, i')$ = nombre minimal d'arcs (ou arêtes) reliant les sommets i et i'.

En théorie des fluides cette distance permet d'estimer le temps nécessaire d'arrivée du flux entre un point de départ et un point d'arrivée. Elle mesure aussi, d'un autre point de vue, l'accessibilité entre les deux sommets.

D'après Sabidussi [1966] la mesure de proximité $C_c(i)$ du sommet i est

$$C_c(i) = \frac{1}{\sum_{i' \neq i} d(i, i')}. \tag{1.7}$$

Plus le sommet i se trouve près des autres sommets plus grande sera sa proximité.

Cependant cette façon de mesurer la proximité n'est valide que si le graphe est connexe. Si le graphe n'est pas connexe tous les sommets ne peuvent pas être atteints à partir d'une partie du reste des sommets. Dans ce cas-là la distance entre eux est infinie.

Comme la définition de proximité de l'équation (1.7) ne prend pas en compte le nombre de sommets du graphe, il devient impossible de comparer sa valeur calculée entre graphes

3. En fait cette notion fait référence à l'approche "médiane" de Fermat Weber. Approche consistant à trouver un point médian, sorte de "résumé" issu d'une collection d'informations multidimensionnelles, en cherchant une solution à "éloignement minimal" par rapport à des situations initiales fixées au départ. Ces situations initiales, qui correspondent à des structures de références ou de comparaison, sont soit un ensemble de points fixes, validés, auxquels on tente de ressembler ou bien représentatifs de buts à atteindre, soit des points de repères témoins, auxquels, soit il faut accéder, soit il est nécessaire de se comparer ou de s'étalonner et qu'il faut globalement approximer "au mieux". Ce principe existe depuis fort longtemps et ne date pas d'aujourd'hui : preuve en est le Problème dit de "Fermat-Weber", introduit par Pierre de Fermat en (1640), qui revient à chercher le point à distance minimale de points fixés et qui a été résolu dans ce cas de 3 points par Evangelista Torricelli en 1647, puis généralisé en 1909 par Alfred Weber au problème de "localisation à distance minimale" sur un nombre fini de "points de repères" ou de situations de départ (voir à ce propos l'article historique et bien étayé de Hiriart-Urruty [2004]).

4. Cette définition de distance provient du problème dit de cheminement (Shortest path problem) consistant à trouver le chemin le plus court entre deux sommets. Un des problèmes les plus étudiés dans la théorie des graphes

de tailles différentes. Pour résoudre ce problème en 1965 Bauchamp a proposé la proximité relative

$$\hat{C}_c(i) = \frac{N-1}{\sum_{i' \neq i} d(i,i')}. \tag{1.8}$$

Cette dernière expression, valable, encore une fois, uniquement pour les graphes connexes, est en réalité une moyenne de distances entre le sommet i et les $N-1$ sommets restants dans le graphe.

Centralité d'intermédiarité *(Betweenness centrality)*

Lorsque Bavelas a introduit la notion de centralité dans Bavelas [1948], il a indiqué qu'un sommet peut être considéré comme central dans un réseau s'il est présent sur la plupart des plus courts chemins reliant deux autres sommets quelconques dans le réseau. Ainsi si deux sommets i' et i'' doivent communiquer entre eux, et que pour cela ils doivent absolument passer par le sommet i, ce dernier devient certainement responsable de cette mise en contact, il est donc, un "intermédiaire".

Dans Freeman [1977] Freeman a donné une expression permettant de calculer l'intermédiarité : étant donné un graphe $G = (V, E)$, la mesure d'intermédiarité du sommet i indique la fréquence de présence du sommet i dans le plus court chemin entre les sommets i' et i'', soit

$$C_B(i) = \sum_{i \neq i' \neq i'' \in V} \frac{\sigma_{i'i''}(i)}{\sigma_{i'i''}}, \tag{1.9}$$

où :

- $C_B(i)$ est la mesure d'intermédiarité du sommet i.
- $\sigma_{i'i''}$ est le nombre des plus courts chemins entre i' et i''.
- $\sigma_{i'i''}(i)$ est le nombre des plus courts chemins entre i' et i'' passant par i.

L'intermédiarité $C_B(i)$ peut être normalisée en divisant par le nombre total de chemins dont ni le point de départ ni le point d'arrivée est i (soit le nombre maximal de chemins qui pourraient contenir i). Pour un graphe orienté, cette quantité est $(N-1)(N-2)$. Pour un graphe non orienté cette quantité est $(N-1)(N-2)/2$. Par exemple, pour un graphe non orienté en forme d'étoile, le sommet central (lequel est contenu dans tous les plus courts chemins possibles) aurait une intermédiarité égale à 1 tandis que les feuilles auraient une intermédiarité nulle.

Centralité spectrale *(Eigenvector centrality)*

Cette mesure de centralité, proposée par Bonacich [1972], assigne des scores relatifs au degré de chaque sommet. Le degré, comme valeur de centralité compte tout simplement le nombre de connexions que chaque sommet possède. La centralité spectrale somme toutes les connexions en donnant plus d'importance aux sommets qui possèdent à leur tour une forte connectivité. Par exemple, dans un réseau social la connexion de deux personnes qui sont eux-mêmes influentes augmente leur degré d'influence dans le réseau, qui

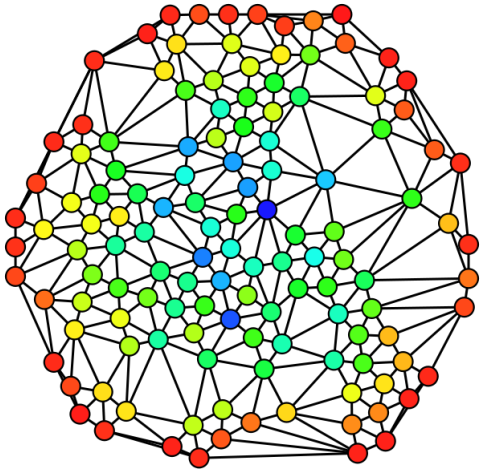

FIGURE 1.7 – L'intermédiarité augmente au fur et à mesure que l'on avance vers le centre, allant du rouge avec une valeur basse de l'intermédiarité au bleu avec une valeur d'intermédiarité maximale.

est plus élevé que celui de deux personnes qui sont connectées mais qui ont peu d'influence.

Étant donné un graphe $G = (V, E)$, soit s_i le score du sommet i, et $\mathbf{A}$ la matrice d'adjacence du graphe. Le score s_i doit être proportionnel à la somme de scores de tous les sommets connectés à i, soit

$$s_i = \frac{1}{\lambda} \sum_{j \in \eta(i)} s_j = \frac{1}{\lambda} \sum_{j=1}^{N} A_{i,j} s_j, \tag{1.10}$$

où :
– $\eta(i)$ est l'ensemble des voisins du sommet i.
– λ est une constante.

Si $\mathbf{s} = (s_1, s_2, ..., s_n)$ est le vecteur de scores, l'équation (1.10) devient

$$\mathbf{s} = \frac{1}{\lambda}\mathbf{A}\mathbf{s} \quad \text{parce que} \quad \mathbf{A}\mathbf{s} = \lambda\mathbf{s}. \tag{1.11}$$

L'équation (1.11) montre que $\mathbf{s}$ est un vecteur propre de $\mathbf{A}$ avec comme valeur propre λ. Comme la matrice d'adjacence est d'ordre N, il y aura N valeurs différentes de λ pour chaque vecteur propre. Cependant les scores s_i doivent être non-négatifs, cela implique selon le théorème de Perron-Frobenius que λ doit être la plus grande valeur propre et $\mathbf{s}$ son vecteur propre associé (cf. Newman [2008]).

Par ailleurs les définitions données ci-dessus sont essentiellement dues à Alex Bavelas extraites de son livre écrit en 1951, ceci dit, elles ne font pas l'unanimité chez les chercheurs en théorie des graphes, des définitions, plus raffinées, ont été données par d'autres, comme par exemple Claude Flament dans le livre « Réseaux de Communications et Structures Sociales », livre de 196 pages, édité par Dunod en (1966)], où il propose une critique constructive des indices de Bavelas et dont le but sous-jacent est de corriger les disparités et biais dûs aux définitions précédentes. On trouvera dans l'article suivant de Pierre Parlebas :

≪Centralité et Compacité d'un graphe ≫, Mathématiques et Sciences Humaines, Tome 39, pp : 5-27, (1972), une présentation simplifiée et pédagogique, tout en étant complète, autour de la notion de « centralité » et de compacité des graphes, où sont reprises les nouvelles définitions de Claude Flament, comparativement à celles de Bavelas.

1.5 Densité d'un graphe et degré moyen

La densité d'un graphe mesure le rapport entre le nombre d'arêtes existantes et le nombre maximal d'arêtes qui pourrait exister. Mathématiquement, la densité d'un graphe non orienté, non pondéré et non réflexif, notée δ, est définie comme

$$\delta = \frac{2M}{N(N-1)} \tag{1.12}$$

ou lorsque le graphe est réflexif

$$\delta = \frac{2M}{N^2}. \tag{1.13}$$

Les deux valeurs sont proches lorsque N tend vers l'infini. Dans ce document, sauf mention contraire, nous considérerons la seconde définition. À partir de ces deux définitions nous pouvons déduire que la densité est toujours comprise entre 0 et 1. Dans le cas d'un graphe complet 1. Un graphe dont la densité est proche de 1 est considéré *dense*. Il n'existe pas une convention universelle par rapport à la valeur de densité à partir de laquelle un graphe est considéré dense. Cependant quelques définitions ont été données par Lee and Streinu [2008] et Streinu and Theran [2009].

À partir de la définition de densité, il est possible de définir le *degré moyen* d'un graphe, noté d_{av} (où "av" vient du mot anglais average qui signifie moyen) :

$$d_{av} = \frac{1}{N}\sum_{i=1}^{N} d_i = \frac{2M}{N} = N\delta. \tag{1.14}$$

La plupart des grands réseaux réels ont une densité peu élevée comme le montre le tableau suivant extrait de la thèse de Guillaume [2004] :

	Internet	Web	Acteurs	Co-signature	Cooccurrence	Protéines
N	75 885	325 729	392 340	16 401	9 297	2 113
M	357 317	1 090 108	15 038 083	29 552	392 066	2 203
d_{av}	9,42	6,69	76,66	3,60	84,34	2,09
δ	$1{,}24 \cdot 10^{-4}$	$2{,}05 \cdot 10^{-5}$	$1{,}95 \cdot 10^{-4}$	$2{,}20 \cdot 10^{-4}$	$9{,}07 \cdot 10^{-3}$	$9{,}87 \cdot 10^{-4}$

Table 1.1 – Densité et degré moyen des réseaux réels.

Les réseaux du tableau 1.1 sont de tailles différentes et issus des domaines variés :
- *Internet* est un réseau d'une carte de l'internet au niveau de routeurs.
- *Web* est le réseau du graphe du Web de l'Université Notre-Dame.
- *Acteurs* est un graphe des Acteurs.
- *Co-signature* est un graphe des relations de co-signature d'articles sur Arxiv.

– *Cooccurrence* est un réseau de la cooccurrence de mots dans la Bible.
– *Protéines* est un réseau d'interaction protéiques.

1.6 Diamètre d'un graphe

Avant de définir le diamètre d'un graphe il est nécessaire de définir la distance entre deux sommets. Dans un graphe la distance entre deux sommets est la longueur d'un plus court chemin entre ces deux sommets.

Le diamètre d'un graphe est l'excentricité maximale de ses sommets, c'est-à-dire la plus grande distance possible qui puisse exister entre deux de ses sommets.

Les premières études sur la distance entre les sommets des grands réseaux sociaux datent de 1929 et ont été effectuées par le Hongrois Frigyes Karinthy sous le nom de théorie de "six degrés de séparation" (aussi appelée Théorie des 6 poignées de main). Cette théorie évoque la possibilité que toute personne sur le globe peut être reliée à n'importe quelle autre, au travers d'une chaîne de relations individuelles comprenant au plus cinq autres maillons. Cette théorie a été reprise par Milgram [1967] à travers l'étude du phénomène dit du "petit monde" (également connu sous le vocable ≪paradoxe de Milgram≫ car ses résultats semblent contraires à l'intuition). L'étude du petit monde est l'hypothèse que chacun puisse être relié à n'importe quel autre individu par une courte chaîne de relations sociales.

Cette théorie peut se démontrer de nos jours avec le réseau social Facebook, qui met en évidence les liens que nous avons avec les autres et les liens que nous avons avec des personnes que nous ne connaissons pas (amis des amis). Elle est encore plus manifeste sur LinkedIn, qui signale le degré de séparation entre deux individus ainsi que les ≪chemins≫ possibles qui relient un individu à un autre à travers leurs réseaux relationnels respectifs. Ainsi, par exemple, sur le réseau social Facebook [5] le degré de séparation de deux individus a été mesuré à 4,74 en 2008.

5. Voir l'article paru dans "20 minutes" entitulé : "Facebook a rétréci le monde, ramenant les ≪six degrés de séparation ≫ à 4,74 en moyenne" à l'adresse http ://www.20minutes.fr/ledirect/828370/facebook-retreci-monde-ramenant-six-degres-separation-474-moyenne.

Chapitre 2

L'Analyse Relationnelle

2.1 Introduction

L'Analyse Relationnelle et Mathématique des Données, en tant que discipline a été développée à l'origine en 1972 au Centre Scientifique IBM de Paris. Le besoin de création de cette discipline nait dans le but de répondre à quelques défis théoriques et applicatifs, liés à la résolution exacte de problèmes d'Optimisation et Recherche Opérationnelle et d'Analyse des données, réputés complexes. Parmi lesquels on peut citer : les " Classements Consensus en Théorie des votes", problème également connu sous le vocable de "Recherche d'ordres médians" ou "Problème de John Kemeny". D'un point de vue théorique, l'ARM étudie les relations binaires et particulièrement les problèmes qui ont trait à leurs mesures d'associations, leurs agrégations et à la détermination de relations consensuelles. Des travaux plus récents ont mis en évidence (voir Labiod [2008]) que le schéma valable pour les problématiques citées plus haut se généralise à des critères récemment proposés pour résoudre des problèmes de théorie des graphes dont par exemple le problème de "modularité optimale dans les grands graphes", qui est en plein essor, principalement du fait du fort développement actuel des "Réseaux Sociaux".

L'idée originale de Condorcet (voir Condorcet [1785]), qu'on peut résumer simplement à l'introduction de la notion de "comparaisons par paires", était plus puissante que certains l'ont d'abord imaginée [1]. En fait, elle préfigurait l'approche relationnelle associée, qui, elle, a permis de formaliser l'ensemble des problématiques précédentes au travers d'une convention de notations unique et unifiée, induisant un nombre important d'axiomes et de propriétés ayant permis d'en faire une Théorie, applicable à d'autres problématiques que la seule recherche d'ordres consensus.

Ainsi, il a été prouvé que le problème de Recherche d'une Relation d'équivalence à "Distance Minimale d'un Graphe Symétrique" (posé par Zahn [1964]), dérivait complètement du critère proposé sous forme littérale en 1785, en théorie des votes, par Antoine Caritat, Marquis de Condorcet (voir Condorcet [1785]).

Nous verrons également par la suite, que l'ARM n'est pas une théorie à part, mais bien une approche qui permet d'unifier et de structurer différents concepts, d'abord d'une façon formelle au travers d'une systématique de notations, mais également d'une façon

1. Voir les travaux de Guilbaud [1952] Premier texte français où l'on introduit les travaux de Condorcet et en particulier le fameux "Effet Condorcet"

plus structurelle, en généralisant des méthodes a priori très différentes les unes des autres. Par exemple, dans un article assez long publié en 1991 (voir Marcotorchino [1991] dont on peut lire des extraits dans Marcotorchino [1989] ou Marcotorchino [2000]), on peut trouver un argumentaire détaillé sur les ponts validés et fondamentaux entre l'Analyse Factorielle des Correspondances et l'ARM.

Il existe un lien direct entre l'Analyse Relationnelle Mathématique et la théorie des graphes, selon le principe qu' "*un graphe peut être considéré comme une structure mathématique servant á modéliser les relations binaires entre objets d'un même ensemble*".

2.2 Principales définitions

Les bases fondamentales de l'ARM reposent sur la définition de la notion de *Relation Binaire*. Voici une définition générale d'une relation binaire :

Définition 2.1. ***Relation Binaire*** *Une relation binaire $\mathcal{R}$ entre deux ensembles E et F (ou de E vers F) est un sous-ensemble du produit cartésien $E \times F$, soit une collection de couples dont la première composante est dans E et la seconde dans F.*

Si les ensembles E et F sont différents, on parle de *Relation Binaire bipartite*. Ce qui ne sera pas l'objet de cette thèse, nous traiterons essentiellement des relations binaires croisant le même ensemble. Dans le cas d'un graphe, l'ensemble de départ et l'ensemble d'arrivée sont identiques c'est l'ensemble des sommets V du graphe et chaque lien (une arête ou un arc) représente une relation binaire entre les deux sommets. Ainsi, un graphe représente une relation binaire croissant le même ensemble :

Définition 2.2. ***Relation Binaire sur le même ensemble***
Une relation binaire $\mathcal{R}$ sur un ensemble V est un sous-ensemble du produit cartésien $V \times V$, noté $G(\mathcal{R})$ et appelé graphe de la relation.

Ainsi si le couple (u, v) (où $u \in E$ et $v \in E$) appartient à ce sous-ensemble alors u et v sont en relation via la relation $\mathcal{R}$, cela s'écrit $u\mathcal{R}v$. Voici quelques exemples pratiques de relations binaires $\mathcal{R}$:

- "Plus grand que..."
- "Plus petit que..."
- "Appartient à la même classe que..."
- "Inférieur ou égal à..."
- "Supérieur ou égal à..."
- "A la même propriété que..."
- etc...

Toute relation binaire $\mathcal{R}$ possède sa relation complémentaire notée $\not{\mathcal{R}}$ dont la définition est la suivante :

Définition 2.3. ***Complémentaire d'une Relation Binaire***
Étant donnée une relation binaire $\mathcal{R}$ sur l'ensemble E, sa relation complémentaire $\not{\mathcal{R}}$ est un sous-ensemble du produit cartésien $E \times E$, tel que $(u, v) \notin G(\mathcal{R})$ (où $u \in E$ et $v \in E$),

cela s'écrit $u \not\mathcal{R} v$.

2.3 Propriétés générales des Relations Binaires

Soit $\mathcal{R}$ une relation binaire qui croise l'ensemble V avec lui même, alors $\mathcal{R}$ est :

- **Réflexive** si et seulement si $\forall i \in V : i\mathcal{R}i$. Chaque objet de V est en relation avec lui-même.
- **Complète** ou **totale** si et seulement si $\forall (i, i') \in V \times V$ tel que $i \neq i'$, soit $i\mathcal{R}i'$ ou $i'\mathcal{R}i$. La relation existe obligatoirement dans l'un des deux sens.
- **Symétrique** si et seulement si $\forall (i, i') \in V \times V : i\mathcal{R}i' \Rightarrow i'\mathcal{R}i$.
- **Asymétrique** si et seulement si $\forall (i, i') \in V \times V : i\mathcal{R}i' \Rightarrow \neg(i'\mathcal{R}i)$.
- **Antisymétrique** si et seulement si $\forall (i, i') \in V \times V : i\mathcal{R}i'$ et $i'\mathcal{R}i \Rightarrow i = i'$.
- **Transitive** si et seulement si $\forall (i, i', i'') \in V \times V \times V$: si $i\mathcal{R}i'$ et $i'\mathcal{R}i'' \Rightarrow i\mathcal{R}i''$.

En fonction des propriétés qu'elles vérifient il existe des relations typées. Par exemple :

- Une Relation de *Pré-ordre* est *Réflexive*, et *Transitive*.
- Une Relation de *Pré-ordre Total* est *Réflexive*, *Transitive* et *Complète*.
- Une Relation *d'Ordre Total* est *Transitive*, *Asymétrique* et *Complète*.

Définition 2.4 (Relation d'équivalence)**.** *Une relation d'équivalence* $\mathcal{R}$ *dans un ensemble* V *est une relation binaire qui est à la fois réflexive, symétrique et transitive.*

Théorème 2.1 (Relation d'équivalence et Partition)**.** *Si* $\mathcal{R}$ *est une Relation d'équivalence sur un ensemble* V*, alors il existe une partition* P *de* V *telle que*

$$i\mathcal{R}i' \Longleftrightarrow C(i) = C(i')$$

Où $C(i)$ *est la classe de* P *qui contient l'élément* i *de* V*. Réciproquement, si* P *est une partition de* V*, alors la relation définie par* $i\mathcal{R}i'$ *implique qu'il existe une classe* C *de* P *qui contient* i *et* i'*.*

Les notions de relation d'équivalence et de partition sont donc fondamentalement équivalentes.

2.3.1 Représentations relationnelles par contraintes linéaires

Désormais nous parlerons de relations binaires croisant le même ensemble V, dont le cardinal est $N = |V|$. Il existe une matrice permettant de caractériser une relation binaire $\mathcal{R}$. Il s'agit de la matrice *unitaire* relationnelle de Condorcet de comparaisons par paires. Cette matrice, notée $\mathbf{C}$ et de taille $(N \times N)$, est définie de la façon suivante

$$c_{ii'} = \begin{cases} 1 & \text{si } i\mathcal{R}i' \, \forall (i, i') \in V \times V, \\ 0 & \text{sinon.} \end{cases} \tag{2.1}$$

Comme nous allons le voir par la suite, la matrice $\mathbf{C}$ est un outil important pour l'ARM. En effet, la méthodologie relationnelle s'appuie fortement sur la définition de cette matrice

globale de similarités.

De façon analogue nous pouvons définir la matrice relationnelle de Condorcet de la relation complémentaire $\not\mathcal{R}$ comme

$$\bar{c}_{ii'} = \begin{cases} 1 & \text{si } i \not\mathcal{R} i' \forall (i,i') \in V \times V, \\ 0 & \text{sinon.} \end{cases} \tag{2.2}$$

Ce qui caractérise une relation binaire sont les propriétés qu'elle vérifie, exprimables de façon générale soit par des expressions logiques soit par des équations mathématiques. En fonction des propriétés vérifiées par une relation $\mathcal{R}$, il existe des relations dites *typées* (c'est-à-dire vérifiant plusieurs propriétés élémentaires). Le Tableau 2.1 montre une liste non exhaustive de propriétés qu'une relation binaire $\mathcal{R}$ pourrait vérifier.

Propriété	Définition logique	Expression mathématique
Réflexivité	$i\mathcal{R}i \qquad \forall i \in V$	$c_{ii} = 1$
Symétrie	$i\mathcal{R}i' \Rightarrow i'\mathcal{R}i \qquad \forall (i,i') \in V \times V$	$c_{ii'} = c_{i'i}$
Asymétrie	$i\mathcal{R}i' \Rightarrow \neg(i'\mathcal{R}i) \qquad \forall (i,i') \in V \times V$	$c_{ii'} + c_{i'i} \leq 1$
Totalité	$i\mathcal{R}i' \vee i'\mathcal{R}i \qquad \forall (i,i') \in V \times V, i \neq i'$	$c_{ii'} + c_{i'i} \geq 1$
Transitivité	$i\mathcal{R}i' \wedge i'\mathcal{R}i'' \Rightarrow i\mathcal{R}i'' \qquad \forall (i,i',i'')$	$c_{ii'} + c_{i'i''} - c_{ii''} \leq 1$

TABLE 2.1 – Principales propriétés vérifiées par une relation binaire.

Voici quelques exemples de relations typées et leurs propriétés respectives :
- Une Relation de *Pré-ordre* est *Réflexive* et *Transitive*.
- Une Relation de *Pré-ordre Total* est *Réflexive*, *Transitive* et *Complète*.
- Une Relation *d'Ordre Total* est *Transitive*, *Asymétrique* et *Complète*.
- Une Relation d'*équivalence* est une relation *réflexive*, *symétrique* et *transitive*[2].

Le domaine d'application de l'Analyse Relationnelle étant assez large nous ne décrirons dans cette section que les deux thématiques les plus répandues et connues : "la recherche de Consensus Électoraux" et "la recherche de Clustering Consensus". Deux thématiques qui, traduites au langage relationnel consistent à rechercher d'une relation d'ordre total et à rechercher d'une relation d'équivalence respectivement. Pour finir ce chapitre, nous décrirons brièvement le premier axe de recherche. Ensuite, nous traiterons de façon assez détaillée le deuxième axe de recherche, car cela correspond au sujet principal de cette thèse. En particulier, modulariser un graphe revient à définir une relation d'équivalence sur son ensemble de sommets. C'est principalement à ce type de relation (Relation d'équivalence) que nous aurons affaire, via un typage relationnel que nous allons analyser en détail par la suite.

2. La condition de transitivité s'interprète comme suit : "si i est dans la même classe d'équivalence que i' et i' est dans la même classe que i'', alors forcement i et i'' sont dans la même classe".

La condition *duale* sur la matrice de Condorcet $\bar{\mathbf{C}}$, est l'*inégalité triangulaire* :

$$\bar{c}_{ii''} \leq \bar{c}_{ii'} + \bar{c}_{i'i''}$$

La relation d'inégalité triangulaire est plutôt liée à des approches "métriques" de distances, alors que la condition duale de transitivité générale est plutôt utilisée sous l'angle relationnel logique.

2.3.2 Recherche de Consensus Électoraux : relation d'ordre

Cet axe de recherche couvre la thématique de la *"Recherche de Consensus Électoraux"*, liée elle-même à *"l'Analyse des Préférences"* et à *"la Théorie du Choix Social"* lancée par Kenneth Arrow avec son célèbre *théorème d'impossibilité*. Toutes ces approches étant reliées à la recherche d'ordre ou d'électeur Médian.

La formulation mathématique de ce problème est la suivante : soient M juges ou votants ($k \in \{1, 2, 3, ..., M\}$) ayant donné leurs préférences sous forme de classements totaux sur N candidats. Le classement de chaque votant définit une relation d'ordre sur l'ensemble des objets (candidats). Une relation d'ordre peut être représentée de façon *unique* grâce à un tableau (ou matrice) d'ordre N, noté $\mathbf{C}^k$ (pour l'électeur k) et appelé *Matrice Relationnelle de Condorcet*. Le terme général de cette matrice est défini de la façon suivante

$$c_{ii'}^k = \begin{cases} 1 & \text{si le candidat } i \text{ est classé avant } i' \text{ par le votant } k, \\ 0 & \text{sinon.} \end{cases} \tag{2.3}$$

La recherche de l'électeur médian consiste à trouver une relation d'ordre médian, qui est représentée à son tour par une matrice $\mathbf{X}$ carrée d'ordre N et dont le terme général est défini de façon analogue aux variables V^k par

$$x_{ii'} = \begin{cases} 1 & \text{si le candidat } i \text{ est classé avant } i', \\ 0 & \text{sinon.} \end{cases} \tag{2.4}$$

En 1959 le mathématicien américain J. Kemeny a posé ce problème de recherche de l'Ordre Médian dans Kemeny [1959], défini comme la recherche de la relation d'ordre total minimisant la norme L^1 de la fonctionnelle suivante (en notations relationnelles)

$$F_K(X) = \sum_{k=1}^{M} |C^k - X|. \tag{2.5}$$

Du fait que les termes généraux des matrices C^k et de X sont binaires (valeurs possibles : 0 ou 1) la norme L^1 est équivalente à la norme L^2 et la Recherche de l'Ordre Médian est équivalente à la Recherche de l'Ordre Moyen selon la transformation suivante

$$F_K(X) = \sum_{k=1}^{M} \left(\sum_{i=1}^{N} \sum_{i'=1}^{N} (c_{ii'} - x_{ii'})^2 \right) = \frac{1}{M} \sum_{i=1}^{N} \sum_{i'=1}^{N} (M - 2c_{ii'}) x_{ii'} + K,$$

où $c_{ii'} = \sum_{k=1}^{M} c_{ii'}^k$, le terme général du *tableau relationnel de Condorcet* ou *Tableau de comparaisons par paires*.

La modélisation proposée par Kemeny est connue également sous le nom de the *Acyclic Subgraph Problem* chez les combinatoristes et les spécialistes de la théorie des Graphes. En 1979 F. Marcotorchino et P. Michaud en rapprochant l'écriture littérale du célèbre article d'Antoine Caritat, Marquis de Condorcet : *"Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix"* (voir Condorcet [1785]) ont montré que le Problème de Kemeny n'était rien d'autre que la fonctionnelle dite du *"Support de Voix de Condorcet"* (voir Marcotorchino and Michaud [1979]), qui une fois traduit dans des notations relationnelles s'écrit comme le problème de la Maximisation de la fonctionnelle suivante

$$F_C(X) = \sum_{k=1}^{M} \left(\sum_{i=1}^{N} \sum_{i'=1}^{N} (c_{ii'}^k x_{ii'} + \bar{x}_{ii'}^k \bar{x}_{ii'}) \right), \tag{2.6}$$

où $\bar{x}_{ii'} = 1 - x_{ii'}$ et $\bar{c}_{ii'} = 1 - c_{ii'}$ sont les relations inverses de X et C^k respectivement. Avec ces notations l'expression (2.6) peut alors se écrire de la façon suivante

$$F_C(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} (c_{ii'} x_{ii'} + \bar{c}_{ii'}(1 - x_{ii'})) = \sum_{i=1}^{N} \sum_{i'=1}^{N} (c_{ii'} - \bar{c}_{ii'}) x_{ii'} + K = \sum_{i=1}^{N} \sum_{i'=1}^{N} (2c_{ii'} - M) x_{ii'} + K. \tag{2.7}$$

Dans la suite on dénotera K une fonction des données d'origine, donc une constante qui n'affectera pas las solution optimale.

Ici $K = \sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{c}_{ii'}$, où la dernière égalité est obtenue en supposant qu'il n'y a pas de données manquantes : $\bar{c}_{ii'} + c_{ii'} = M \quad i, i'$. A partir de cette dernière expression on remarque facilement que maximiser le critère de Condorcet F_C revient à minimiser le critère de Kemeny F_K (expression (2.5)).

La solution du problème de Condorcet (2.7) sans contraintes sur $\mathbf{X}$ est trivialement de classer i avant i' si une majorité de votants préfèrent i à i' (majorité condorcéenne par paires), soit $x_{ii'} = 1$ si $2cii' – M > 0$ et $x_{ii'} = 0$ sinon. Cependant la solution ainsi obtenue n'est pas transitive et on aboutit au fameux *effet Condorcet* ou *Paradoxe de Condorcet*.

Dès lors pour pallier cet inconvénient, il faut OBLIGATOIREMENT rajouter des contraintes qui forcent à $\mathbf{X}$ à représenter une Relation d'ORDRE TOTAL c'est-à-dire Réflexive, Transitive, Totale et antisymétrique. Ce qui se traduit mathématiquement par les contraintes linéaires :

$$\begin{array}{rcl} x_{ii'} \in \{0,1\} & & \text{Binarité} \\ x_{ii} = 1 & \forall i & \text{Réflexivité} \\ x_{ii'} - x_{i'i} = 1 & \forall (i,i') & \text{Antsymétrie et Totalité} \\ x_{ii'} + x_{i'i''} - x_{ii''} \leq 1 & \forall (i,i',i'') & \text{Transitivité.} \end{array} \tag{2.8}$$

La Maximisation du critère de Condorcet (2.6) sous les contraintes linéaires (2.8) est un problème de Programmation Linéaire en Nombres entiers (0-1). Sa solution est appelée désormais *Solution de la Recherche de l'Ordre Médian de Condorcet sous contraintes*.

Cette solution vérifie les Axiomes du Deuxième Théorème d'Arrow (version modifiée du premier Théorème, proposée en 1986 dans Arrow and Raynaud [1986]) qui consiste en une Relaxation de la condition n°3 [3] (sur 5) de son Théorème sur l'impossibilité d'un classement collectif consensus (voir Arrow [1963]) et qui lui a valu son prix Nobel en 1972.

La solution optimale du problème "Médian Condorcéen" sous contraintes a été trouvée pour la première fois de façon concrète et systématique par une approche Programmation

3. Condition dite de l'indépendance par paires vis-à-vis des alternatives extérieures : le choix entre deux alternatives ne dépend que des préférences individuelles entre ces alternatives.

Linéaire en utilisant des procédures *dual du dual*, ceci est présenté dans le livre de Marcotorchino and Michaud [1979]. Aujourd'hui il existe des heuristiques ad-hoc pour résoudre ce problème y compris dans le logiciel Open source R.

2.3.3 La recherche de Clustering Consensus : relation d'équivalence

Le deuxième axe important de recherche en Analyse Relationnelle est le "Clustering Condorcéen" ou "Problème des Partitions Médianes" ou "Problème des Partitions Centrales", ce problème revient à la mode de nos jours grâce principalement au développement récent et rapide des réseaux sociaux.

Cette fois-ci au lieu d'avoir à manipuler des classements ou préférences d'électeurs nous disposons de M variables catégorielles $J = \{V^1, V^2, ..., V^M\}$ décrivant N objets, aussi appelés éléments [4].

Chaque variable k possède p_k modalités, et $P = \sum_{k=1}^{M} p_k$ est le nombre total de modalités. Les objets peuvent être, par exemple, un ensemble de voitures caractérisées par certains attributs, comme la couleur {rouge,blanc,bleu} ; la marque {Renault, Toyota, BMW, etc...}.

Comme chaque objet peut appartenir à une et une seule catégorie de chaque descripteur, chaque variable définit ainsi une relation d'équivalence sur l'ensemble des objets. Il existe 3 façons possibles de représenter cette variable, en fait 3 codages différents :

1. **Codage linéaire** : La variable est représentée comme un vecteur de $\mathbb{R}^N$ dont l'élément i décrit la catégorie prise par l'objet i.
2. **Codage disjonctif complet** : La variable est représentée sous forme de matrice, notée $\mathbf{K}$, de taille $(N \times p)$ dont l'élément k_{ij} est une variable de présence-absence et il est donné par

$$k_{ij} = \begin{cases} 1 & \text{si l'objet } i \text{ possède la modalité } j, \\ 0 & \text{sinon.} \end{cases} \tag{2.9}$$

3. **Codage Relationnel** : étant donné que la variable à représenter est une relation d'équivalence, celle-ci peut être représentée par la matrice relationnelle de Condorcet $\mathbf{C}^k$ associée. Ainsi, le terme général de $c^k_{ii'}$ du tableau de Condorcet vaut dans ce cas-là

$$c^k_{ii'} = \begin{cases} 1 & \text{si } i \text{ et } i' \text{ appartiennent à la même classe de la variable } k, \\ 0 & \text{sinon.} \end{cases} \tag{2.10}$$

A titre d'exemple nous illustrons sur la FIG.2.1 les 3 codages possibles d'une variable, par exemple la *nationalité*, qui décrit ici 5 individus : A, B, C, D, E. Dans cet exemple, la variable *nationalité* possède 3 modalités : russe (RU), italienne (IT) et française (FR).

4. N pouvant atteindre des valeurs égales à plusieurs millions dans les problématiques réelles : "CRM (Customer Relationship Management" et Marketing bancaires, par exemple.

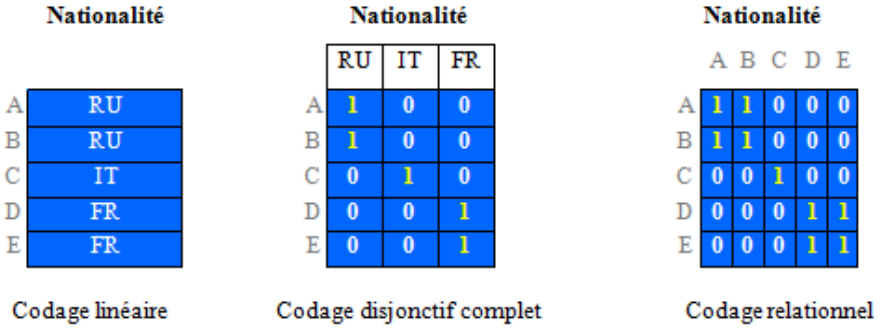

FIGURE 2.1 – 3 codages possibles d'une variable catégorielle (relation d'équivalence).

Ces trois notations ou codages contiennent pratiquement la même information. Quoiqu'en ce qui concerne le troisième codage, il semblerait exister une perte de précision sur le *label* de la classe d'appartenance de chaque individu. Il faut nuancer cette affirmation car la redondance d'information dans le cas relationnel permet des désambigüisations, ce qui fait qu'elle possède au final la même information que dans les cas précédents. Le troisième codage permet entre autres :

- de travailler sur l'espace des individus, ce qui autorise l'addition de plusieurs tableaux de variables représentant chacun une relation d'équivalence. Ce qui n'est pas possible avec la deuxième notation (tableau disjonctif complet), car le nombre de colonnes du tableau dépend du nombre de modalités de la variable qu'il représente ; et bien entendu, avec la première notation (codage linéaire) car l'addition dans ce cas-là n'a pas de sens.
- de ne pas être dépendant du nombre de modalités de la variable qui se trouve implicitement contenu dans le tableau. Ainsi lorsque nous voulons définir une relation d'équivalence optimale inconnue X s'approchant au mieux des données d'entrée, nous ne sommes pas obligés de fixer le nombre de modalités (classes) a priori.
- de tenir compte de la multi-appartenance d'un objet à plusieurs classes.

Si l'on garde le même formalisme que celui introduit lors de la définition du Critère de Condorcet (2.6) et que l'on remplace la Relation Binaire $\mathbf{X}$ d'ordre Total du cas précédent par une relation d'équivalence on étend le modèle Condorcéen de Recherche de(s) l'électeur(s) "Médian(s)" à celui de la recherche d'une (des) Partition(s) Centrale(s) ou "Médiane(s)".

Ce problème de Recherche des Partitions Centrales a été posé la première fois par Régnier [1966], suivi ensuite par Mirkin and Cherny [1970]. L'identification de la totale similarité du Problème de la recherche d'Electeurs Concensus ou d'Electeurs Médians et de celui de la Recherche de Partitions Centrales ou Médianes a été réalisé par Marcotorchino and Michaud [1979]. L'inconnue $\mathbf{X}$ est cette fois définie par

$$x_{ii'} = \begin{cases} 1 & \text{si } i \text{ et } i' \text{ appartiennent à la même classe,} \\ 0 & \text{sinon.} \end{cases} \tag{2.11}$$

Ainsi, on aboutit de facto au Modèle de Maximisation du Critère de Condorcet (équation (2.6)), mais cette fois sous des contraintes linéaires sur $\mathbf{X}$ caractérisant une relation d'équivalence à savoir :

$$\begin{aligned} x_{ii'} \in \{0,1\} \quad & \quad \text{Binarité} \\ x_{ii} = 1 \quad & \forall i \quad \text{Réflexivité} \\ x_{ii'} - x_{i'i} = 0 \quad & \forall (i,i') \quad \text{Symétrie} \\ x_{ii'} + x_{i'i''} - x_{ii''} \leq 1 \quad & \forall (i,i',i'') \quad \text{Transitivité.} \end{aligned} \tag{2.12}$$

Ce modèle revient à trouver la partition $\mathbf{X}$ située en moyenne à Distance de la "Différence Symétrique" minimale d'un ensemble de M variables qualitatives C^k. Nous verrons dans le chapitre 5 que dans le cas où l'on cherche juste la Partition approximant au mieux une relation binaire symétrique (c'est-à-dire un Graphe non orienté et non pondéré), donnée a priori, on tombe sur le Problème de Zahn (voir Zahn [1964]).

Un avantage très important de l'écriture Relationnelle du critère de Condorcet (2.6) est qu'il ne dépend pas du nombre de classes de la partition optimale $\mathbf{X}$, donc, on n'est pas obligé de le fixer a priori contrairement aux méthodes K-Means.

La relation inverse de $\mathbf{X}$, soit $\bar{\mathbf{X}}$, peut s'interpréter comme

$$\bar{x}_{ii'} = \begin{cases} 1 & \text{si } i \text{ et } i' \text{ n'appartiennent pas à la même classe,} \\ 0 & \text{sinon.} \end{cases} \tag{2.13}$$

Il existe aussi une forme duale au critère de Condorcet [5], qui devient alors une fonction à minimiser et non pas à maximiser

$$F_C(\bar{X}) = \sum_{k=1}^{M} \left(\sum_{i=1}^{N} \sum_{i'=1}^{N} (\bar{c}^k_{ii'} x_{ii'} + c^k_{ii'} \bar{x}_{ii'}) \right). \tag{2.14}$$

D'autre part l'écriture (2.7) reste valable dans le cas de recherche d'une partition.

Ce problème de recherche de la partition médiane a été résolu de façon exacte par la technique du «dual du dual» avec relaxation des contraintes de binarité pure, chez IBM Research en 1980 (avec une taille maximum où $N = 150$ et $M = 25$), un jeux de données relativement petit de nos jours.

Matrice de Condorcet Pondérée Nous définissons maintenant une matrice qui joue un rôle de médiation entre l'Analyse Factorielle et l'ARM, il s'agit du tableau de Condorcet pondéré, noté $\hat{\mathbf{C}}$. La matrice Relationnelle pondérée d'une relation binaire $\mathcal{R}$ est un tableau $N \times N$ dont l'élément général est défini comme

5. Le problème de "Clustering Consensus" fut récemment traité à nouveau par Gionis et al. [2007] sous le nom de "Clustering aggregation". La fonction à optimiser proposée par les auteurs correspond à la version duale du critère de Condorcet.

$$\hat{c}_{ii'} = \frac{2c_{ii'}}{c_{i.} + c_{.i'}}, \forall i, i' \in V \times V \qquad \text{avec} \qquad c_{i.} = \sum_{i'=1}^{N} c_{ii'}, \forall i \in V. \tag{2.15}$$

Dans le cas où $\mathcal{R}$ représente une relation d'équivalence, des simplifications importantes se produisent dans la formule (2.15). En effet, si $c_{ii'} = 1$ (i et i' sont dans la même classe), alors $c_{i.} = c_{.i'}$ et cette dernière quantité représente l'effectif de la classe contenant i. Donc l'expression (2.15) se simplifie en

$$\hat{c}_{ii'} = \frac{c_{ii'}}{c_{i.}}. \tag{2.16}$$

La matrice $\hat{\mathbf{C}}$ possède des propriétés importantes : elle est symétrique, bi-stochastique, idempotente, à blocs diagonalisation constante et elle possède aussi la propriété de fuzzyness [6].

Le nombre de modalités est contenue implicitement dans le tableau $\mathbf{C}$ et il peut être obtenu de deux façons :

1. En faisant la somme des termes diagonaux de $\hat{\mathbf{C}}$ on obtient

$$\sum_{i=1}^{N} \hat{c}_{ii} = \sum_{i=1}^{N} \frac{1}{c_{i.}} = p. \tag{2.17}$$

2. En faisant la somme des carrés des termes de $\hat{\mathbf{C}}$ on obtient

$$\sum_{i=1}^{N} \sum_{i'=1}^{N} \hat{c}_{ii'}^2 = \sum_{i=1}^{N} \sum_{i'=1}^{N} \frac{c_{ii'}}{c_{i.}c_{.i'}} = p, \tag{2.18}$$

où p est le nombre de classes d'équivalence de la partition $\mathbf{C}$.

Un dernier résultat important concernant la matrice de "Condorcet Pondérée" est sa complémentaire $\bar{\hat{C}}$, définie par

$$\bar{\hat{c}}_{ii'} = \frac{\hat{c}_{ii} + \hat{c}_{i'i'}}{2} - \hat{c}_{ii'}. \tag{2.19}$$

Cette matrice peut s'interpréter comme un écart à la situation d'*auto similarité maximale* [7].

6. Propriété d'appartenance à l'intervalle $[0, 1]$

7. Cette matrice n'est rien d'autre, à une constante près, que la *distance du* χ^2 *(chi2)*, introduite par J.P. Benzécri en "Analyse Factorielle des Correspondances" en 1973, (voir Benzécri [1973a] et Benzécri [1973b]). A savoir

$$d_{\chi^2}(ii') = \frac{2N}{M} \bar{\hat{c}}_{ii'}.$$

Cette expression qui lie "Analyse Relationnelle" et "Analyse Factorielle" est exploitée, étendue et discutée en détail dans les articles Marcotorchino [1989] et Marcotorchino [1991].

Conclusion : Un graphe étant un objet mathématique servant à modéliser les relations binaires entre objets d'un même ensemble, le problème de modularisation de graphes peut être modélisé à l'aide du formalisme de l'Analyse Relationnelle Mathématique. Ce problème revient à chercher une partition P sur l'ensemble de sommets du graphe V, qui n'est autre qu'une relation d'équivalence, en ayant comme donnée d'entrée sa matrice d'adjacence. Ainsi, la Recherche de Communautés dans les réseaux sociaux, relève également de l'analyse Condorcéenne Relationnelle.

Chapitre 3

Communauté et modularisation

3.1 Introduction

Les réseaux réels ne sont pas des graphes aléatoires car ils présentent des hétérogénéités importantes. Ils recèlent souvent un ordre et une organisation implicites. En plus, la distribution d'arêtes est localement hétérogène. Étant très élevée à l'intérieur de certains groupes de sommets et faible à l'extérieur ou entre ces groupes.

Le problème de partitionnement des graphes n'est pas un sujet récent. Il a déjà été abordé dans des disciplines les plus diverses. Ainsi en sociométrie, la première analyse de la structure d'un réseau social a été faite par Weis and Jacobson [1955]. Ils ont étudié les principaux concepts caractérisant la structure d'une organisation. Leur but était de définir des groupes de travail dans une Agence publique (*Government Agency*). Chaque groupe était séparé du reste du réseau en supprimant les membres travaillant avec fonctionnaires des différents groupes. Ces derniers étaient considérés comme des connecteurs entre les différents groupes. Pour les auteurs, chaque groupe jouait un rôle important dans la structure de l'organisation.

Dans le livre *Quantitative methods in politics* de Rice [1928], l'auteur cherchait à classifier la population selon la similarité de leur préférence pour un parti politique. Nous allons voir dans la section suivante qu'il existe un lien étroit entre la méthode de détection de communautés et la définition de communauté.

3.2 Définition de la notion de communauté

La formulation mathématique du problème de détection des communautés repose sur le partitionnement (ou clustering) de graphes. Ce dernier n'est pas bien défini car il n'existe pas une définition universelle de *communauté*. En effet, différents critères de clustering de graphes ont été proposés au fil du temps pour modéliser des phénomènes issus de domaines différents, chacun reposant implicitement sur une définition propre, légèrement différente, à chaque fois, de la notion de *communauté*. De plus, le problème de partitionnement d'un graphe possède une application beaucoup plus large que la recherche de communautés dans les réseaux sociaux. Car un graphe sert à modéliser non seulement un réseau social mais également tout autre réseau : biologique, informatique, etc...Dans un graphe un groupe de sommets peut être, suivant le contexte ou le domaine d'application, appelé *communauté*,

mais aussi *classe*, *cluster* ou *module*[1]. Voici quelques définitions de communautés trouvées dans la littérature :

- Une définition très citée dans la littérature est celle donnée par Radicchi et al. [2004]. Les auteurs formulent deux définitions de *communauté* :

  1. Définition de la notion de *communauté* dans le sens fort : un sous-graphe est une communauté au sens fort si chacun de ses sommets a plus de connexions au sein de la communauté qu'avec le reste du réseau.

  2. Définition de la notion de *communauté* dans le sens faible : un sous-graphe est une communauté au sens faible si pour toute la communauté la somme de poids d'arêtes intra-communauté multipliée par deux est plus grande que la somme des poids d'arêtes inter-communauté ou vers le reste du réseau.

     Il est clair qu'une communauté au sens fort est également une communauté au sens faible, alors que l'inverse n'est pas forcement vrai.

- Groupes de sommets densément connectés, avec quelques rares connexions entre groupes.
- Groupes de sommets qui interagissent plus entre eux qu'avec le reste du réseau.
- Groupes de sommets qui partagent une caractéristique commune ou qui poursuivent un but commun.
- Ensemble de sommets qui communiquent plus entre eux qu'avec le reste du réseau.
- Ensemble de sommets homogènes entre eux mais hétérogènes avec les sommets restants.
- etc...

Ces définitions, bien que différentes, possèdent un dénominateur commun : une forte densité d'arêtes intra-classe et une faible densité d'arêtes inter-classes (quoique l'un implique l'autre). Donc, l'identification de structure de classe n'est possible que si le graphe est dense, i.e. s'il existe une quantité importante d'arêtes par rapport au nombre de sommets. Ainsi dans un *arbre*[2] ou un graphe en forme de grille (*lattice ou grid graphe en anglais*)la notion de module n'existe pas. La figure 3.1 montre des exemples de graphes où la définition de communauté n'a pas de sens, bien évidement il n'existe pas de groupes de sommets qui soient remarquablement plus connectés que d'autres.

En contrepartie, le graphe de la figure 3.2[3] illustre clairement la notion de communautés. On voit bien que les 3 groupes retrouvés sont fortement liés et qu'il existe peu de connections entre eux.

1. Dans le présent document nous utiliserons ces termes de façon indistincte comme s'il s'agissait de synonymes
2. Nous rappelons qu'un arbre est un graphe minimalement connecté, i.e. $M = N - 1$
3. La figure a été extraite de Girvan and Newman [2002]

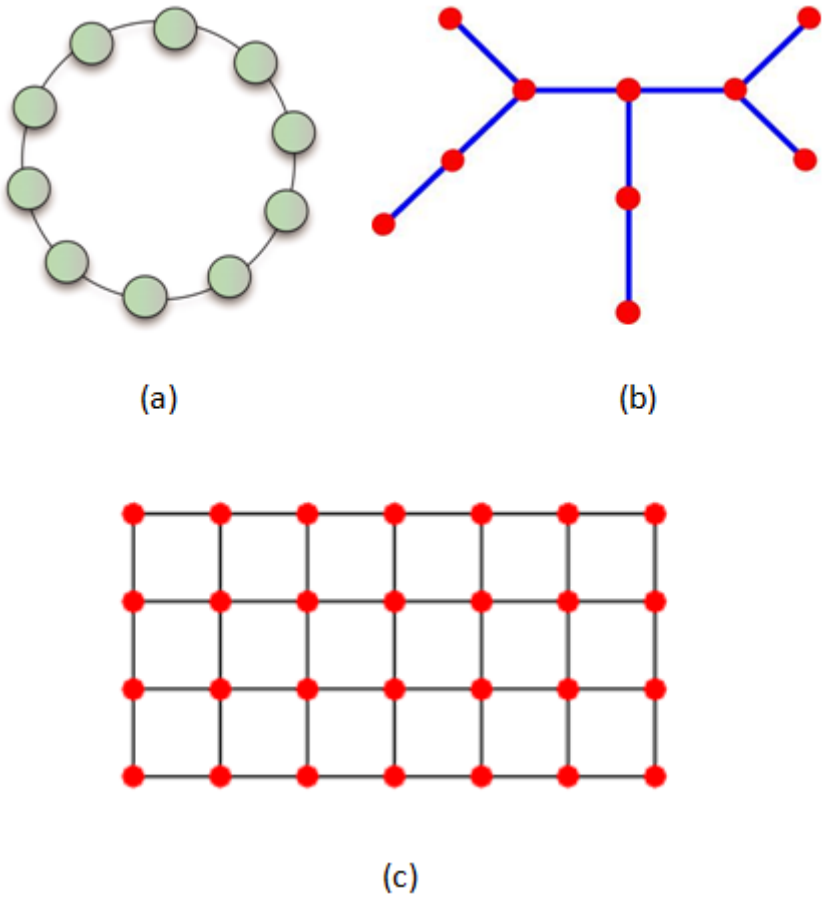

Figure 3.1 – (a) Anneau treillis graphe (*Ring lattice graph*), (b)Arbre, (c)Graphe en forme de grille *(grid graph)*

3.3 Détection de communautés et modularisation

Lorsque la taille du graphe (le nombre de sommets) devient importante il devient difficile d'étudier sa structure globalement. Donc, décomposer les sommets en sous-graphes devient nécessaire afin de comprendre la structure réelle du graphe. Ces sous-ensembles de sommets s'appellent des "communautés". Modulariser ou partitionner un graphe signifie chercher les communautés qui le composent.

Quelque soit la définition de communauté la tâche est la même : "Pour un graphe donné, nous souhaitons le décomposer en sous-graphes de telle sorte que les sommets de chaque sous-graphe aient plus à voir entre eux qu'avec les sommets du reste du réseau".

Modulariser un graphe revient donc à trouver une partition P de ses sommets de telle sorte que les éléments de P soient le plus possible "densément connectés". Les parties de P ainsi obtenues sont les communautés cherchées. Ainsi le graphe est partitionné en κ classes.

La détection de communautés a de multiples applications :

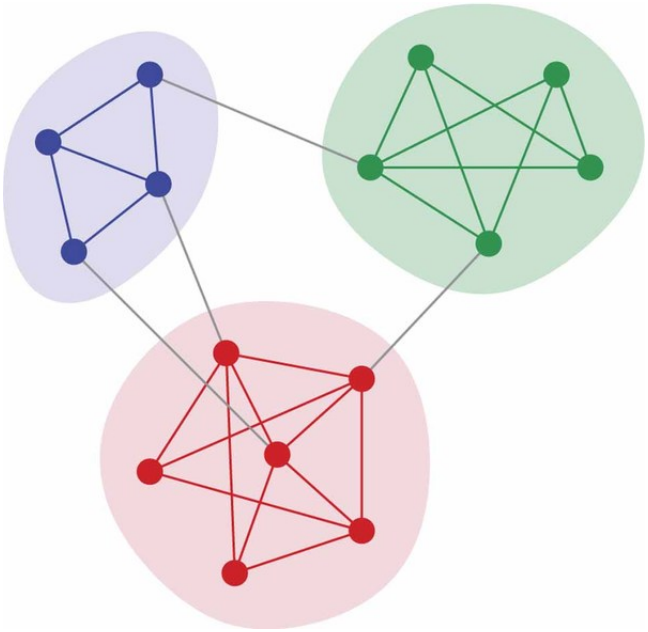

FIGURE 3.2 – Exemple de graphe qui possède une structure de communautés.

- Fréquemment les sommets densément connectés partagent une remarquable propriété commune. Par exemple, dans le cas de réseaux sociaux cette propriété peut être un intérêt commun ; dans le cas des pages web, les communautés partagent parfois une thématique commune. Donc, en analysant les objets qui composent une même communauté on peut leurs attribuer des propriétés intrinsèques et caractéristiques de la communauté elle -même.
- Étant donné que les objets qui appartiennent à une même communauté sont plus homogènes que le réseau dans sa globalité. Étudier chaque communauté séparément peut nous permettre de repérer des caractéristiques qui ne sont pas facilement repérables si l'on étudie le réseau au niveau global.
- Chaque communauté peut être condensée (résumée) en un seul "méta-sommet" (meilleur représentant de la communauté [4]), permettant une analyse du réseau à un niveau plus grossier, et une focalisation sur la structure de niveau supérieur. Cette approche peut servir pour visualiser un réseau de façon simplifiée, en perdant le minimum d'information.
- Les sites internet dédiés à la vente de produits de consommation cherchent à trouver chez leurs acheteurs des groupes de clients qui ont des intérêts et des désirs d'achats semblables et sont géographiquement proches les uns des autres. Ainsi chaque groupe recevrait une attention et un service différent de la part du site internet.
- Si le graphe est grand, chercher à le comprendre dans sa globalité devient très complexe, voire impossible. Il est, alors, indispensable de pouvoir le "partitionner" en sous-ensembles gérables et faciles à comprendre.
- Découvrir les modules d'un graphe permet d'identifier la fonction de chaque sommet dans le module. Ainsi un sommet qui possède une position centrale (i.e. il est adjacent

4. Ceci renvoie aux notions de centralité (de différentes sortes) que nous avons explicitées précédemment.

à plusieurs sommets dans la classe) peut avoir une fonction importante de contrôle de la stabilité au sein du groupe ; tandis que les sommets situés à la frontière de la classe jouent un rôle important de médiation et d'échange avec les autres communautés (voir Barabási and Frangos [2002], Barabási [2012] et Viennet [2009]). Par exemple dans des réseaux comme FaceBook ou Internet, on trouve des sommets, appelés *hubs*, qui ont beaucoup de liens et en même temps certains sommets qui ont très peu de liens. En fait, la plupart des réseaux réels sont très hétérogènes [5]..

– Dans un réseau informatique, la détection de modules peut servir à connaître quelle est la meilleure façon de répartir les tâches associées aux processeurs afin de minimiser les communications entre eux et permettre une haute performance de calcul.

Cette procédure de modularisation de graphes nécessite l'élaboration d'un critère pertinent qui puisse juger la qualité du partitionnement. Comme mentionné précédemment plusieurs critères ont été proposés. La plupart de ces critères prennent en compte la valeur de 4 quantités importantes :

– Nombre d'arêtes intra-classe ; la plupart de critères qui font intervenir cette quantité sont des critères à maximiser. Il s'agit de la quantité représentant les accords positifs entre les données et la variable cherchée $\mathbf{X}$.

– Nombre d'arêtes intra-classe manquantes pour en faire un graphe complet : cette quantité élémentaire compte les arêtes manquantes à l'intérieur de la classe pour que celle-ci devienne une clique ou un graphe complet.

– Nombre d'arêtes inter-classes : quantité nommée dans la littérature anglophone *cut*, coupure, car il s'agit du nombre d'arêtes qu'il faut couper pour obtenir les classes séparées. La plupart des critères faisant intervenir le *cut* s'expriment comme une fonction à minimiser, par exemple le "ratio cut" de Wei and Cheng [1989] et Wei and Cheng [1991], ou bien la *coupe normalisée* ("normalized cut") de Shi and Malik [2000].

– Nombre d'arêtes complètes inter-classes manquantes. Cette quantité élémentaire est apparentée à la quantité précédente, et représente les accords négatifs entre les données et la Relation cherchée inverse $\bar{\mathbf{X}}$.

Le tableau suivant montre l'écriture relationnelle de ces 4 quantités. Chacune joue un rôle important dans la formulation des critères de partitionnement de graphes :

5. Les réseaux de ce type sont aussi connus sous le nom de réseaux "sans échelle"

Quantité	**Écriture Relationnelle**
Arêtes intra-classe	$\frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}x_{ii'}$
Arêtes intra-classe manquantes	$\frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{a}_{ii'}x_{ii'}$
Arêtes inter-classes : *Cut*	$\frac{1}{2}\sum_{i}^{N}\sum_{i'}^{N} a_{ii'}\bar{x}_{ii'}$
Accords négatifs	$\frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{a}_{ii'}\bar{x}_{ii'}$

Certains critères s'écrivent comme l'addition et/ou la soustraction de certaines parmi les quantités présentées ci-dessus. D'autres critères maximisent ou minimisent des ratios de ces différentes quantités.

Dans la suite nous étudierons différents critères de modularisation. Sauf indication contraire nous traiterons les graphes non-orientés, non pondérés et non réflexifs. On trouvera un résumé sur la modularisation des graphes orientés dans Malliaros and Vazirgiannis [2013].

Chapitre 4

Propriétés des Critères de modularisation

4.1 Propriétés vérifiées par des Critères de Partitionnement

La liste de propriétés présentées par la suite n'est pas une liste exhaustive. Il s'agit des propriétés fondamentales et générales vérifiables par certains critères de partitionnement ou de classement. Ces propriétés sont formelles ou structurelles et induisent (tout au moins pour certaines d'entre elles) des finalités plus ou moins intuitives, qu'il s'agit de bien interpréter et comprendre. En voici quelques unes parmi les plus incontournables :

4.1.1 Propriété de Linéarité

Un problème de classification est **linéaire** par rapport à la variable inconnue $\mathbf{X}$, si le critère associé (≪fonction économique ≫ou fonction qualité) s'écrit sous la forme

$$F(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \phi(a_{ii'}) x_{ii'} + K, \tag{4.1}$$

où la fonction $\phi(a_{ii'})$ dépend uniquement des données d'entrée (dans notre cas, de la matrice d'adjacence du graphe).

4.1.2 Propriété de Séparabilité

Un critère de modularisation ou de classification est séparable au sens : "variables-données" si le critère associé peut s'écrire sous la forme :

$$F(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \phi(a_{ii'}) \psi(x_{ii'}) + K, \tag{4.2}$$

où les quantités $\phi(a_{ii'})$ ne dépendent que des données initiales du problème considéré et où $\psi(x_{ii'})$ est une fonction de la relation inconnue X comme par exemple :

- Si $\psi(x_{ii'}) = x_{ii'}$ le critère possède aussi la propriété de linéarité.
- Si $\psi(x)$ est, par exemple, une fonction de la relation $\hat{\mathbf{X}}$ pondérée :

$$\psi(x_{ii'}) = \hat{x}_{ii'} = \frac{x_{ii'}}{x_{i.}} = \begin{cases} \frac{1}{x_{i.}} & \text{si } x_{ii'} = 1, \\ 0 & \text{sinon.} \end{cases} \tag{4.3}$$

Dans ce cas le critère n'est plus à proprement parler "linéaire" en $x_{ii'}$, mais il est néanmoins séparable.

La propriété de "séparabilité " est très importante du fait qu'elle permettra de simplifier beaucoup d'expressions de critères qui seront présentées par la suite.

4.1.3 Propriété d'Equilibre Général

Un problème de classification est **équilibré** par rapport à la variable inconnue $\mathbf{X}$, si le critère associé peut s'écrire sous la forme :

$$F(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \phi(a_{ii'})\psi(x_{ii'}) + \sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{\phi}(a_{ii'})\bar{\psi}(\bar{x}_{ii'}) + K, \tag{4.4}$$

où :

- $\phi(.)$ et $\bar{\phi}(.)$ sont des fonctions dépendantes des données d'entrée et non pas de l'inconnue $\mathbf{X}$. Il s'agit de fonctions positives $\phi(a_{ii'}) \geq 0$ et $\bar{\phi}(a_{ii'}) \geq 0$ non toutes nulles, ce qui implique $\sum_{i=1}^{N}\sum_{i'=1}^{N} \phi_{ii} > 0$ et $\sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{\phi}_{ii} > 0$.

- $\psi(.)$ et $\bar{\psi}(.)$ sont des fonctions dépendantes et croissantes des inconnues $\mathbf{X}$ et $\bar{\mathbf{X}}$ respectivement.

Si le problème de classification correspond à une maximisation, nous aurons dans quelques cas particuliers $\phi(a_{ii'})$ est croissante en $a_{ii'}$ et/ou $\bar{\phi}(a_{ii'})$ est croissante en $\bar{a}_{ii'}$. De façon analogue, si le critère de classification est à minimiser dans certains cas $\phi(a_{ii'})$ peut être décroissante en $a_{ii'}$ et/ou $\bar{\phi}(a_{ii'})$ peut être décroissante en $\bar{a}_{ii'}$.

La notion d'équilibre vient du fait que l'équation (4.4) est composée de façon symètrique :

1. D'accords positifs (ou attractions positives), la quantité : $\sum_{i=1}^{N}\sum_{i'=1}^{N} \phi(a_{ii'})\psi(x_{ii'})$.
2. D'accords négatifs (ou attractions négatives), la quantité : $\sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{\phi}(\bar{a}_{ii'})\bar{\psi}(\bar{x}_{ii'})$.

4.1.4 Propriété d'Equilibre Général pour les critères linéaires

Un critère de modularisation est équilibré linéairement par rapport à la variable relationnelle $\mathbf{X}$ dans le cas particulier où $\psi(x_{ii'}) = x_{ii'}$ et $\bar{\psi}(\bar{x}_{ii'}) = \bar{x}_{ii'}$. L'expression (4.4) devient alors :

$$F(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \phi(a_{ii'})x_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{\phi}(a_{ii'})\bar{x}_{ii'} + K. \tag{4.5}$$

Dans l'équation (4.5) le critère est composé de façon symétrique des fonctionnelles linéaires suivantes :

1. Les accords positifs (ou attractions positives), la quantité : $\sum_{i=1}^{N}\sum_{i'=1}^{N}\phi(a_{ii'})x_{ii'}$.
2. Les accords négatifs (ou attractions négatives), la quantité : $\sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{\phi}(a_{ii'})\bar{x}_{ii'}$.

En effectuant le changement de variable $\bar{x}_{ii'} = 1 - x_{ii'}$ dans l'équation (4.5), cette formule peut se réécrire comme

$$F(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(\phi_{ii'} - \bar{\phi}_{ii'})x_{ii'} + K. \tag{4.6}$$

Dans le cas où le critère est à maximiser. La présence et/ou l'absence des termes $\phi_{ii'}$ et $\bar{\phi}_{ii'}$ joue un role très important dans la solution optimale du critère.

En effet, à partir de l'expression (4.5) il est facile de vérifier que si le terme $\bar{\phi}_{ii'} = 0\,\forall i, i'$, i.e. si le terme d'accords négatifs n'existe pas la solution qui maximise $F(X)$ est la partition **grossière** où tous les sommets sont réunis en une seule et unique classe, donc $\kappa = 1$ et $x_{ii'} = 1 \quad \forall (i, i')$ et $F(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\phi_{ii'}$.

De façon analogue, dans (4.5) si le terme $\phi_{ii'} = 0\,\forall i, i'$, i.e. en absence du terme d'accords positifs la solution optimale est la partition **triviale** où tous les sommets sont isolés les uns par rapport aux autres, donc $\kappa = N$ et $x_{ii'} = 0$ si $i \neq i'$ et $x_{ii} = 1\,\forall i$ et $F(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{\phi}_{ii}$.

> En conclusion, l'optimisation de tout critère linéaire ne vérifiant pas la propriété d'équilibre général rendra une solution soit triviale soit grossière ; obligeant ainsi à l'utilisateur à fixer a priori le nombre de classes de la partition cherchée.

À partir des résultats précédents nous pouvons déduire que les valeurs que les fonctions ϕ et $\bar{\phi}$ prennent permettent de créer une sorte de *balance* ou d'équilibre entre le fait de générer plusieurs classes (partition triviale, $\kappa = N$) et le fait de générer peu de classes (partition grossière, $\kappa = 1$). C'est ces fonctions que nous allons caractériser au chapitre 6 pour les critères que nous allons présenter au chapitre 5.

En tenant compte de ces résultats nous définissons deux niveaux d'équilibre pour les critères possédant la propriété d'équilibre général :

Définition 4.1 (**Propriété d'équilibre Local**). *Tout critère linéaire qui vérifie la propriété d'équilibre général et dont les fonctions $\phi_{ii'}$ et $\bar{\phi}_{ii'}$ vérifient*

$$\phi_{ii'} + \bar{\phi}_{ii'} = K \quad \forall\, i, i'$$

possède la propriété d'équilibre local.

Cela implique que pour toute paire de sommets (i, i'), donc au niveau local, lorsque $\phi_{ii'}$ augmente $\bar{\phi}_{ii'}$ diminue et vice-versa.

Définition 4.2 (**Propriété d'équilibre Global**). *Tout critère linéaire qui vérifie la propriété d'équilibre général et dont les fonctions $\phi_{ii'}$ et $\bar{\phi}_{ii'}$ vérifient*

$$\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\phi_{ii'}+\bar{\phi}_{ii'}\right)=K$$

possède la propriété d'équilibre global.

La définition d'équilibre global implique que pour tout le réseau, donc au niveau global, la somme des valeurs prises par $\phi_{ii'}$ plus la somme des valeurs prises par $\bar{\phi}_{ii'}$ soit constante.

Nous pouvons déduire des deux définitions précédentes qu'un critère vérifiant la propriété d'équilibre local vérifie aussi la propriété d'équilibre global. En revanche, la vérification de la propriété d'équilibre global ne garantit pas que le critère soit équilibré localement.

Un cas particulier des critères vérifiant la propriété d'équilibre global est un modèle dit **nul**. Un critère se basant sur un **modèle nul** vérifie la propriété suivante :

$$\sum_{i=1}^{N}\sum_{i'=1}^{N}\phi_{ii'}=\sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{\phi}_{ii'} \tag{4.7}$$

Il ne peut pas exister un critère vérifiant la propriété d'équilibre local et étant un modèle nul simultanément. En effet, cela impliquerait que $\phi_{ii'}=\bar{\phi}_{ii'} \quad \forall(i,i')$ et par conséquent, selon l'expression (4.6) le critère vaudrait alors $F(X)=\sum_{i=1}^{N}\sum_{i'=1}^{N}(\phi_{ii'}-\bar{\phi}_{ii'})x_{ii'}=0$ ce qui n'aurait aucun sens. Par conséquent, les critères possédant la propriété d'équilibre local et les modèles nuls sont deux sous-ensembles disjoints de l'ensemble de critères équilibrés globalement comme le montre le schéma 4.1.

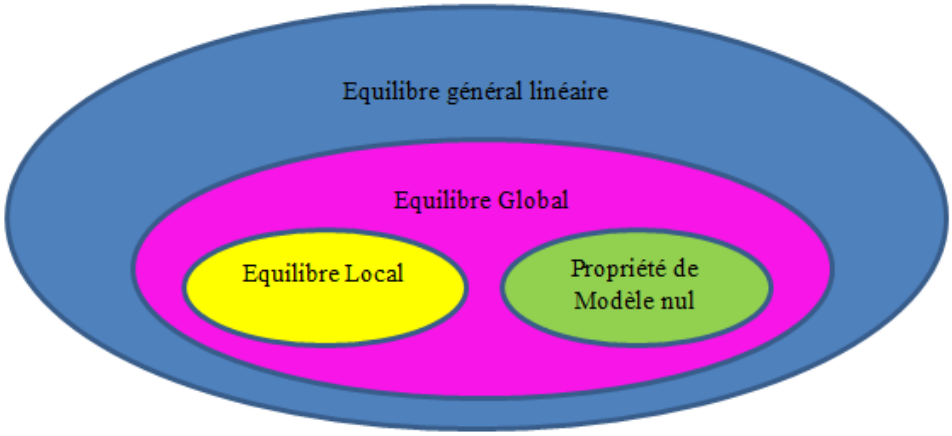

FIGURE 4.1 – Diagramme de Venn des critères vérifiant la propriété d'équilibre général.

Un critère de modularisation qui repose sur la définition de modèle nul possède une **limite de résolution**. La limite de résolution est une caractéristique du critère qui fait que son optimisation ne permet pas la détection de communautés en dessous d'une certaine

échelle qui dépend de caractéristiques globales du réseau comme la taille du graphe par exemple, même s'il n'existe aucune ambigüité par rapport à l'existence de communautés, comme par exemple la détection des cliques connectées par une seule arête. La limite de résolution est une conséquence de la définition globale du critère qui a pour conséquence que le critère ne soit pas invariant d'échelle.

Le fait qu'un critère possède une limite résolution est considéré un inconvénient. Cette propriété a été introduite dans l'article très connu de Fortunato and Barthelemy [2006], où les auteurs ont étudié la limite de résolution du critère de modularisation le plus connu, la modularité de Newman-Girvan.

Au chapitre suivant nous présenterons certains critères de modularisation et nous étudierons leurs propriétés. Au chapitre 6 nous montrerons comment la limite de résolution empêche aux critères se basant sur un modèle nul d'identifier certains groupes de sommets densément connectés.

Chapitre 5

Critères de Modularisation

5.1 Introduction

Dans cette partie nous allons privilégier la mise en exergue d'une structuration logique à l'expression d'un nombre important de critères de modularisation de graphes. En effet, différents critères de clustering de graphes ont été proposés au fil du temps pour modéliser des phénomènes issus de domaines divers.

Comme nous l'avons vu précédemment, modulariser un graphe revient à trouver une relation $\mathbf{X}$ (une partition sur l'ensemble de sommets) la *plus proche possible* de $\mathbf{A}$. Pour évaluer cette *proximité* il faut définir soit une fonction de $\mathbf{A}$ et de $\mathbf{X}$ qui mesure une distance (ou un écart), soit une similarité ou une proximité entre ces deux matrices. C'est cette fonction que nous appellerons *critère de modularisation*. Ce critère sera à minimiser dans le cas d'un écart ou à maximiser dans le cas d'une proximité, il sera noté de façon générale $F(\mathbf{A}, \mathbf{X})$. Ainsi tout critère sera écrit en notation relationnelle de la façon suivante :

$$\max_X \text{ ou } \min_X F(A, X). \tag{5.1}$$

Bien évidement s'il n'existait aucune contrainte sur l'inconnue $\mathbf{X}$ la meilleure solution de ce problème serait de poser $\mathbf{X} = \mathbf{A}$ mais $\mathbf{X}$ doit représenter une partition, i.e. elle doit de facto posséder les propriétés d'une relation d'équivalence, à savoir : symétrie, réflexivité et transitivité. Ainsi $\mathbf{X}$ doit vérifier les contraintes :

$$\begin{array}{rcl} x_{ii'} \in \{0,1\} & & \text{Binarité} \\ x_{ii} = 1 & \forall i & \text{Réflexivité} \\ x_{ii'} - x_{i'i} = 0 & \forall (i,i') & \text{Symétrie} \\ x_{ii'} + x_{i'i''} - x_{ii''} \leq 1 & \forall (i,i',i'') & \text{Transitivité.} \end{array} \tag{5.2}$$

$\mathbf{A}$ n'appartient pas, en général, à l'ensemble des solutions possibles pour un graphe non-orienté et non pondéré, la matrice $\mathbf{A}$ garantit seulement les propriétés de binarité et de symétrie.

Nous séparons les critères en deux catégories : linéaires et non-linéaires puisque la propriété de linéarité joue un rôle très important tant dans l'obtention de la solution du problème que dans la compréhension des résultats associés.

Il est important de noter ici que nous ne présenterons qu'une liste non-exhaustive de critères de classification ou de modularisation.

5.2 Critères linéaires en X

Le tableau 5.2 montre l'écriture relationnelle des critères linéaires, qui sont par voie de conséquence, également séparables.

5.2.1 Le critère de Zahn-Condorcet (1964,1785)

C. T. Zahn dans son article fondamental "Approximating Symmetric Relations by Equivalence Relations" (cf.Zahn [1964]) se pose pour la première fois la question de trouver une relation d'équivalence X qui s'approche le mieux possible d'une relation binaire symétrique donnée R, définie sur les paires d'éléments d'un ensemble fini V. Pour satisfaire un tel but il définit la fonction de *distance* ρ suivante entre les deux relations à minimiser.

$$\rho(X) = |X - R| + |R - X|, \tag{5.3}$$

où :

- R relation binaire connue.
- X est une relation d'équivalence qui doit approximer au mieux R au sens de la distance ρ.
- La notation $|B|$ signifie le cardinal de l'ensemble B.
- La notation $A - B \equiv A \cap \bar{B}$.

Dans Zahn [1964] l'auteur se rend compte que ce la relation R peut être facilement modélisée à partir d'un graphe non-orienté et sans boucles où les sommets seraient les éléments de l'ensemble V et les relations entre eux seraient les arêtes.

Pour définir la fonction d'éloignement 5.3 Zahn traite la relation R comme un sous-ensemble du produit cartésien $V \times V$. Les paires $(v_i, v_i') \in V \times V$ qui ne sont pas en relation peuvent être dénotées par $\bar{R}$, le complément de R dans $V \times V$. L'équation 5.3 peut s'écrire de la façon suivante :

$$\rho(X) = |X \cap \bar{R}| + |R \cap \bar{X}|.$$

Ainsi l'équation 5.3 n'est autre que le critère de Condorcet pour $k = 1$:

$$F_{ZC}(X) = \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} (\bar{a}_{ii'} x_{ii'} + a_{ii'} \bar{x}_{ii'}), \tag{5.4}$$

où $x_{ii'}$ doit vérifier les contraintes d'une relation d'équivalence (5.2)

L'équation (5.4) peut aussi s'écrire comme une fonction à maximiser :

$$\begin{aligned} F_{ZC}(X) &= \tfrac{1}{2} \textstyle\sum_{i=1}^{N} \sum_{i'=1}^{N} (\bar{a}_{ii'} x_{ii'} + a_{ii'} \bar{x}_{ii'}) \\ &= \tfrac{1}{2} \textstyle\sum_{i=1}^{N} \sum_{i'=1}^{N} (\bar{a}_{ii'} x_{ii'} + a_{ii'} (1 - x_{ii'})) \\ &= \tfrac{1}{2} \textstyle\sum_{i=1}^{N} \sum_{i'=1}^{N} (\bar{a}_{ii'} - a_{ii'}) x_{ii'} + a_{ii'}), \end{aligned}$$

ce qui équivaut à maximiser :

$$F_{ZC}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(a_{ii'} - \bar{a}_{ii'})x_{ii'} - K. \tag{5.5}$$

A partir de cette dernière expression nous pouvons déduire ses propriétés :
- Il est linéaire, donc séparable.
- Il possède la propriété d'équilibre général donc il rend une solution optimale non triviale ni grossière.
- Il est aussi équilibré localement car $\phi_{ii'} + \bar{\phi}_{ii'} = a_{ii'} + \bar{a}_{ii'} = 1$. Par conséquence, il est aussi équilibré globalement : $\sum_{i=1}^{N}\sum_{i'=1}^{N}(a_{ii'} + \bar{a}_{ii'}) = N^2$.

Le théorème suivant caractérise la densité d'arêtes des classes obtenues via l'optimisation du critère de Zahn-Condorcet :

Théorème 5.1 (Les classes obtenues en maximisant le critère de Zahn-Condorcet)**.** *Étant donné un graphe connecté $G = (V, E)$, non pondéré et non orienté, le partitionnement obtenu en maximisant le critère de Zahn-Condorcet possède la propriété suivante : le nombre d'arêtes à l'intérieur de chaque classe d'équivalence est supérieur ou égal à la moitié du nombre maximal d'arêtes intra-classe possibles, c'est-à-dire le nombre d'arêtes existant dans le cas où les éléments de la classe forment un graphe complet.*

Démonstration. En tenant compte des contraintes de réflexivité et symétrie de la variable $x_{ii'}$ (i.e. $x_{ii} = 1 \forall i$ et $x_{ii'} = x_{i'i}$), l'équation (5.5) s'écrit

$F_{ZC}(X) = \sum_{i>i'}(a_{ii'} - \bar{a}_{ii'})x_{ii'} + N^2 - 2M - N.$

S'il n'y a pas de données manquantes [1] nous avons :

- $\sum_{i>i'} a_{ii'}x_{ii'}$ est le nombre d'arêtes intra-classe pour toutes les classes.
- $\sum_{i>i'} \bar{a}_{ii'}x_{ii'}$ est le nombre d'arêtes manquantes dans chaque classe pour que les éléments de la classe forment un graphe complet.

Si l'on note E_j le nombre total d'arêtes à l'intérieur de la classe j. Le nombre total d'arêtes manquantes pour que la classe j soit un graphe complet sera : $\left(\frac{n_j(n_j-1)}{2} - E_j\right)$. Avec ces notations le critère de Zahn-Condorcet se réécrit :

$F_{ZC}(\mathcal{C}) = \sum_{j=1}^{\kappa}(E_j - (\frac{n_j(n_j-1)}{2} - Ej)) + N^2 - 2M - N,$

soit
$F_{ZC}(\mathcal{C}) = \sum_{j=1}^{\kappa}(2E_j - \frac{n_j(n_j-1)}{2}) + N^2 - 2M - N.$

Le terme $(2E_j - \frac{n_j(n_j-1)}{2})$ représente la contribution de la classe j à la valeur du critère. Pour chaque classe j de la partition optimale $\mathcal{C}$ cette contribution doit être positive, c'est-à-dire :

1. Dans la suite de ce document cette hypothèse sera considérée vérifiée.

$(2E_j - \frac{n_j(n_j-1)}{2}) \geq 0$, soit $E_j \geq \frac{n_j(n_j-1)}{4}$.

S'il existe une classe j telle que $(2E_j - \frac{n_j^2}{2}) < 0$, cette classe ne fait pas partie de la partition optimale car sa contribution est négative. Il est toujours possible d'augmenter la valeur du critère en séparant les sommets de j. Par exemple, rien qu'en isolant tous les sommets de j la contribution de chaque classe obtenue devient nulle. □

Le théorème 5.1 montre que la densité d'arêtes des classes obtenues via l'optimisation du critère de Zahn-Condorcet est supérieure ou égale à 50%. Ce résultat est une conséquence de "*La règle de la majorité absolue de Condorcet*" en théorie des votes.

Remarque importante : le théorème 5.1 s'applique à chaque classe d'équivalence du partitionnement optimal et non pas à chaque paire de sommets d'une classe.

5.2.2 Le critère paramétré d'Owsiński-Zadrożny (1986)

Dans certains cas la condition de majorité absolue intra-classe imposée par le critère de Zahn-Condorcet peut paraître sévère et exigeante. Tout dépend fortement de la définition que l'on donne à la notion de *communauté* ou du besoin que l'on a d'obtenir des communautés denses. Si l'on appelle α le pourcentage minimal requis d'arêtes intra-classe, on peut introduire ce paramètre dans la fonction à maximiser qui s'écrit alors :

$$F_{OZ}(X) = \sum_{i=1}^{N}((1-\alpha)a_{ii'}x_{ii'} + \alpha\bar{a}_{ii'}\bar{x}_{ii'}) avec \qquad 0 < \alpha < 1. \tag{5.6}$$

Bien évidemment sous les contraintes d'une relation d'équivalence, équation (5.2). Cette dernière écriture constitue une généralisation du critère de Condorcet, déjà proposée par Owsiński and Zadrożny [1986].

Les propriétés de ce critère sont :

- Il est linéaire, donc séparable.
- Il possède la propriété d'équilibre général donc il rend une solution optimale non triviale ni grossière.
- il est aussi équilibré globalement : $\sum_{i=1}^{N}\sum_{i'=1}^{N}((1-\alpha)a_{ii'} + \alpha\bar{a}_{ii'}) = (1-\alpha)2M + \alpha(N^2 - 2M)$.
  Il n'est pas équilibré localement sauf pour $\alpha = 0.5$ où ce critère est équivalent au critère de Zahn-Condorcet.

La constante α définit l'équilibre entre la composante d'attractions positives : $\sum_{ii'} a_{ii'}x_{ii'}$ et la composante d'attractions négatives : $\sum_{ii'} \bar{a}_{ii'}\bar{x}_{ii'}$. Ce terme dépend de la définition que l'utilisateur donne à la notion de communauté. Si α est proche de 1 plus denses seront les connexions intra-classe. Si $\alpha = 0.5$ on obtient la fonction de Zahn-Condorcet.

L'équation (5.6) peut se réécrire comme

$$F_{OZ}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}((a_{ii'} - \alpha)x_{ii'}) + K \qquad 0 < \alpha < 1. \tag{5.7}$$

Ici nous avons $K = \alpha \sum_{i=1}^{N} \sum_{i'=1}^{N} \bar{a}_{ii'}$. Bien que cette dernière quantité dépende de α, elle est constante car α est fixé à l'avance par l'utilisateur.

Nous pouvons aussi caractériser les obtenues via l'optimisation du critère d'Owsiński-Zadrożny :

Théorème 5.2 (Les classes obtenues en maximisant le critère d'Owsiński-Zadrożny)**.** *Étant donné un graphe connecté $G = (V, E)$, non pondéré et non orienté, le partitionnement obtenu en maximisant le critère d'Owsiński-Zadrożny possède la propriété suivante : le nombre d'arêtes à l'intérieur de chaque classe d'équivalence est supérieur ou égal à $\alpha\%$ du nombre maximal d'arêtes intra-classe possibles.*

Démonstration. La démonstration est analogue à celle du théorème 5.1 si l'on remplace $\frac{1}{2}$ par α. □

Selon le théorème 5.2 la densité d'arêtes des classes obtenues via l'optimisation du critère de d'Owsiński-Zadrożny sera supérieure ou égale à $\alpha\%$.

5.2.3 Le critère de Newman-Girvan (2004) : la modularité proprement dite

La *modularité* de Newman-Girvan est aujourd'hui le critère de modularisation le plus connu et par conséquent il est assez souvent utilisé en tant que fonction qualité dans plusieurs algorithmes de classification des graphes.

Les auteurs ont proposé ce critère dans Newman and Girvan [2004] comme une mesure de la force de la structure communautaire d'un graphe. En principe, ils avaient employé un certain nombre d'algorithmes pour modulariser des graphes réels. Un de ces algorithmes est le très connu algorithme de "Girvan-Newman" (voir Girvan and Newman [2002]). Cette méthode consiste à enlever itérativement les arêtes possédant une mesure de centralité d'intermédiarité (*betweenness* en anglais) élevée, i.e. détecter les arêtes à couper (*cuts*). En résumé, le principe était le suivant : les arêtes inter-classes jouant le rôle d'intermédiaires entre 2 communautés et ces arêtes inter-classes étant plus rares que les autres, elles possèdent une forte valeur d'intermédiarité. Par ailleurs le résultat obtenu à la sortie de chaque algorithme était un dendogramme.

Cependant cette démarche présentait plusieurs inconvénients. D'une part, les coupures ne révélaient pas correctement la notion de structure de communauté. D'autre part, il fallait décider à quel niveau couper le dendogramme fourni par l'algorithme.

C'est pour faire face à ces problèmes que les auteurs ont défini un critère qui maximise l'écart entre le graphe original et sa version aléatoire correspondante en partant du principe qu'un graphe aléatoire ne possède pas de structures communautaires. En effet, le principe de construction de ce critère énoncé dans Newman and Girvan [2004] est : *Only if the number of between-group edges is significantly lower than would be expected purely by chance can we justifiably claim to have found significant community structure*" (ce n'est seulement que si le nombre d'arêtes entre deux groupes distincts est nettement plus faible que ce qui est prévu par le hasard pur que nous pouvons légitimement prétendre avoir trouvé une structure communautaire significative). Ce principe implique que tous

les graphes n'ont pas une structure communautaire intrinsèque.

La fonction à maximiser proposée par Newman and Girvan [2004], plus connue aujourd'hui sous le nom de *modularité*, cherche à maximiser la différence entre le nombre d'arêtes intra-classe et l'espérance de cette valeur dans un graphe avec la même distribution des degrés mais où les arêtes seraient placées de façon aléatoire sans référence aux communautés.

Soit $\mathbf{e}$ une matrice $\kappa \times \kappa$ (où κ est le nombre de classes de la partition cherchée) dont l'élément $e_{jj'}$ représente la fraction d'arêtes reliant les sommets de la classe j aux sommets de la classe j'. La fonction à maximiser proposée par Newman and Girvan [2004], plus connue aujourd'hui comme "fonction de modularité", est donnée par l'expression suivante :

$$F_{NG}(\mathbf{e}) = \sum_{j}^{\kappa} \left(e_{jj} - \left(\sum_{j'}^{\kappa} e_{jj'} \right)^2 \right) = Tr(\mathbf{e}) - \left(\sum_{j}^{\kappa} \sum_{j'}^{\kappa} [\mathbf{e}^2]_{jj'} \right), \tag{5.8}$$

où $[\mathbf{e}^2]_{jj'}$ représente l'élément jj' de la matrice $\mathbf{e}^2$, donc le deuxième terme de (5.8) n'est autre que la somme des éléments de la matrice $\mathbf{e}^2$. Dans l'expression (5.8) le premier terme représente la proportion d'arêtes intra-classe. Le deuxième terme représente l'espérance d'arêtes intra-classe dans un graphe aléatoire possédant le même nombre d'arêtes et la même distribution des degrés. Ainsi, obtenir une valeur de la fonction de modularité proche de l'unité signifie que le partitionnement obtenu possède des communautés densément connectées et un écart important entre la distribution réelle d'arêtes et celle d'un graphe aléatoire.

Une autre formulation de ce critère utilisée souvent par d'autres auteurs, notamment par Brandes et al. [2008] est :

$$F_{NG}(\kappa) = \sum_{j=1}^{\kappa} \left(\frac{|E(\mathcal{C}_j)|}{M} - \left(\frac{\sum_{i \in \mathcal{C}_j} d_i}{2M} \right)^2 \right), \tag{5.9}$$

où M est le nombre total d'arêtes (ou la somme des poids dans le cas d'un graphe pondéré), $E(\mathcal{C}_j)$ est l'ensemble d'arêtes intra-classe de la classe $\mathcal{C}_j$ et d_i est le degré du sommet i.

En notations relationnelles l'expression de ce critère est la suivante [2] (voir aussi Labiod et al. [2010]) :

$$F_{NG}(X) = \frac{1}{2M} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(a_{ii'} - \frac{a_{i.} a_{.i'}}{2M} \right) x_{ii'}, \tag{5.10}$$

avec $\mathbf{X}$ vérifiant également les contraintes caractéristiques d'une relation d'équivalence, équation (5.2).

Où le terme $\frac{1}{2M}$ est une constante de normalisation servant à comparer la modularité de graphes de tailles différentes. Par conséquent la modularité appartient à l'intervalle

2. Dans la formulation originale proposée par Newman-Girvan la variable $\mathbf{X}$ apparaît comme une sorte de *mesure de Dirac* $\delta(i, i')$ ou comme une *variable indicatrice*.

$\left[-\frac{1}{2}, 1\right]$. La quantité $a_{i.} = a_{.i} = \sum_{i'=1}^{N} a_{ii'}$ représente le degré du sommet i : $d(i)$. Le terme $\frac{a_{i.}a_{.i'}}{2M}$ représente l'espérance du nombre d'arêtes $a_{ii'}$ entre les sommets i et i' (ou le poids des arêtes dans le cas d'un graphe pondéré) dans la version aléatoire du graphe, i.e. un graphe possédant le même nombre total d'arêtes $\sum_{i=1}^{N}\sum_{i'=1}^{N} \frac{a_{i.}a_{.i'}}{2M} = 2M$ et la même distribution des degrés $\sum_{i'=1}^{N} \frac{a_{i.}a_{.i'}}{2M} = a_{i.} = d_i \quad \forall i$ mais où les arêtes sont placées de façon aléatoire sans égard pour les communautés. [3] En effet, sans aucune affinité entre les sommets, étant donné un sommet u, le nombre ou (le poids d'arêtes) reliant u à un autre sommet v est proportionnel au nombre total d'arêtes partant de v, donc à son degré d_v. Cette définition de graphe aléatoire révèle la notion d'indépendance statistique dans la distribution d'arêtes.

L'écriture relationnelle de l'expression (5.10) permet de **faire disparaitre le nombre de classes de la partition cherchée**, ainsi on n'est pas obligé de le fixer à l'avance. De plus, elle met en évidence les principales propriétés du critère :

- Linéarité et par conséquent séparabilité.
- Equilibre général linéaire, donc le critère possède une solution optimale différente de la partition triviale et de la partition grossière (sauf dans certaines exceptions).
- Ce critère est équilibré globalement et plus particulièrement il s'agit d'un **modèle nul** : $\sum_{i=1}^{N}\sum_{i'=1}^{N} \phi_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{\phi}_{ii'} = 2M$. Il possède, donc une limite de résolution ne permettant pas la détection de communautés en dessous d'une certaine échelle qui dépend du nombre total d'arêtes et le degré d'interconnectivité des modules. En effet, comme $\frac{a_{i.}a_{.i'}}{2M}$, ce critère aura tendance à créer de grandes classes ou de petites classes ceci selon la distribution des degrés du graphe. Au chapitre suivant nous caractériserons le comportement de ce terme.

La limite de résolution de ce critère a été mise en évidence dans Fortunato and Barthelemy [2006] (voir aussi Good et al. [2010]). Certains auteurs ont travaillé aussi sur quelques approches pour résoudre ce problème (voir Reichardt and Bornholdt [2006] et Arenas et al. [2008]).

D'autres propriétés intéressantes de ce critère sont discutées dans Brandes et al. [2008]. Notamment la propriété de *non-locality* (non-localité), qui met en évidence que la "modularité" de Newman-Girvan, étant un critère formulé de façon globale, l'ajout d'un sommet avec un seul voisin peut changer complètement la partition optimale obtenue avant ledit ajout. De Montgolfier et al. [2012] ont montré aussi que des graphes très réguliers ne possédant aucune structure communautaire naturelle (grilles, hypercubes,...) ont une modularité asymptotiquement égale à 1.

Il est intéressant de remarquer que le terme entre parenthèses de l'équation (5.10) est un écart à la situation d'indépendance, comme l'indice d'association entre 2 variables qualitatives introduit il y a quelques années par Belson [1959] ou encore le numérateur de la

3. Cette définition de la notion de graphe aléatoire correspond à la solution optimale du modèle de flux d'entropie d'Alain Wilson (*Flows Entropy Model*) introduit dans Wilson [1967] dont l'objectif est de déterminer la distribution de flux d'échanges entre individus qui maximise l'entropie sous les contraintes de marges fixées et de conservation de masse. Dans le graphe aléatoire de Newman-Girvan, ces contraintes correspondent à la conservation de la distribution des degrés et du nombre total d'arêtes.

mesure d'association du χ^2 d'écart à l'indépendance calculé sur un tableau de contingence. En effet, la matrice d'adjacence $\mathbf{A}$ d'un graphe (pondéré ou pas) peut être vue comme un tableau de contingences qui croise 2 variables à N modalités observées sur $2M$ objets (extrémités ou bouts de chaque arête) [4]. Le terme général de cette matrice $\mathbf{A}$ serait, dans ce cas-là, interprété comme :

$a_{ii'}$: Nombre d'extrémités d'arêtes reliant le sommet i au sommet i'.

L'indice de Belson $\mathcal{B}(X,Y)$ (cf. Belson [1959]) pour un tableau de contingence de terme général n_{uv} qui croise 2 variables X et Y à p et q modalités respectivement observées sur N objets s'écrit :

$$\mathcal{B}(X,Y) = \sum_{u=1}^{p}\sum_{v=1}^{q}\left(n_{uv} - \frac{n_{u.}n_{.v}}{N}\right)^2. \tag{5.11}$$

Belson effectue la somme des écarts au carré pour ne pas sommer et soustraire des écarts dont la somme par définition est nulle. Ce critère appliqué au tableau de contingence A sera :

$$\mathcal{B}(X,Y) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{a_{i.}a_{.i'}}{2M}\right)^2. \tag{5.12}$$

5.2.4 La Version équilibrée du critère de Newman-Girvan (2013)

De façon analogue au critère de Newman-Girvan qui contient un terme d'accords positifs, le nombre d'arêtes intra-classe moins son espérance dans le cas d'un graphe vérifiant la structure d'indépendance (terme, entre parenthèses, de l'équation (5.10), la Modularité Équilibrée contient de plus un terme d'accord négatifs : le nombre d'arêtes inter-classes moins l'espérance de cette quantité dans la cas d'un graphe aléatoire vérifiant la structure d'indépendance. L'expression de ce critère, proposée par moi-même et énoncée dans Conde-Céspedes and Marcotorchino [2013] (voir aussi Conde-Céspedes and Marcotorchino [2013]), est la suivante :

$$F_{BM}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\left(a_{ii'} - \frac{a_{i.}a_{.i'}}{2M}\right)x_{ii'} + \left(\bar{a}_{ii'} - \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M}\right)\bar{x}_{ii'}\right), \tag{5.13}$$

où BM signifie *Balanced modularity* en anglais. Où $\mathbf{X}$ doit vérifier les contraintes d'une relation d'équivalence (5.2).

En notant $p_{ii'} = \frac{a_{i.}a_{.i'}}{2M}$ et $\bar{p}_{ii'} = \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M}$. De la même façon que le terme d'accords positifs vérifie $\sum_{i=1}^{N}\sum_{i'=1}^{N} p_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'} = 2M$, le terme d'accords négatifs vérifie $\sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{p}_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N}\frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M} = \sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{a}_{ii'} = N^2-2M.$

4. Il est important de remarquer que la matrice $\mathbf{A}$ est un tableau de contingence spécial car elle est symétrique.

Ce critère possède des caractéristiques voisines de celui de Newman-Girvan. Il est linéaire, donc séparable, il possède la propriété d' Equilibre général linéaire, et il est équilibré globalement et plus particulièrement il s'agit d'un **modèle nul** :

$$\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'}+\frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M}\right)=\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\bar{a}_{ii'}+\frac{a_{i.}a_{.i'}}{2M}\right)=N^2$$

.

Il possède, donc une **limite de résolution**.

5.2.5 Le critère d'Écart à l'Indétermination (2013)

Ce critère part du principe que l'on peut faire le parallèle entre la construction de critères en environnement contingentiel et ce qui se passe dans le domaine de la théorie des graphes.

Il y a deux standards de formation de critères de contingence qui sont définis à partir d'un tableau de contingences croisant deux variables catégorielles A et X (qualitatives, relations d'équivalence)[5] à p et q modalités ou catégories respectivement :

$A\backslash X$	1	...	v	...	q	Total
1	n_{11}	...	n_{1v}	...	n_{1q}	$n_{1.}$
⋮	⋮		⋮		⋮	⋮
u	n_{u1}	...	n_{uv}	...	n_{uq}	$n_{u.}$
⋮	⋮		⋮		⋮	⋮
p	n_{p1}	...	n_{pv}	...	n_{pq}	$n_{p.}$
Total	$n_{.1}$	...	$n_{.v}$	...	$n_{.q}$	$n_{..}$

Table 5.1 – Tableau de contingence

- n_{uv} est le nombre d'objets ayant les modalité u et v de A et X respectivement.
- $n_{u.}$ est le nombre d'objets ayant la modalité u de A.
- $n_{.v}$ est le nombre d'objets ayant la modalité v de X.
- $n_{..}=N$ est le nombre total d'objets.

On définit alors deux principes :

1. Le principe d'**écart à l'indépendance** où l'on compare chaque quantité n_{uv} à $\frac{n_{u.}n_{.v}}{N}$. Dans le cas d'indépendance parfaite entre les deux variables A et X nous avons $n_{uv}=\left(\frac{n_{u.}n_{.v}}{N}\right)\ \forall(u,v)$, quantité qui annule l'indice de Belson (équation (5.11)) ou le numérateur de la statistique du χ^2 utilisée pour tester l'indépendance entre deux variables aléatoires. Cette quantité correspond aussi à la solution optimale du modèle de flux d'entropie d'Alain Wilson (voir Wilson [1967], Wilson [1969] et Wilson [1970]). Wilson considère un système dont les éléments ne possèdent aucune

5. Ici A représente une variable qualitative ou une relation d'équivalence et non pas la matrice d'adjacence d'un graphe qui n'est pas, en général, une relation d'équivalence sinon une relation binaire symétrique dans le cas de graphes non orientés.

affinité entre eux et l'objectif est de déterminer la distribution de flux d'échange normalisés $\pi_{uv} = \frac{n_{uv}}{N}$ (avec $\pi_{uv} > 0 \forall (u,v)$) entre individus qui maximise l'entropie du système : $-\sum_{u=1}^{p}\sum_{v=1}^{q}\pi_{uv}\ln\pi_{uv}$ sous les contraintes de marges fixées et de conservation de masse. En effet, sans aucune affinité particulière entre lignes et colonnes, étant donné une colonne v le flux d'une ligne quelconque u à cette colonne est proportionnelle à sa marginale $n_{u.}$. Idem si l'on fixe une ligne u le flux d'une colonne quelconque v vers u sera proportionnelle à sa marginale $n_{.v}$ (voir Marcotorchino [2013] et Marcotorchino and Conde-Céspedes [2013] pour plus de détails).

2. Le principe d'**écart à l'indétermination** où l'on compare n_{uv} à $\left(\frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{N}{pq}\right)$. Dans le cas d'une situation d'indétermination parfaite entre A et X nous avons $n_{uv} = \frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{N}{pq} \forall (u,v)$ quantité qui annule le numérateur du coefficient de Janson-Vegelius (voir Janson and Vegelius [1982]), indice qui sert, entre autres, à mesurer la similarité entre deux partitions (voir Youness and Saporta [2004]) [6] (voir Janson and Vegelius [1982]) : $\sum_{u=1}^{p}\sum_{v=1}^{q}\left(n_{uv} - \left(\frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{N}{pq}\right)\right)^2$. Cette quantité correspond aussi à la solution optimale du problème de *Commerce Minimal* ou *Minimal Trade Model* en anglais (voir Stemmelen [1977], Marcotorchino [1984a]). Le modèle de *Commerce Minimal* cherche à minimiser l'écart quadratique des valeurs n_{uv} à la distribution uniforme : $\min_{\pi}\sum_{u=1}^{p}\sum_{v=1}^{q}\left(\pi_{uv} - \frac{1}{pq}\right)^2$ sous les contraintes de marginales fixées et de conservation de masse (voir Marcotorchino [2013] pour plus de détails).

Le premier principe tient compte des valeurs marginales du tableau de contingence sous forme multiplicative, alors que le second principe tient compte des valeurs marginales sous forme additive. On trouvera dans Marcotorchino [2013] une intéressante synthèse de la dualité entre ces deux principes.

En se servant des formules de passage des notations contingentielles aux notations relationnelles (voir annexe A pour une liste non- exhaustive des formules de transfert, pour plus de détails sur leur obtention et démonstrations voir Kendall and Stuart [1961] et Marcotorchino [1984a]) on obtient la dualité suivante entre critères d'écart carré à l'indétermination et écart carré à l'indépendance (voir Ah-Pine and Marcotorchino [2007]) :

$$\sum_{u=1}^{p}\sum_{v=1}^{q}\left(n_{uv} - \frac{n_{u.}n_{.v}}{N}\right)^2 = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{a_{i.}}{N} - \frac{a_{.i'}}{N} + \frac{a_{..}}{N^2}\right)\left(x_{ii'} - \frac{x_{i.}}{N} - \frac{x_{.i'}}{N} + \frac{x_{..}}{N^2}\right) \quad (5.14)$$

$$\sum_{u=1}^{p}\sum_{v=1}^{q}\left(n_{uv} - \frac{n_{u.}}{q} - \frac{n_{.v}}{p} + \frac{N}{pq}\right)^2 = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{1}{p}\right)\left(x_{ii'} - \frac{1}{q}\right) \quad (5.15)$$

6. Le coefficient de Janson-Vegelius peut être interprété comme un coefficient de corrélation entre deux relations d'équivalence. (voir des développements dans Idrissi [2000] et Ah-Pine and Marcotorchino [2010] :

$$JV(A,X) = \frac{\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{1}{p}\right)\left(x_{ii'} - \frac{1}{q}\right)}{\sqrt{\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{1}{p}\right)^2}\sqrt{\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(x_{ii'} - \frac{1}{q}\right)^2}}.$$

Ce coefficient varie de -1 à +1 et s'annule si les variables vérifient la structure d'indétermination.

où les termes $a_{ii'}$ et $x_{ii'}$ représentent les termes généraux des matrices relationnelles de Condorcet des relations d'équivalence décrites par les variables A et X ; et $a_{..} = \sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}$ et $x_{..} = \sum_{i=1}^{N}\sum_{i'=1}^{N} x_{ii'}$.

On voit apparaître en sous-jacent le principe de décomposition matricielle de Torgerson du tableau de contingence qui apparaît à droite (côté relationnel) dans la formule (5.14) et à gauche (côté contingentiel) dans la formule (5.15) utilisée dans le domaine du Positionnement Multidimensionnel ou *Multidimensional scaling (MDS)* en anglais [7].

Focalisons-nous désormais sur la situation d'indétermination. En cas d'indétermination parfaite et à partir de la formule (5.15) nous obtenons l'expression suivante :

$$\sum_{u=1}^{p}\sum_{v=1}^{q}\left[n_{uv} - \left(\frac{n_{u.}}{q} + \frac{n_{.v}}{p} - \frac{N}{pq}\right)\right]^2 = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{1}{p}\right)\left(x_{ii'} - \frac{1}{q}\right) = 0. \tag{5.16}$$

A partir de l'expression (5.16) on peut déduire l'expression suivante

$$(p-1)(q-1)\sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}x_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{a}_{ii'}\bar{x}_{ii'} = (p-1)\sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}\bar{x}_{ii'} + (q-1)\sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{a}_{ii'}x_{ii'}. \tag{5.17}$$

Pour $p = 2$ et $q = 2$ on obtient :

$$\sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}x_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{a}_{ii'}\bar{x}_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}\bar{x}_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{a}_{ii'}x_{ii'}.$$

Le cas où il y a autant de voix de support *"pour"* que de voix de support *"contre"*, c'est-à-dire précisément la structure d'indétermination vraie qui a donné son nom à cette approche (voir Marcotorchino [1985] et Ah-Pine [2007] pour plus de détails). La formule (5.17) s'appelle alors structure d'indétermination générale pondérée.

En revenant au problème de modularisation de graphes, de façon analogue à la construction du critère de Newman-Girvan qui se base sur la maximisation de l'écart à l'indépendance et en tenant compte la dualité entre la structure d'indépendance et celle d'indétermination en statistiques de contingences, nous pouvons définir un nouveau critère de modularisation qui se base sur la maximisation de l'écart à l'indétermination.

Comme dans la section 5.2.3 voyons la matrice d'adjacence $\mathbf{A}$ d'un graphe non-orienté comme un tableau de contingence croisant deux variables, que nous appellerons A_1 et A_2, à N modalités chacune décrivant $2M$ objets, soit les extrémités de chaque arête, (car $\mathbf{A}$ est d'ordre N et la somme de ses termes $\sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'} = 2M$). Il est important de remarquer qu'il ne s'agit pas d'un tableau de contingence quelconque, tout d'abord il croise deux variables ayant le même nombre de modalités [8] N (le nombre de sommets) et il est

7. Le Positionnement Multidimensionnel est un ensemble de techniques d'analyse multivariée en statistiques utilisées souvent dans le domaine de la visualisation d'information pour explorer les similarités et les dissimilarités dans les données. La donnée d'entrée est une matrice de dissimilarités ou distances entre chaque paire de N objets.

8. Cela ne s'applique pas aux graphes bipartites dont les sommets sont partitionnés en deux sous-ensembles tels que chaque arête ait une extrémité dans un ensemble et l'autre dans l'autre ensemble

symétrique (car le graphe est non-orienté), en plus si le graphe est non-pondéré ce tableau est binaire.

Le terme général de la matrice d'adjacence vue comme un tableau de contingence peut être interprété de la façon suivante [9] :

$$a_{ii'} = \text{nombre d'extrémités d'arêtes sortant du sommet } i \text{ vers le sommet } i'.$$

Ainsi, une extrémité ou bout d'arête possède la modalité i de A_1 et la modalité i' de A_2 si elle sort du sommet i et si l'autre extrémité de son arête sort du sommet i'. Et comme chaque arête a deux extrémités et le graphe est non-orienté nous avons : $a_{i'i} = a_{ii'}$.

Sur ce tableau de contingence nous définissons un critère de modularisation qui maximise l'écart entre $a_{ii'}$ et $\left(\frac{a_{i.}}{n} + \frac{a_{.i'}}{N} - \frac{2M}{N^2}\right)$ que nous appellerons *Critère d'Écart à l'Indétermination* ou *Deviation to Indetermination : DI* en anglais (voir Marcotorchino [2013]) :

$$F_{DI}(X) = \frac{1}{2M} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(a_{ii'} - \frac{a_{i.}}{N} - \frac{a_{.i'}}{N} + \frac{2M}{N^2} \right) x_{ii'}, \tag{5.18}$$

où $\mathbf{X}$ doit vérifier les contraintes d'une relation d'équivalence, équation (5.2).

Nous assumons la condition de positivité suivante :

$$N(a_{i.} + a_{.i'}) \geq 2M \qquad \forall i, i'. \tag{5.19}$$

La condition (5.19) garantit la positivité du poids d'arêtes, dans le modèle vérifiant la structure d'indétermination. Comme nous le verrons au chapitre 6, le non respect de cette condition peut entrainer que la solution optimale contienne des classes à composantes non connectées.

Ce critère possède des propriétés voisines de celui de Newman-Girvan : il est linéaire, séparable et il est aussi un **modèle nul** car $\sum_{i=1}^{N} \sum_{i'=1}^{N} \left(a_{ii'} - \frac{a_{i.}}{N} - \frac{a_{.i'}}{N} + \frac{2M}{N^2}\right) = 0$, donc il a une limite de résolution.

Nous pouvons interpréter la situation d'indétermination comme un graphe qui possède les mêmes propriétés que le graphe d'origine en ce qui concerne le nombre total d'arêtes, $\sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2}\right) = 2M$, et la distribution des degrés [10] $\sum_{i'=1}^{N} \left(\frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2}\right) = a_{i.} = d_i$ mais où une fois le degré de chaque sommet fixé (ou la somme du poids des arêtes incidentes dans le cas d'un graphe pondéré), celui-ci est distribué de façon équitable entre toutes les arêtes du graphe tout en conservant la somme totale d'arêtes ou de poids total. Autrement dit, les degrés d_i et $d_{i'}$ étant fixés le poids de l'arête reliant les sommets i et i', soit $a_{ii'}$, sera un N-ème du degré de i, soit $\frac{d_i}{N}$, plus un N-ème du degré de i, soit $\frac{d_i}{N}$, moins le degré total équiréparti entre toutes les arêtes $\frac{2M}{N^2}$.

Nous parlons de graphe pondéré car les seuls graphes non-pondérés ayant la structure d'indétermination sont les graphes où aucun sommet n'est connecté, i.e. $a_{ii'} = 0 \, \forall i, i'$ et

9. Cette façon d'interpréter le terme général de la matrice d'adjacence équivaut à traiter le graphe non-orienté comme un graphe orienté symétrique.

10. Ces contraintes sur la distribution d'arêtes correspondent aux contraintes de conservation de masse et de marginales fixées du problème de *Commerce Minimal*.

une clique, i.e. $a_{ii'} = 1 \forall i, i'$ où tous les sommets sont connectés les uns avec les autres. La figure 5.1 montre un graphe pondéré dont les poids des arêtes vérifient la structure d'indétermination.

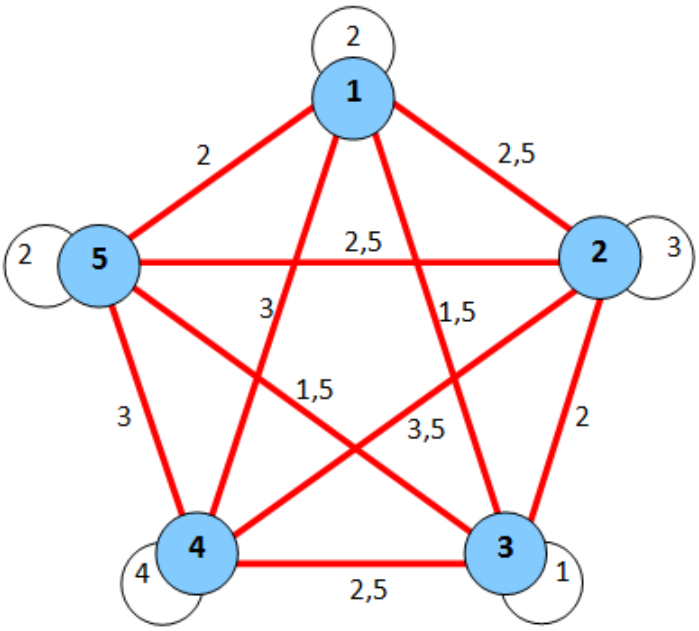

Figure 5.1 – Graphe pondéré vérifiant la structure d'indétermination.

Pour mieux comprendre la structure de l'écart à l'indétermination dans les graphes considérons alors un graphe pondéré mais non-orienté, dont le terme général de la matrice d'adjacence ou matrice des poids $\mathbf{W}$ est donné par $w_{ii'} = \left(\frac{w_{i.}}{N} + \frac{w_{.i'}}{N} - \frac{2M}{N^2}\right)$ où $w_{i.} = \sum_{i'=1}^{N} w_{ii'}$ est le poids total d'arêtes partant du sommet i et où $2M = \sum_{i=1}^{N}\sum_{i'=1}^{N} w_{ii'}$. Il est facile de démontrer que $\forall i \neq i', j \neq j' \quad (w_{ii'} + w_{jj'}) = (w_{ji'} + w_{ij'})$, c'est-à-dire que pour tout sous-tableau de taille 2×2 extrait de la matrice d'adjacence, la somme de ses termes diagonaux est égale à la somme de ses termes anti-diagonaux. Cette condition d'équilibre interne s'appuie sur ce que certains (dont Alan Hoffman dans Hoffman [1963]) ont appelé les "conditions de Monge d'un tableau" faisant référence aux travaux de Gaspard Monge définis dans Monge [1781] de la façon suivante :

Définition 5.1 (Tableau Monge et Tableau Anti-Monge)**.** *Une matrice $\boldsymbol{C}$ de taille $p \times q$ et de terme général c_{uv} est un tableau Monge si $\boldsymbol{C}$ vérifie la propriété :*

$$c_{uv} + c_{u'v'} \leq c_{uv'} + c_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\ 1 \leq v < v' \leq q.$$

Réciproquement, une matrice $\boldsymbol{C}$ est appelée un tableau anti- Monge si celle-ci vérifie :

$$c_{uv} + c_{u'v'} \geq c_{uv'} + c_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\ 1 \leq v < v' \leq q.$$

En l'occurrence la matrice de poids du graphe vérifiant la structure d'indétermination est Monge et AntiMonge. Ainsi nous en déduisons :

$$w_{ii} + w_{i'i'} = w_{i'i} + w_{ii'} \quad \forall i, i'$$

.

De plus, comme le graphe est non orienté la matrice de poids est symétrique $w_{i'i} = w_{ii'}$, cela implique que le poids de chaque arête vaut :

$$w_{ii'} = \frac{w_{ii} + w_{i'i'}}{2}. \tag{5.20}$$

Ainsi le poids de l'arête reliant les sommets i et i' est la moyenne arithmétique des poids de boucles des sommets se trouvant à chaque extrémité.

De façon analogue, la matrice de poids $\mathbf{W}$ d'un graphe dont le terme général vaut $w_{ii'} = \left(\frac{w_{i.}w_{.i'}}{2M}\right)$ vérifie $\forall i \neq i', j \neq j' \quad (w_{ii'}w_{jj'}) = (w_{ji'}w_{ij'})$, c'est-à-dire que pour tout sous-tableau de taille 2×2 extrait de la matrice d'adjacence, le produit de ses termes diagonaux est égale au produit de ses termes anti-diagonaux. En effet la matrice de poids est à la fois un tableau log-Monge et anti log-Monge dont la définition est la suivante (voir Marcotorchino [2013]) :

Définition 5.2 (Tableau Log-Monge et Tableau Anti Log-Monge)**.** *Une matrice $\boldsymbol{C}$ de taille $p \times q$ et de terme général $c_{uv} > 0$ est un tableau Log-Monge si $\boldsymbol{C}$ vérifie la propriété :*

$$\ln c_{uv} + \ln c_{u'v'} \leq \ln c_{uv'} + \ln c_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\ 1 \leq v < v' \leq q. \tag{5.21}$$

Réciproquement, une matrice $\boldsymbol{C}$ est appelée un tableau anti log-Monge si celle-ci vérifie :

$$\ln c_{uv} + \ln c_{u'v'} \geq \ln c_{uv'} + \ln c_{u'v} \quad \forall\, 1 \leq u < u' \leq p,\ 1 \leq v < v' \leq q. \tag{5.22}$$

A partir de ces résultats nous déduisons :

$$w_{ii}w_{i'i'} = w_{i'i}w_{ii'} \quad \forall i, i'$$

.

De plus, comme le graphe est non orienté, la matrice de poids est symétrique $w_{i'i} = w_{ii'}$, cela implique que le poids de chaque arête vaut :

$$w_{ii'} = \sqrt{w_{ii}w_{i'i'}}. \tag{5.23}$$

Donc, si la distribution d'arêtes du graphe vérifie la structure d'indépendance, le poids de l'arête reliant les sommets i et i' est la moyenne géométrique des poids de boucles des sommets se trouvant à chaque extrémité. La figure 5.2 illustre ces résultats :

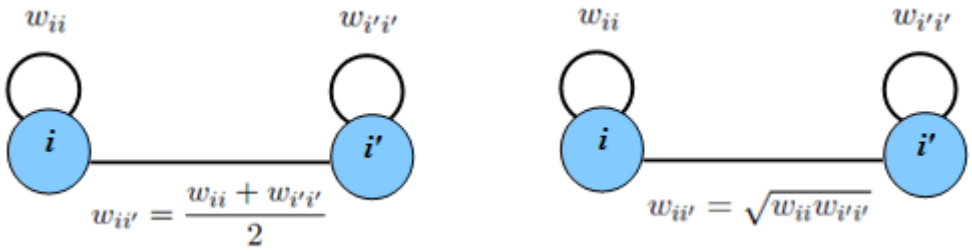

FIGURE 5.2 – A gauche le poids d'une arête dans le cas d'INDETERMINATION, à droite le poids d'une arête dans le cas d'INDEPENDANCE.

Nous allons étudier plus en détail la différence entre le critère de Newman-Girvan et l'Écart à l'Indétermination au chapitre 6 en nous basant sur la dualité entre ces deux principes.

Plus sur la dualité indépendance-indétermination

Maintenant focalisons-nous sur la dualité indépendance-indétermination quand on passe de l'environnement contingentiel vers l'environnement relationnel et vice versa. En se servant de la symétrie des matrices $\mathbf{A}$ et $\mathbf{X}$, en développant le membre droit de l'expression (5.14) et après regroupement et simplification des termes de même nature, celui-ci peut s'écrire sous la forme équivalente suivante :

$$\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'}-\frac{a_{i.}}{N}-\frac{a_{.i'}}{N}+\frac{a_{..}}{N^2}\right)\left(x_{ii'}-\frac{x_{i.}}{N}-\frac{x_{.i'}}{N}+\frac{x_{..}}{N^2}\right)=\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'}-\frac{a_{i.}}{N}-\frac{a_{.i'}}{N}+\frac{a_{..}}{N^2}\right)x_{ii'}. \tag{5.24}$$

Sous cette forme il est intéressant de remarquer que le membre droit de (5.24) est *formellement équivalent* au critère d'Écart à l'Indétermination proposé précédemment. Ainsi si la matrice d'adjacence du graphe à modulariser $\mathbf{A}$ représente une relation d'équivalence à p_A modalités [11] maximiser le critère d'Écart à l'Indétermination en notations relationnelles équivaut à maximiser l'écart à l'indépendance en notations contingentielles entre les relations d'équivalence $\mathbf{A}$ et $\mathbf{X}$:

$$F_{DI}=\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'}-\frac{a_{i.}}{N}-\frac{a_{.i'}}{N}+\frac{2M}{N^2}\right)x_{ii'}=\sum_{u=1}^{p_A}\sum_{v=1}^{\kappa}\left(n_{uv}-\frac{n_{u.}n_{.v}}{N}\right)^2, \tag{5.25}$$

car $a_{..}=\sum_{i=1}^{N}\sum_{i'=1}^{N}a_{ii'}=2M$.
La matrice $\mathbf{A}$ ne représentant pas une relation d'équivalence nous ne pouvons pas parler d'écriture contingentielle du critère. Cependant nous pouvons nous servir de l'expression (5.25) pour trouver une représentation du critère d'Écart à l'Indétermination dans un environnement *semi-contingentiel*. Un environnement qui croise non pas deux relations d'équivalence mais une relation binaire symétrique d'une part avec une relation d'équivalence d'autre part. Dans notre cas particulier, cette relation binaire symétrique sera représentée par l'espace des arêtes du graphe présentes dans la matrice d'adjacence et la relation d'équivalence sera la décomposition du graphe en classes d'équivalence telle que celle que nous cherchons à obtenir, ceci se faisant de la façon suivante :

On définit la matrice $\tilde{\mathbf{A}}$ comme $\tilde{\mathbf{A}}=\mathbf{A}+M\mathbf{I}_N$, où $\mathbf{I}_N$ est la matrice identité en dimension N.

On définit de plus la matrice $\tilde{\mathbf{A}}^j$ pour $j\in\{1,2,...,M\}$ comme la matrice de la relation d'équivalence à $(N-1)$ classes : 1 classe contenant les 2 sommets reliés par l'arête j et $(N-2)$ classes composées chacune d'un seul sommet, chacun des $(N-2)$ sommets restants.

Clairement la matrice $\tilde{\mathbf{A}}$ est la somme de M matrices représentant une relation d'équivalence, soit mathématiquement $\tilde{\mathbf{A}}=\sum_j^M\tilde{\mathbf{A}}^j$. A titre d'exemple, considérons le graphe suivant :
Dont la matrice d'adjacence est :

11. Ce qui n'est pas le cas car elle n'est pas transitive, sauf si le graphe est décomposé en cliques non connectées. Ce qui n'est pas le type de graphes qui font l'objet de notre étude, sinon le problème de modularisation n'aurait pas de sens.

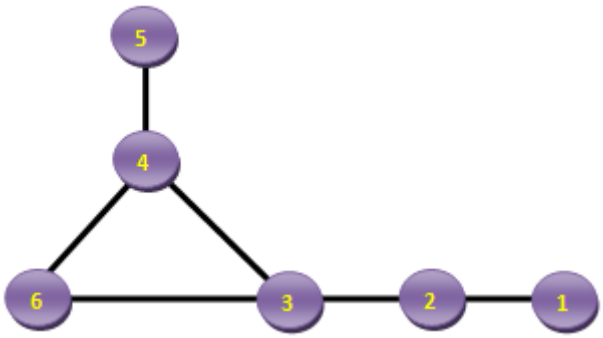

FIGURE 5.3 – Graphe non orienté à 6 sommets

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}$$

La matrice $\tilde{\mathbf{A}}$ et, par exemple, la matrice $\tilde{\mathbf{A}}^1$ de l'arête reliant les sommets 1 et 2 seront :

$$\tilde{\mathbf{A}} = \begin{pmatrix} 6 & 1 & 0 & 0 & 0 & 0 \\ 1 & 6 & 1 & 0 & 0 & 0 \\ 0 & 1 & 6 & 1 & 0 & 1 \\ 0 & 0 & 1 & 6 & 1 & 1 \\ 0 & 0 & 0 & 1 & 6 & 0 \\ 0 & 0 & 1 & 1 & 0 & 6 \end{pmatrix} \qquad \tilde{\mathbf{A}}^1 = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Les matrices $\tilde{\mathbf{A}}^j$ et $\mathbf{X}$ définissant chacune une relation d'équivalence sur les N sommets, elles admettent une notation contingentielle aussi :

$$\sum_{u=1}^{N-1}\sum_{v=1}^{\kappa}\left(n_{uv}^j - \frac{n_{u.}^j n_{.v}}{N}\right)^2 = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\tilde{a}_{ii'}^j - \frac{\tilde{a}_{i.}^j}{N} - \frac{\tilde{a}_{.i'}^j}{N} + \frac{(N+2)}{N^2}\right)x_{ii'}, \tag{5.26}$$

puisque $\forall j \sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}^j = (N+2)$. Comme l'expression (5.26) est vérifiée pour chaque arête, nous pouvons l'appliquer pour chacune et faire la somme pour les M arêtes :

$$\sum_{j=1}^{M}\sum_{u=1}^{N-1}\sum_{v=1}^{\kappa}\left(n_{uv}^j - \frac{n_{u.}^j n_{.v}}{N}\right)^2 = \sum_{j=1}^{M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\tilde{a}_{ii'}^j - \frac{\tilde{a}_{i.}^j}{N} - \frac{\tilde{a}_{.i'}^j}{N} + \frac{(N+2)}{N^2}\right)x_{ii'}. \tag{5.27}$$

Développons d'abord le membre gauche de cette dernière expression. Les arêtes peuvent être séparées en deux groupes : *coupées* et *non-coupées* par la partition $\mathbf{X}$. En notant *cut*

le nombre d'arêtes coupées, nous avons :

$$\sum_{u=1}^{N-1}\sum_{v=1}^{\kappa}\left[\left(\sum_{j=1}^{\text{cut}}\left(n_{uv}^{j}-\frac{n_{u.}^{j}n_{.v}}{N}\right)^{2}\right)+\left(\sum_{j=\text{cut}+1}^{M}\left(n_{uv}^{j}-\frac{n_{u.}^{j}n_{.v}}{N}\right)^{2}\right)\right]=$$

$$\sum_{u=1}^{N-1}\sum_{v=1}^{\kappa}\left[\sum_{j=1}^{\text{cut}}(n_{uv}^{j})^{2}+\sum_{j=\text{cut}+1}^{M}(n_{uv}^{j})^{2}-\frac{2}{N}\left(\sum_{j=1}^{\text{cut}}n_{uv}^{j}n_{u.}^{j}n_{.v}+\sum_{j=\text{cut}+1}^{M}n_{uv}^{j}n_{u.}^{j}n_{.v}\right)+\sum_{j=1}^{M}\frac{(n_{u.}^{j})^{2}n_{.v}^{2}}{N^{2}}\right]. \tag{5.28}$$

Les termes généraux des tableaux de contingence croisant les relations $\tilde{\mathbf{A}}^{j}$ et $\mathbf{X}$ ont des caractéristiques particulières selon que l'arête j ait été coupée ou non-coupée. Tout d'abord, concernant les marginales, nous avons $n_{u.}^{j}=\{1,...,1,2,1,...,1\}$ car les variables $\tilde{\mathbf{A}}^{j}$ possèdent une classe à 2 sommets et les autres classes constituent des sommets isolés. De ce fait deux cas peuvent se présenter :

1. Si l'arête j est coupée tous les termes du tableau de contingence croisant $\tilde{\mathbf{A}}^{j}$ et $\mathbf{X}$ sont soit égaux à l'unité soit nuls, tout en respectant les marginales. Ceci s'interprète comme le fait que les deux extrémités de l'arête j se trouvent dans deux classes différentes.

2. Si l'arête j est non-coupée tous les termes du tableau de contingence croisant $\tilde{\mathbf{A}}^{j}$ et $\mathbf{X}$ sont soit égaux à l'unité ou nuls et il existe un seul terme égal à 2. Les deux sommets à chaque extrémité de l'arête j.

A titre d'exemple, considérons une partition du graphe de la figure 5.3 en 2 classes, une $\mathcal{C}_1=\{1,2\}$ contenant les sommets 1 et 2 et une deuxième contenant le reste de sommets $\mathcal{C}_2=\{3,4,5,6\}$. Il existe donc une seule coupure, celle de l'arête reliant les sommets 2 et 3. En numérotant les arêtes par ordre croissant selon le numéro des sommets auxquels elles sont attachées, les tableaux de contingence entre la relation d'équivalence décrite par l'arête j, soit $\tilde{\mathbf{A}}^{j}$ et la partition $\mathbf{X}$ notés $\tilde{\mathbf{A}}^{j}$ respectivement :

$$\tilde{\mathbf{B}}^{1}=\begin{pmatrix}2&0\\0&1\\0&1\\0&1\\0&1\end{pmatrix}\quad\tilde{\mathbf{B}}^{2}=\begin{pmatrix}1&0\\1&1\\0&1\\0&1\\0&1\end{pmatrix}\quad\tilde{\mathbf{B}}^{3}=\tilde{\mathbf{B}}^{4}=\tilde{\mathbf{B}}^{5}=\tilde{\mathbf{B}}^{6}=\begin{pmatrix}1&0\\1&0\\0&1\\0&2\\0&1\end{pmatrix}$$

Clairement le tableau $\tilde{\mathbf{B}}^{2}$ correspond à l'arête coupée.

En tenant compte de ces remarques les termes de la dernière expression de (5.28) se simplifient de la façon suivante :

- Lorsque l'arête j est coupée, tous les termes de son tableau de contingence étant nuls ou égaux à l'unité, nous avons : $(n_{uv}^{j})^{2}=n_{uv}^{j}$ et alors $\sum_{j=1}^{\text{cut}}\sum_{u=1}^{N-1}\sum_{v=1}^{\kappa}(n_{uv}^{j})^{2}=\sum_{j=1}^{\text{cut}}N=N\text{cut}$.

- Lorsque j n'est pas coupée, le tableau de contingence $\tilde{\mathbf{B}}^j$ possède $(N-2)$ termes égaux à l'unité et un terme égal à 2, donc la somme des carrés des termes vaut $(N+2)$, de ce fait nous avons : $\sum_{u=1}^{N-1}\sum_{j=\text{cut}+1}^{M}\sum_{v=1}^{\kappa}(n_{uv}^j)^2 = \sum_{j=\text{cut}+1}^{M}(N+2) = (M-\text{cut})(N+2)$.

- Compte tenu de la structure du tableau de contingence $\tilde{\mathbf{B}}^j$ lorsque j est coupée, le terme $\sum_{u=1}^{N-1}\sum_{v=1}^{\kappa} n_{uv}^j n_{u.}^j n_{.v} = \sum_{v=1}^{\kappa} n_{.v}^2 + (n^{j1}+n^{j2})$ où n^{j1} et n^{j2} représente l'effectif des classes contenant les deux sommets se trouvant à chaque extrémité de l'arête j.
- De façon analogue au cas précédent, lorsque l'arête j n'est pas coupée, étant donné que les sommets se trouvant à ses deux extrémités sont dans la même classe, nous avons : $\sum_{u=1}^{N-1}\sum_{v=1}^{\kappa} n_{uv}^j n_{u.}^j n_{.v} = \sum_{v=1}^{\kappa} n_{.v}^2 + 2n^j$ où n^j dénote la taille de la classe contenant l'arête j.

- nous avons toujours $\sum_{u=1}^{(N-1)} n_{u.}^2 = (N+2)$.

Par conséquence nous obtenons que (5.28) s'exprime par :

$$\text{cut}N+(M-\text{cut})(N+2)-\frac{2}{N}M\sum_{v=1}^{\kappa}n_{.v}^2-\frac{2}{N}\left(\sum_{j=1}^{\text{cut}}(n^{j1}+n^{j2})+2\sum_{j=\text{cut}+1}^{M}n^j\right)+\frac{M}{N^2}(N+2)\sum_{v=1}^{\kappa}n_{.v}^2,$$

expression qui peut encore se simplifier de la façon suivante :

$$TG = MN + 2(M-\text{cut}) + \left(\frac{2M}{N^2}-\frac{M}{N}\right)\sum_{v=1}^{\kappa}n_{.v}^2 - \frac{2}{N}\left(\sum_{j=1}^{\text{cut}}(n^{j1}+n^{j2})+2\sum_{j=\text{cut}+1}^{M}n^j\right), \tag{5.29}$$

où TG signifie terme gauche, maintenant développons le terme droit TD de (5.27) :

$$TD = \sum_{i=1}^{N}\sum_{i'=1}^{N}x_{ii'}\sum_{j=1}^{M}\left(\tilde{a}_{ii'}^j - \frac{\tilde{a}_{i.}^j}{N} - \frac{\tilde{a}_{.i'}^j}{N} + \frac{(N+2)}{N^2}\right).$$

En nous servant de la définition de la matrice $\tilde{\mathbf{A}}$ et de l'égalité suivante :

$$\tilde{\mathbf{A}} = \textstyle\sum_j^M \tilde{\mathbf{A}}^j = \mathbf{A} + M\mathbf{I}.$$

Nous déduisons les propriétés suivantes :

- Le terme général de $\tilde{\mathbf{A}}$ vaut $\tilde{a}_{ii'} = \sum_j^M \tilde{a}_{ii'}^j = \begin{cases} a_{ii'} & \text{si} \quad i \neq i', \\ M & \text{sinon.} \end{cases}$

- A partir du résultat précédent : $\tilde{a}_{i.} = \tilde{a}_{.i} = \sum_{i'=1}^{N}\sum_j^M \tilde{a}_{ii'}^j = (a_{i.}+M)$.

Si on remplace dans notre expression TD nous obtenons :

$$TD = \sum_{i \neq i'}^{N} a_{ii'} x_{ii'} + \sum_{i=1}^{N} M x_{ii} - \sum_{i=1}^{N}\sum_{i'=1}^{N} \frac{(a_{i.} + a_{.i'} + 2M) x_{ii'}}{N} + \sum_{i=1}^{N}\sum_{i'=1}^{N} \frac{M(N+2)}{N^2} x_{ii'}.$$

Après développement et en tenant compte que $a_{ii} = 0 \quad \forall i$, nous obtenons l'expression suivante :

$$TD = \sum_{i=1}^{N}\sum_{i'=1}^{N} \left(a_{ii'} - \frac{(a_{i.} + a_{.i'})}{N} + \frac{2M}{N^2} \right) x_{ii'} + MN - \frac{M}{N} \sum_{i=1}^{N}\sum_{i'=1}^{N} x_{ii'}. \tag{5.30}$$

Maintenant en égalant $TG = TD$ (expressions (5.29) et(5.30)) et après simplification de termes similaires et en considérant que $\sum_{v=1}^{\kappa} n_{.v}^2 = \sum_{i=1}^{N}\sum_{i'=1}^{N} x_{ii'}$ (voir les formules de transfert, annexe A) nous obtenons :

$$2(M - \text{cut}) - \frac{2}{N} \sum_{j=1}^{M} (n^{j1} + n^{j2}) + \frac{2M}{N^2} \sum_{v=1}^{\kappa} n_{.v}^2 = \sum_{i=1}^{N}\sum_{i'=1}^{N} \left(a_{ii'} - \frac{(a_{i.} + a_{.i'})}{N} + \frac{2M}{N^2} \right) x_{ii'}. \tag{5.31}$$

A partir de cette dernière expression nous pouvons déduire l'expression de critère d'Écart à l'Indétermination (5.18) dans un espace *semi-contingentiel* qui croise une relation binaire symétrique [12] avec une relation d'équivalence, qui est fonction des tailles des classes de la partition cherchée et d'arêtes :

$$F_{DI} = 2(M - \text{cut}) - \frac{2}{N} \sum_{j=1}^{M} (n^{j1} + n^{j2}) + \frac{2M}{N^2} \sum_{v=1}^{\kappa} n_{.v}^2. \tag{5.32}$$

A partir de cette expression nous pouvons déduire les formules de transfert suivantes :

Notation Relationnelle ⇔ Semi-Contingentielle	**Interprétation**
$\frac{1}{2}\sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'} x_{ii'} = (M - \text{cut})$	Arêtes intra-classe.
$\frac{1}{2}\sum_{i}^{N}\sum_{i'}^{N} a_{ii'} \bar{x}_{ii'} = \text{cut}$	Arêtes inter-classes
$\sum_{i=1}^{N}\sum_{i'=1}^{N} x_{ii'} = \sum_{v=1}^{\kappa} n_{.v}^2$	Somme des carrés de tailles de classes
$\sum_{i=1}^{N}\sum_{i'=1}^{N} a_{i.} x_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N} a_{.i'} x_{ii'} = \sum_{j=1}^{M} (n^{j1} + n^{j2})$	Tailles de classes contenant chaque extrémité d'arête

L'écriture (5.32) permet aussi d'écrire le critère d'Écart à l'Indétermination en fonction de la matrice $\mathbf{e}$ définie précédemment et utilisée dans la définition originale du critère de Newman-Girvan (5.8) :

12. Il est important de rappeler que ce résultat est valable pour un graphe non-pondéré, non-orienté et non-réflexif.

$$F_{DI}(\mathbf{e}) = Tr(\mathbf{e}) - \frac{1}{N}\sum_{j}^{\kappa}\sum_{j'}^{\kappa}[\mathbf{e}\mathbf{D}_\kappa + \mathbf{D}_\kappa\mathbf{e}]_{jj'} + \frac{2M}{N^2}\sum_{j}^{\kappa}\sum_{j'}^{\kappa}[\mathbf{D}_\kappa^2]_{jj'}, \tag{5.33}$$

où la matrice $\mathbf{D}_\kappa = \text{diag}(n_1, \ldots, n_j, \ldots, n_\kappa)$ est une matrice diagonale $\kappa \times \kappa$ contenant les tailles de classes de $\mathbf{X}$ sur la diagonale principale.

5.2.6 Le critère d'Écart à l'Uniformité

Ce critère a une construction voisine au critère de Newman-Girvan. Il maximise l'écart entre le nombre d'arêtes intra-classe et sa version aléatoire dans le cas où tous les sommets possèderaient le même degré, à savoir le degré moyen du graphe. Un critère se basant sur le même principe sous le nom de *critère d'écart à la moyenne* a été déjà introduit dans le cadre de classification automatique par Chah [1983]. Cette fois la version aléatoire correspond à un graphe où les arêtes sont distribuées selon une loi uniforme et il n'existe pas de contraintes concernant le respect du degré de chaque sommet.

Le cas hypothétique où tous les sommets sont repartis selon une loi uniforme, correspond au cas où nous avons une absence totale d'information par rapport à la situation relative à la distribution d'arêtes, mis à part le nombre total d'arêtes lui-même. Un exemple de graphe suivant une telle distribution des arêtes est ce qu'on appelle un "grid lattice", c'est un graphe qui ne possède aucune structure communautaire, disons un graphe en forme de grille ou treillis (lattice graph en anglais) :

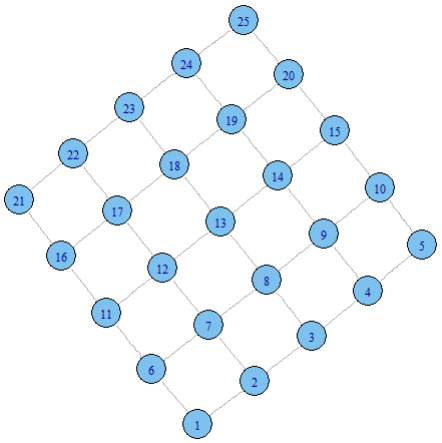

FIGURE 5.4 – Graphes dont tous les sommets ont le même degré.

Le critère d'Écart à l'Uniformité s'écrit alors [13] :

13. La construction du graphe suivant la structure d'uniformité, est analogue à la solution du plus simple PSIS (Program of Spatial Interaction System) en absence de contraintes. En effet, quand on n'a aucune connaissance concernant la distribution des degrés, celle ci vérifie le principe de Laplace dit "d'indifférence" (principle of insufficient reason) : The likelihood of the diverse states of the Nature is identical because we do not know any information as for their relative likelihood and to consider that the world trade is uniformly distributed inside the system. La distribution uniforme correspond au plus haut degré de désordre qu'un système peut avoir selon l'entropie de Boltzmann and Shannon.

$$F_{\text{Unif}}(X) = \frac{1}{2M} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(a_{ii'} - \frac{2M}{N^2} \right) x_{ii'}, \tag{5.34}$$

où la variable $\mathbf{X}$ doit vérifier les contraintes d'une relation d'équivalence (5.2).

La quantité $\delta = \frac{2M}{N^2}$ est précisément la densité ou taux d'occupation d'arêtes, toujours comprise entre 0 à 1, définie au chapitre 1 et dont quelques valeurs empiriques pour des graphes réels ont été données dans le tableau 1.1.

Ce critère est linéaire, séparable, et équilibré globalement car il s'agit d'un **modèle nul** car $\sum_{i=1}^{N} \sum_{i'=1}^{N} a_{ii'} = \sum_{i=1}^{N} \sum_{i'=1}^{N} \frac{2M}{N^2} = 2M$. Il possède donc, une limite de résolution.

Comme nous avons vu le critère de Zahn-Condorcet n'est pas un modèle nul. Cependant lorsque $\alpha = \delta$ le critère d'Owsiński-Zadrożny pondéré le terme d'accords positifs et le terme d'accords négatifs du critère de Zahn-Condorcet de façon à rendre le critère résultant un modèle nul, en effet :

$$\sum_{i=1}^{N} \sum_{i'=1}^{N} (1-\delta) a_{ii'} = \sum_{i=1}^{N} \sum_{i'=1}^{N} \delta \bar{a}_{ii'} = \left(2M - \frac{4M^2}{N^2} \right)$$

.

Le critère résultant sera :

$$F_{OZ,\alpha=\delta}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left((1-\delta) a_{ii'} x_{ii'} + \delta \bar{a}_{ii'} \bar{x}_{ii'} \right). \tag{5.35}$$

En réécrivant l'expression (5.35) nous retrouvons l'Écart à l'Uniformité à une constante près :

$$F_{OZ,\alpha=\delta}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} (a_{ii'} - \delta) x_{ii'} + K$$

.

5.2.7 Le critère de *Correlation clustering* de Demaine et Immorlica (2002)

Le problème dit de "correlation clustering" fut posé initialement par N. Bansal, A. Blum, et S. Chawla dans Bansal et al. [2002] :"Étant donné un graphe complet d'ordre N où chaque arête possède soit une étiquette + si ses sommets sont considérés comme similaires soit une étiquette − si ses sommets sont considérés comme différents". Le but est de trouver une partition qui :

- Soit maximise le nombre d'accords : nombre d'arêtes + intra-classe ainsi que le nombre d'arêtes − inter-classes.
- Soit minimise le nombre de désaccords : nombre d'arêtes + inter-classes ainsi que le nombre d'arêtes − intra-classe.

Étant donné un graphe $G = (V, E)$ pondéré avec des poids réels, i.e. $w_{ii'} \in \mathbb{R}$ (à la fois positifs et négatifs), le but est de trouver une partition des sommets de façon à minimiser les arêtes à poids positif inter-classes et les arêtes à poids négatif intra-classe [14]. Les grands poids positifs représentent une forte corrélation entre les points extrêmes alors que les grands poids négatifs représentent une forte répulsion, et les poids à valeur absolue proche de zéro représentent peu d'information. Pour résoudre ce problème, dans Demaine and Immorlica [2003] (voir aussi Demaine et al. [2006]) les auteurs proposent la fonction de coût à minimiser suivante :

$$F_{CC}(\mathbb{P}) = cost(\mathbb{P}) = cost_p(\mathbb{P}) + cost_m(\mathbb{P}) \tag{5.36}$$

(nous notons cette fonction F_{CC} pour *correlation clustering*),
où :

- $\mathbb{P}$ est une partition de V : $\mathbb{P} = \{C_1, C_2, ..., C_\kappa\}$.
- $cost(\mathbb{P})$ coût total de la partition $\mathbb{P}$.
- $cost_p(\mathbb{P}) = \sum\{|w_{ii'}| : (i, i') \in E; w_{ii'} > 0; \forall j, |\{i, i'\} \cap C_j| \leq 1\}$; soit la somme des poids positifs entre deux sommets qui ne sont pas dans la même classe.
- $cost_m(\mathbb{P}) = \sum_{\{} |w_{ii'}| : (i, i') \in E; w_{ii'} < 0; \exists j, |\{i, i'\} \cap C_j| = 2\}$; soit la somme des poids négatifs entre deux sommets qui sont dans la même classe.

Soit $\mathbf{Y}$ une matrice d'ordre N représentant la variable relationnelle dont le terme général est défini de la façon suivante :

$$y_{ii'} : \begin{cases} 1 \text{ si } i \text{ et } i' \text{ ne sont pas dans la même classe,} \\ 0 \text{ si } i \text{ et } i' \text{ sont dans la même classe.} \end{cases} \tag{5.37}$$

Clairement la variable $Y_{ii'} = \bar{X}_{ii'}, \forall i, i' \in V$. En notations relationnelles les deux termes de l'équation (5.36) deviennent :

$$cost_p(Y) = \frac{1}{2} \sum_{i,i'} w_{ii'} y_{ii'} \mathbb{1}_{(w_{ii'}>0)} \tag{5.38}$$

$$cost_m(Y) = \frac{1}{2} \sum_{i,i'} |w_{ii'}|(1 - y_{ii'}) \mathbb{1}_{(w_{ii'}<0)}, \tag{5.39}$$

où :

- Le coefficient 1/2 vient du fait que l'on somme les poids deux fois.
- $\mathbb{1}_Y$ est la fonction indicatrice de l'ensemble Y.
- Le terme $cost_p(\mathbb{S})$ représente la somme des poids positifs inter-classes.
- Le terme $cost_m(\mathbb{S})$ correspond à la somme des poids négatifs intra-classe.

Avec ces notations le coût total à minimiser sera :

14. Ou à maximiser les arêtes à poids positif intra-classe et les arêtes à poids négatif inter-classes

$$\begin{aligned} F_{CC}(Y) &= \sum_{i,i'} w_{ii'} y_{ii'} \mathbb{1}_{(w_{ii'}>0)} + \sum_{i,i'} |w_{ii'}|(1-y_{ii'})\mathbb{1}_{(w_{ii'}<0)} \\ &= \sum_{i,i'} w_{ii'} y_{ii'} \mathbb{1}_{(w_{ii'}>0)} + \sum_{i,i'} |w_{ii'}| \bar{y}_{ii'} \mathbb{1}_{(w_{ii'}<0)} \\ &= \sum_{i,i'} w^+_{ii'} y_{ii'} + \sum_{i,i'} w^-_{ii'} \bar{y}_{ii'}, \end{aligned} \tag{5.40}$$

avec :
- $w^+_{ii'} = w_{ii'} \mathbb{1}_{(w_{ii'}>0)}$.
- $w^-_{ii'} = |w_{ii'}| \mathbb{1}_{(w_{ii'}<0)}$.

L'équation (5.40) est la forme duale de Condorcet présentée dans l'équation (2.14). En effet, en remplaçant $\mathbf{Y}$ par $\bar{\mathbf{X}}$ et $\bar{\mathbf{Y}}$ par $\mathbf{X}$:

$$F_{CC}(X) = \sum_{i,i'} w^+_{ii'} \bar{x}_{ii'} + \sum_{i,i'} w^-_{ii'} x_{ii'}. \tag{5.41}$$

L'équation (5.41) est une formulation très voisine du critère "Dual de Condorcet" pour un graphe pondéré avec des poids réels. Cette réécriture a déjà été proposée dans les travaux de Labiod [2008].

L'expression (5.41) montre que ce critère est linéaire, séparable il possède la propriété d'équilibre général. En revanche, le fait qu'il soit équilibré localement ou globalement dépendra des valeurs prises par les poids $w^+_{ii'}$ et $w^-_{ii'}$.

L'expression (5.40) peut encore être simplifiée :

$$\begin{aligned} F_{CC}(Y) &= \sum_{i,i'} w_{ii'} y_{ii'} \mathbb{1}_{(w_{ii'}>0)} + \sum_{i,i'} |w_{ii'}|(1-y_{ii'})\mathbb{1}_{(w_{ii'}<0)} \\ &= \sum_{i,i'} w_{ii'} y_{ii'} \mathbb{1}_{(w_{ii'}>0)} - \sum_{i,i'} w_{ii'}(1-y_{ii'})\mathbb{1}_{(w_{ii'}<0)} \\ &= \sum_{i,i'} w_{ii'} y_{ii'} \mathbb{1}_{(w_{ii'}>0)} - \sum_{i,i'} w_{ii'} \mathbb{1}_{(w_{ii'}<0)} + \sum_{i,i'} w_{ii'} y_{ii'} \mathbb{1}_{(w_{ii'}<0)}. \end{aligned}$$

Le terme $\sum_{i,i'} w_{ii'} \mathbb{1}_{(w_{ii'}<0)}$ étant une constante il peut être enlevé de la fonction coût, ainsi la fonction à minimiser devient :

$$F_{CC}(Y) = \sum_{i,i'} w_{ii'} y_{ii'} \mathbb{1}_{(w_{ii'}>0)} + \sum_{i,i'} w_{ii'} y_{ii'} \mathbb{1}_{(w_{ii'}<0)} = \sum_{i,i'} w_{ii'} y_{ii'}. \tag{5.42}$$

Nous allons montrer que cette expression est une écriture très proche de l'expression (2.7) :

$$\begin{aligned} F_{CC}(Y) = \sum_{i,i'} w_{ii'} y_{ii'} = \sum_{ii'} (w^+_{ii'} - w^-_{ii'}) y_{ii'} &= \sum_{ii'} (w^+_{ii'} - w^-_{ii'}) \bar{x}_{ii'} \\ &= \sum_{ii'} (w^+_{ii'} - w^-_{ii'})(1 - x_{ii'}) \end{aligned}$$

Ce qui équivaut à maximiser l'expression :

$$F_{CC}(X) = \sum_{ii'} (w^+_{ii'} - w^-_{ii'}) x_{ii'}.$$

Cette expression n'est autre que l'expression du critère de Condorcet (2.7) avec : $c_{ii'} = w^+_{ii'}$ et $\bar{c}_{ii'} = w^-_{ii'}$.

5.2.8 Le critère de Condorcet pondéré en A (1991)

Ce critère a été introduit pour la première fois dans Marcotorchino [1991] afin de faire la liaison entre l'Analyse Relationnelle et Analyse Factorielle. Ce critère cherche à maximiser l'expression suivante :

$$F_{CPond}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(\hat{a}_{ii'} x_{ii'} + \bar{\hat{a}}_{ii'} \bar{x}_{ii'} \right), \tag{5.43}$$

où $\hat{a}_{ii'}$ et $\bar{\hat{a}}_{ii'}$ sont respectivement définis au travers des équations (2.15) et (2.19).

Pour garantir que l'optimisation du critère (5.43) permette d'obtenir une partition, $\mathbf{X}$ doit vérifier les contraintes d'une relation d'équivalence, énoncées dans (5.2).

Compte tenu de la définition de $\bar{\hat{\mathbf{A}}}$ (équation (2.19)), maximiser l'expression (5.43) revient à maximiser l'expression :

$$F_{CPond}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(2\hat{a}_{ii'} - \frac{\hat{a}_{ii} + \hat{a}_{i'i'}}{2} \right) x_{ii'}. \tag{5.44}$$

L'écriture relationnelle (5.44) met en évidence la linéarité de ce critère et par conséquent sa séparabilité. À partir de l'expression (5.44) nous déduisons que le critère est possède la propriété d'équilibre général si sa matrice d'adjacence vérifie la condition : $\sum_{i=1}^{N} \hat{a}_{ii} > 0 \Leftrightarrow \sum_{i=1}^{N} \frac{a_{ii}}{a_{i.}} > 0$. Le degré de chaque sommet étant toujours strictement positif, cette condition implique que le graphe doit être réflexif, i.e. les sommets doivent avoir des boucles. Comme mentionné dans la définition de la propriété d'équilibre linéaire (voir chapitre 4) le non-respect de cette condition a pour conséquence l'obtention de la solution grossière où tous les sommets sont classés dans une seule classe. Ainsi, l'utilisation de ce critère se restreint aux graphes non pondérés et réflexifs. Comme nous le verrons au chapitre 7, si le graphe n'est pas réflexif nous lui rendrons réflexif en ajoutant des boucles sur chaque sommet avant d'employer ce critère.

Ce critère vérifie aussi la propriété fondamentale de la métrique du χ^2, à savoir : l'*équivalence Distributionnelle*. La solution optimale de ce critère n'est pas triviale et est obtenue sans fixer le nombre de classes de la partition cherchée, comme dans le contexte du Critère de Condorcet.

Critère	**Notation Relationnelle**
Zahn-Condorcet (1964,1785)	$F_{ZC}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(a_{ii'}x_{ii'} + \bar{a}_{ii'}\bar{x}_{ii'})$
Owsiński-Zadrożny(1986)	$F_{OZ}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}((1-\alpha)a_{ii'}x_{ii'} + \alpha\bar{a}_{ii'}\bar{x}_{ii'})$ avec $0 < \alpha < 1$
Condorcet pondéré en **A** (1991)	$F_{CPond}(X, \hat{A}) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(\hat{a}_{ii'}x_{ii'} + \bar{\hat{a}}_{ii'}\bar{x}_{ii'})$
Demaine-Immorlica (2002)	$F_{CC}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(w^{+}_{ii'}x_{ii'} + w^{-}_{ii'}\bar{x}_{ii'})$
Newman-Girvan (2004)	$F_{NG}(X) = \frac{1}{2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{a_{i.}a_{.i'}}{2M}\right)x_{ii'}$
Modularité Équilibrée (2013)	$F_{BM}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}((a_{ii'} - p_{ii'})x_{ii'} + (\bar{a}_{ii'} - \bar{p}_{ii'})\bar{x}_{ii'})$ avec $p_{ii'} = \frac{a_{i.}a_{.i'}}{2M}$ et $\bar{p}_{ii'} = \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M}$
Écart à l'Indétermination (2013)	$F_{DI}(X) = \frac{1}{2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{a_{i.}}{N} - \frac{a_{.i'}}{N} + \frac{2M}{N^2}\right)x_{ii'}$
Écart à l'Uniformité (2013)	$F_{\text{Unif}}(X) = \frac{1}{2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{2M}{N^2}\right)x_{ii'}$

Table 5.2 – Critères linéaires de modularisation en notations relationnelles.

5.3 Les critères séparables de fonctions non-linéaires de X

Le Tableau 5.3 montre les critères séparables de fonctions non-linéaires de **X** étudiés dans cette section.

5.3.1 Le critère de Mancoridis-Gansner (1998)

Le critère de Mancoridis-Gansner (voir Mancoridis et al. [1998]) représente une fonction économique d'une partition en classes disjointes du graphe qu'il s'agit de maximiser et qui consiste à maximiser les connexions intra-classe tout en minimisant les connexions inter-classes.

Dans Mancoridis et al. [1998], les auteurs ont proposé un modèle d'optimisation visant à récupérer automatiquement la structure modulaire d'un programme informatique à partir de son code source (puisque ceci était en fait le sujet premier des auteurs, qui travaillaient dans le domaine de la programmation dite de "cluster-programming") .

Leur but était de comprendre la structure des composantes du programme surtout pour les logiciels les plus utilisés pour lesquels, du fait qu'ils contiennent plusieurs centaines de milliers de lignes de code encapsulées dans plusieurs modules, il est indispensable de trouver les structures de décomposition du code en éléments plus simples et plus indépendants les uns des autres. Ceci permet aux programmeurs de faire face à la problématique de

l'interprétation des lignes de code au travers de l'identification des regroupements (clustering) de procédures connexes en "modules" (ou classes) plus indépendantes les unes des autres.

Le modèle de S. Mancoridis et Y. Gansner crée une vue hiérarchique de l'organisation du système, basée uniquement sur les composantes et les relations qui existent dans le code source. Dans un premier temps, leur modèle représente les composantes du système et les relations entre elles comme un graphe de dépendances. Les composantes comprennent les classes, les variables, les macros et les structures de données [15] ; tandis que les relations courantes comprennent l'importation, l'exportation, l'héritage ou l'appel à une procédure [16]. Ensuite, une fois modélisée leur problématique sous forme relationnelle, ils ont proposé un critère et des algorithmes ad-hoc pour partitionner le grand graphe du code ainsi obtenu.

Le but étant la modularisation automatique des composantes du programme en groupes (classes), le critère de S. Mancoridis et Y. Gansner se base sur un compromis entre l'inter et l'intra-connectivité. Ce critère minimise simultanément l'inter-connectivité (les connexions entre les composantes de deux groupes distincts), tout en maximisant l'intra-connectivité (les connexions entre les composantes d'une même classe).

1. **Intra-connectivité** : elle mesure la connectivité entre deux composantes d'une même classe. Une valeur d'intra-connectivité élevée indique que les composantes à l'intérieur de la classe sont fortement connectées, la modification du code d'une composante affectera alors principalement les autres membres de la classe. L'expression de l'intra-connectivité [17] A_j de la classe j à N_j composantes et m_j arcs intra-classe est :

$$A_j = \frac{m_j}{N_j^2}. \tag{5.45}$$

   Cette mesure représente la fraction du nombre d'arcs existants dans la classe j par rapport au nombre maximal d'arcs qui pourraient exister [18], soit N_j^2. Cette dernière quantité est calculée pour un graphe orienté avec des boucles, donc le nombre maximal d'arêtes est la taille de la classe au carré. La valeur d'intra-connectivité est comprise entre 0 et 1. Plus grande est la valeur de l'intra-connectivité plus proche est la classe d'un graphe complet.

2. **Inter-connectivité** : est une mesure de connectivité entre deux groupes ou classes différents. Une valeur d'interconnectivité faible indique que les groupes ou classes sont, dans une large mesure, indépendants. Par conséquent, les modifications faites au code d'une composante affecteront de façon négligeable les composantes d'une

15. Ici les mots : classes, variables, macros et structures de données sont employés du point de vue informatique.

16. Du point de vue informatique.

17. A ne pas confondre avec la matrice d'adjacence du graphe **A**.

18. Cette valeur correspond au nombre maximal d'arêtes qui peuvent exister dans un graphe orienté et réflexif. En effet, dans Mancoridis et al. [1998] les auteurs ont formulé le critère pour un graphe orienté avec des boucles. Cependant dans Delest et al. [2006], les auteurs ont repris ce critère et ils ont modifié le calcul de l'intra-connectivité à un graphe non-orienté et non-réflexif, soit $A_j = \frac{m_j}{\binom{N_j}{2}} = \frac{m_j}{\frac{N_j(N_j-1)}{2}}$
Dans ce document nous nous traiterons seulement la version originale du critère de Mancoridis-Gansner.

autre classe. L'expression de l'inter-connectivité $E_{jj'}$ entre les classes j et j' de taille N_j et N'_j respectivement et $\epsilon_{jj'}$ arcs inter-classes est donnée par :

$$E_{jj'} = \begin{cases} 0 & \text{si } j = j', \\ \frac{\epsilon_{jj'}}{2N_jN_{j'}} & \text{si } j \neq j'. \end{cases} \tag{5.46}$$

Cette mesure est la fraction du nombre d'arcs existants entre les sommets de la classe j et les sommets de la classe j' par rapport au nombre maximal d'arcs qui peuvent exister entre ces deux classes, soit $2N_jN_{j'}$. La valeur 2 vient du fait que pour un graphe orienté il peut y avoir deux arcs reliant la même paire de sommets. L'inter-connectivité est comprise entre 0 et 1. Dans le cas idéal $E_{jj'}$ est nulle, il n'existe donc aucun lien entre les composantes de la classe j et les composantes de la classe j'.

La fonction objectif qui maximise à la fois les connexions intra-classe et minimise les connexions inter-classes afin d'obtenir une partition en κ classes de l'ensemble de composantes s'écrit :

$$F_{MG} = \frac{1}{\kappa}\sum_{j=1}^{\kappa} A_j - \frac{1}{\frac{\kappa(\kappa-1)}{2}} \sum_{j,j'=1}^{\kappa} E_{jj'} \quad \text{si } \kappa > 1. \tag{5.47}$$

Le premier terme de l'équation (5.47) représente la moyenne de l'intra-connectivité des κ classes. Le deuxième terme représente la moyenne d'inter-connectivité entre toutes les paires distantes des classes, soit $\frac{\kappa(\kappa-1)}{2}$. On peut également maximiser l'opposé de l'inter-connectivité ce qui revient à la minimiser. Les valeurs de ce critère vont de -1 (aucune connexion intra-classe) à 1 (aucune connexion entre deux classes distinctes).

À partir de l'équation (2.11) de la variable relationnelle $\mathbf{X}$, il est possible de réécrire l'équation (5.47) sous la forme simplifiée suivante :

$$F_{MG} = \frac{1}{\kappa}\sum_{j=1}^{\kappa} A_j - \frac{1}{\frac{\kappa(\kappa-1)}{2}} \sum_{j,j'=1}^{\kappa} E_{jj'} \quad \text{si } \kappa > 1,$$

avec la possibilité de simplifier encore cette expression en utilisant les notations relationnelles :

1. Compte tenu de l'équation (2.11), la sommation du premier terme de (5.47) s'exprime comme : $\sum_i^N \sum_{i'}^N \frac{a_{ii'}x_{ii'}}{x_{i.}x_{.i'}}$.
2. Compte tenu de l'équation (2.13), la sommation du deuxième terme de (5.47) s'exprime comme : $\sum_i^N \sum_{i'}^N \frac{a_{ii'}\bar{x}_{ii'}}{2x_{i.}x_{.i'}}$.

Où $a_{ii'}$ est le terme général de la matrice d'adjacence $\mathbf{A}$.

L'équation (5.47) s'écrit alors :

$$F_{MG} = \frac{1}{\kappa}\sum_{i}^{N}\sum_{i'}^{N} \frac{a_{ii'}x_{ii'}}{x_{i.}x_{.i'}} - \frac{1}{\kappa(\kappa-1)} \sum_{i}^{N}\sum_{i'}^{N} \frac{a_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}} \quad \text{si } \kappa > 1. \tag{5.48}$$

Dans l'équation (5.48) il est important de voir que le nombre de classes κ a disparu comme borne indicielle des sommations. En effet, la définition de la variable relationnelle

$\mathbf{X}$ fait que, dans le cas du premier terme l'on somme tous les liens internes de chaque classe (intra-classe) ; de façon analogue, pour le deuxième terme la définition de $\bar{\mathbf{X}}$ fait que l'on somme tous les liens externes à chaque classe (inter-classes). D'autre part le terme $x_{.i} = x_{i.} = \sum_{i'=1}^{N} x_{ii'} \forall i, i' \in V \times V$ représente la taille de la classe contenant i.

L'équation (5.48) nous montre que ce critère est non-linéaire, séparable et non-équilibré. Nous allons voir par la suite comment le rendre linéaire, non pas par rapport à $\mathbf{X}$ mais par rapport au terme général de la *Matrice de densité* ou *Matrice de taux d'occupation*, définie comme suit :

Définition 5.3. *[Matrice de densité ou de taux d'occupation :$\mathbf{U}$] Soit $\mathbf{U}$ une matrice carrée d'ordre N, de terme général :*

$$u_{ii'} = \frac{x_{ii'}}{x_{i.}x_{.i'}} \qquad \bar{u}_{ii'} = \frac{\bar{x}_{ii'}}{x_{i.}x_{.i'}}.$$

Lorsque les sommets d'un graphe sont numérotés de manière aléatoire, il n'existe apparemment aucune structure de communauté dans sa matrice d'adjacence. Cependant une fois qu'un processus de classification a été éffectué, si nous réorganisons les sommets en fonction de leurs classes respectives, nous obtenons une matrice d'adjacence diagonalisée par blocs, où chaque bloc est une classe d'équivalence. (Voir d'exemple dans la figure 5.5).

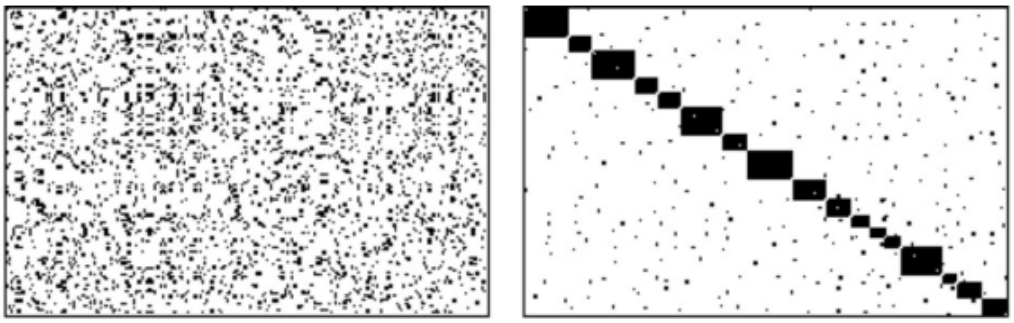

FIGURE 5.5 – À gauche : matrice d'adjacence d'un graphe à $N = 210$ sommets et $M = 1505$ arêtes. S'il existe un lien entre deux sommets il y a un point, sinon il y a un espace blanc. À droite : la matrice d'adjacence du graphe après avoir subi un processus de classification, et avoir ordonné les sommets en fonction de leurs classes respectives ; le résultat est une matrice diagonale par blocs où chaque bloc représente une classe d'équivalence, il y en a 17 au total, de tailles différentes et les connexions hors blocs sont dispersées (voir Schaeffer [2007]).

Dans le cas idéal où il n'existe pas de liens inter-classes et tous les éléments de chaque classe forment un graphe complet (i.e. il y en a un lien pour chaque paire d'éléments de la classe), on obtient une matrice diagonale par blocs, la matrice relationnelle $\mathbf{X}$. Chaque bloc de cette matrice est une sous-matrice carrée dont l'ordre est la taille de la classe que le bloc représente.

La quantité $u_{ii'} = \frac{x_{ii'}}{x_{i.}x_{.i'}}$ représente un taux d'occupation ou une densité. Si i et i' sont dans la même classe, le numérateur vaut 1 et le dénominateur est la *surface* du bloc (la classe) contenant i et i'. Ci-dessous nous montrons un exemple sur un graphe à 7 sommets.

La matrice d'adjacence du graphe de la figure 5.6 ainsi que la matrice relationnelle $\mathbf{X}$ sont présentées ci après. Après avoir effectué une classification en maximisant le critère de Zahn-Condorcet on trouve $\kappa = 3$ classes de taille respective 3, 2 et 2 :

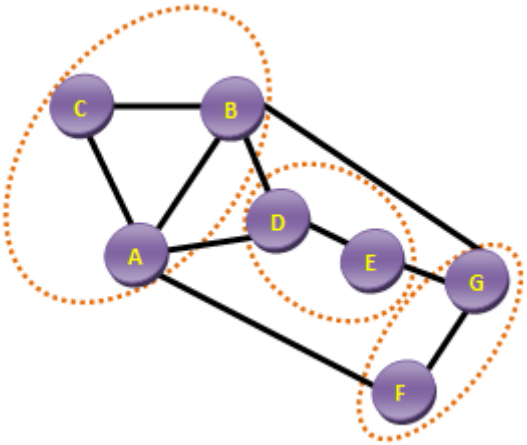

Figure 5.6 – Graphe à 7 sommets. Les lignes en pointillé délimitent les classes

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$\mathbf{U} = \begin{pmatrix} \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & 0 & 0 & 0 & 0 \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & 0 & 0 & 0 & 0 \\ \frac{1}{9} & \frac{1}{9} & \frac{1}{9} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} \end{pmatrix} \qquad \bar{\mathbf{U}} = \begin{pmatrix} 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & 0 & 0 & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{4} & \frac{1}{4} & 0 & 0 \end{pmatrix}$$

Quelques propriétés caractéristiques de la matrice $\mathbf{U}$:

1. Elle est symétrique.

   Cette propriété est une conséquence de la définition de la matrice $\boldsymbol{U}$. *En effet,* $u_{ii'} = \frac{x_{ii'}}{x_{i.}x_{.i'}} = \frac{x_{i'i}}{x_{.i'}x_{i.}} = \frac{x_{i'i}}{x_{i'.}x_{.i}} = u_{i'i}$ *Donc la matrice* $\boldsymbol{U}$ *est symétrique.*

2. $u_{ii'} \in \{0, 1, \frac{1}{4}, \frac{1}{9}, \ldots, \frac{1}{N^2}\} \qquad \forall i, i' \in V \times V.$

   Démonstration. Pour prouver cette propriété nous considérons deux cas possibles :
   - *Soit* i *et* i' *n'appartiennent pas à la même classe et dans ce cas-là* $x_{ii'} = 0$ *et par conséquent* $u_{ii'} = 0$.

- *Soit i et i' sont dans la même classe et dans ce cas-là $x_{ii'} = 1$ et $u_{ii'} = \frac{1}{x_{i.}x_{.i'}} = \frac{1}{x_{i.}^2}$ car $(x_{i.} = x_{.i'})$ et comme $x_{i.} \geq 1$ car il s'agit du cardinal de la classe contenant i, il vient $0 < u_{ii'} = \frac{1}{x_{i.}^2} \leq 1$.*

□

Quelques propriétés supplémentaires de la matrice $\mathbf{U}$:

3. La somme des éléments de chaque bloc vaut 1.

Démonstration. Supposons que l'on veuille sommer les éléments du bloc représentant une classe $\mathcal{C}$ quelconque :

$$\sum_{i,i' \in \mathcal{C}} u_{ii'} = \sum_{i,i' \in \mathcal{C}} \frac{x_{ii'}}{x_{i.}x_{.i'}} = \sum_{i=1}^{|\mathcal{C}|} \sum_{i'=1}^{|\mathcal{C}|} \frac{1}{|\mathcal{C}|^2} = 1.$$

□

4. Comme il y a κ classes, la somme des éléments de $\mathbf{U}$ vaut κ (conséquence de la propriété précédente), soit, $\sum_{i,i'} u_{ii'} = \kappa$. Il s'agit d'une propriété intrinsèque de la matrice d'une relation d'équivalence deux fois pondérée.

Démonstration. À partir de la démonstration de la propriété précédente si l'on fait la somme sur le κ blocs. Sinon directement :

$$\sum_{i,i'} u_{ii'} = \sum_{i=1}^{N} \sum_{i'=1}^{N} \frac{x_{ii'}}{x_{i.}x_{.i'}} = \sum_{i=1}^{N} \frac{1}{x_{i.}} \sum_{i'=1}^{N} \frac{x_{ii'}}{x_{.i'}} = \sum_{i=1}^{N} \frac{1}{x_{i.}} \sum_{i'=1}^{N} \hat{x}_{ii'} = \sum_{i=1}^{N} \frac{1}{x_{i.}} = \kappa.$$

L'avant dernière égalité vient du fait que la matrice $\hat{\boldsymbol{X}}$ est bi-stochastique.

□

5. Elle est diagonale par blocs.

Comme nous avons vu, après permutation simultanée des lignes et des colonnes de $\boldsymbol{U}$, elle est décomposable en blocs symétriques diagonaux, chacun des blocs étant composé de valeurs constantes égales à l'inverse du carré de la taille de chaque classe elle peut prendre les valeurs suivantes par ordre croissant :

$$\left\{ \frac{1}{n_1^2}, \frac{1}{n_2^2} \cdots \frac{1}{n_\kappa^2} \right\},$$

où $n_1, n_2, ..., n_\kappa$ désignent les tailles des classes et où ces dernières sont rangées par ordre croissant suivant : $n_\kappa \leq n_{\kappa-1} \leq, ..., \leq n_2 \leq n_1$. Hors blocs, la matrice est composée de valeurs nulles.

6. Chacun de ses termes est le carré du terme général de la matrice relationnelle pondérée inconnue : $u_{ii'} = \hat{x}_{ii'}^2 \qquad \forall (i, i') \in V \times V$.

Démonstration. Il y a deux cas possibles :

- *Soit $x_{ii'} = 0$ et dans ce cas on a bien évidement $\hat{x}_{ii'} = 0$ et $u_{ii'} = 0$.*
- *Soit $x_{ii'} = 1$, cela signifie que i et i' sont dans la même classe et par conséquent $x_{i.} = x_{.i'}$, de plus, comme $x_{ii'}^2 = x_{ii'}$, il vient : $u_{ii'} = \frac{x_{ii'}}{x_{i.}x_{.i'}} = \frac{x_{ii'}^2}{x_{i.}^2} = \hat{x}_{ii'}^2$.*

□

À partir de la propriété 4 il est possible de déduire une expression pour la somme des éléments de la matrice $\bar{\mathbf{U}}$ comme suit :

$$\sum_{i,i'} \bar{u}_{ii'} = \kappa(\kappa - 1). \tag{5.49}$$

Démonstration. $\sum_{i,i'} \bar{u}_{ii'} = \sum_{i,i'} \frac{\bar{x}_{ii'}}{x_{i.}x_{.i'}} = \sum_{i,i'} \frac{1-x_{ii'}}{x_{i.}x_{.i'}} = \sum_{i,i'} \frac{1}{x_{i.}x_{.i'}} - \sum_{i,i'} \frac{x_{ii'}}{x_{i.}x_{.i'}} = \sum_i \frac{1}{x_{i.}} \sum_{i'} \frac{1}{x_{.i'}} - \kappa = \kappa\kappa - \kappa = \kappa(\kappa - 1)$.

□

En réécrivant l'équation (5.48) et si l'on considère que le graphe est non-pondéré :

$F_{MG} = \frac{1}{\kappa} \sum_i^N \sum_{i'}^N \frac{a_{ii'}x_{ii'}}{x_{i.}x_{.i'}} - \frac{1}{\kappa(\kappa-1)} \sum_i^N \sum_{i'}^N \frac{a_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}}$

$= \frac{1}{\kappa} \sum_i^N \sum_{i'}^N \frac{a_{ii'}x_{ii'}}{x_{i.}x_{.i'}} - \frac{1}{\kappa(\kappa-1)} \sum_i^N \sum_{i'}^N \frac{(1-\bar{a}_{ii'})\bar{x}_{ii'}}{x_{i.}x_{.i'}}$

$= \frac{1}{\kappa} \sum_i^N \sum_{i'}^N \frac{a_{ii'}x_{ii'}}{x_{i.}x_{.i'}} - \frac{1}{\kappa(\kappa-1)} \sum_i^N \sum_{i'}^N \frac{\bar{x}_{ii'}}{x_{i.}x_{.i'}} + \frac{1}{\kappa(\kappa-1)} \sum_i^N \sum_{i'}^N \frac{\bar{a}_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}}$.

Avec les résultats trouvés dans l'équation (5.49) nous obtenons :

$= \frac{1}{\kappa} \sum_i^N \sum_{i'}^N \frac{a_{ii'}x_{ii'}}{x_{i.}x_{.i'}} - \frac{\kappa(\kappa-1)}{\kappa(\kappa-1)} + \frac{1}{\kappa(\kappa-1)} \sum_i^N \sum_{i'}^N \frac{\bar{a}_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}}$.

Donc, maximiser le critère de Mancoridis-Gansner revient à maximiser :

$$F_{MG}(X) = \frac{1}{\kappa} \sum_i^N \sum_{i'}^N \frac{a_{ii'}x_{ii'}}{x_{i.}x_{.i'}} + \frac{1}{\kappa(\kappa-1)} \sum_i^N \sum_{i'}^N \frac{\bar{a}_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}} + K$$

.

Avec la notation de la matrice relationnelle de densité $\mathbf{U}$ nous avons :

$$F_{MG}(U) = \frac{1}{\kappa} \sum_{i=1}^N \sum_{i'=1}^N a_{ii'}u_{ii'} + \frac{1}{\kappa(\kappa-1)} \sum_{i=1}^N \sum_{i'=1}^N \bar{a}_{ii'}\bar{u}_{ii'} \tag{5.50}$$

.

Lorsque $\kappa = 2$ on retrouve le critère de Condorcet pondéré deux fois sur $\mathbf{X}$:

$$F_{MG}(U) = \sum_i^N \sum_{i'}^N a_{ii'}u_{ii'} + \sum_i^N \sum_{i'}^N \bar{a}_{ii'}\bar{u}_{ii'}. \tag{5.51}$$

Lorsque $\kappa > 2$ comme $\frac{1}{\kappa} > \frac{1}{\kappa(\kappa-1)}$, l'importance attribuée aux accords positifs est $(\kappa - 1)$ fois celle accordée aux accords négatifs.

La quantité $(\kappa - 1)$ étant positive [19] maximiser le critère de l'équation (5.50) revient à maximiser ce même critère multiplié fois $(\kappa - 1)$:

19. Puisque $\kappa \in [2, N]$, est un entier positif et désigne le nombre de classes.

$$F_{MG}(U) = \frac{\kappa - 1}{\kappa} \sum_{i}^{N} \sum_{i'}^{N} a_{ii'} u_{ii'} + \frac{1}{\kappa} \sum_{i}^{N} \sum_{i'}^{N} \bar{a}_{ii'} \bar{u}_{ii'}. \tag{5.52}$$

La somme des coefficients pondérant le terme d'accords positifs plus celui d'accords négatifs est égal à l'unité : $\frac{\kappa-1}{\kappa} + \frac{1}{\kappa} = 1 \ \forall \kappa$. Ainsi l'importance accordée aux accords positifs est $(\kappa - 1)$ fois celle accordée aux accords négatifs.

En notant $r = \frac{\kappa-1}{\kappa}$ et $1 - r = \frac{1}{\kappa}$ il est possible d'obtenir une généralisation des critères de classification à maximiser :

$$F_{MG}(U) = r \sum_{i}^{N} \sum_{i'}^{N} a_{ii'} z_{ii'} + (1 - r) \sum_{i}^{N} \sum_{i'}^{N} \bar{a}_{ii'} \bar{z}_{ii'}. \tag{5.53}$$

Cette généralisation fut proposée par Owsiński and Zadrożny [1986]. Dans le cas où :

- Nous considérons le critère de Zahn-Condorcet : $r = 1/2$ et $z_{ii'} = x_{ii'}$.
- Nous considérons le critère de Mancoridis-Gansner : $r = \frac{\kappa-1}{\kappa}$ avec κ égal au nombre de classes de la partition ; et $z_{ii'} = \frac{x_{ii'}}{x_{i.} x_{.i'}}$.

On voit immédiatement qu'hormis le paramétrage en "r", la "philosophie" du critère est en filiation directe avec le critère de Condorcet, il s'agit d'une généralisation de ce principe.

5.3.2 Le critère de Ratio-Cuts de Wei-Cheng (1989)

Le critère "Ratio-cuts" de Wei and Cheng [1989] est apparu dans le domaine du partitionnement et du placement de circuits électriques. Un bon partitionnement (décomposition) peut considérablement améliorer la performance du circuit et réduire les coûts de mise en place de celui-ci. Ce besoin est né suite à l'apparition des circuits intégrés VLSI [20] (Intégration à très grande échelle) au début des années 80. Considérant que la plupart de représentations de circuits ont tendance à mettre des composants de fonctionnalités similaires dans un même groupe fortement connecté Wei and Cheng [1989] ont proposé le critère *ratio-cut*.

Pour cette problématique le graphe $G = (V, E)$ constitue le circuit électronique. Les sommets V sont les modules ou composantes du circuit et les arêtes E sont les signaux. Dans un premier temps, les auteurs ont testé comme critère de partitionnement la minimisation des coupures en fixant le nombre de classes égal à 2. Ainsi le but était de minimiser le *cut* tout en maximisant le flux de signaux. Cependant, l'optimisation de ce critère générait 2 sous-circuits de tailles très inégales : une classe à 1, 2 ou 3 sommets et une classe avec le reste des sommets. C'est dans cette direction que les auteurs ont proposé le *Ratio-cut* : un critère qui satisfait deux objectifs, la minimisation des coupures et l'équipartition.

La fonction à minimiser énoncée par Wei and Cheng [1989] cherche à trouver la meilleure partition en 2 sous-ensembles disjoints $\mathcal{C}_1$ et $\mathcal{C}_2$ en minimisant la fonction suivante :

20. La technologie VLSI permet de supporter plus de 100 000 composants électroniques sur une même puce.

$$F_{Rcut}(\mathcal{C}_1, \mathcal{C}_2) = \frac{e(\mathcal{C}_1, \mathcal{C}_2)}{|\mathcal{C}_1| * |\mathcal{C}_2|}, \tag{5.54}$$

où $e(\mathcal{C}_1, \mathcal{C}_2)$ est le nombre d'arêtes entre les classes $\mathcal{C}_1$ et $\mathcal{C}_2$, donc le nombre de coupures, *cuts*.

Ainsi le ratio-cut permet de trouver une partition naturelle : le numérateur minimise les coupures, tandis que le dénominateur favorise une partition équitable.

En notation relationnelle, les coupures sont quantifiées par l'expression $\frac{1}{2}\sum_i^N \sum_{i'}^N a_{ii'}\bar{x}_{ii'}$, et la taille d'une classe de l'objet i est tout simplement la quantité $x_{i.} = \sum_{i'=1}^N x_{ii'}$. Ainsi le critère *Ratio-cut* pour κ classes s'écrit en notations relationnelles comme la quantité à minimiser suivante :

$$F_{Rcut}(X) = \sum_i^N \sum_{i'}^N \frac{a_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}}, \tag{5.55}$$

où $\mathbf{X}$, comme dans les cas précédents, doit satisfaire les contraintes linéaires d'une relation d'équivalence données par l'équation (5.2).

On reconnaît immédiatement le terme de gauche de (5.48). Il s'agit du terme général de la matrice $\bar{\mathbf{U}}$, complémentaire de la matrice de taux de densité $\mathbf{U}$, deux fois pondérée par la taille des classes. L'écriture relationnelle met en évidence que le critère est non-linéaire para rapport à $\mathbf{X}$ et non-équilibré mais il est néanmoins séparable.

Ainsi le ratio-cut permet de trouver une partition naturelle : le numérateur minimise les coupures, tandis que le dénominateur favorise une partition équitable.

Il s'agit du nombre d'arêtes inter-classes pondéré par la taille des classes. En effet, la partie variable de (5.48) représente le terme général de la matrice $\bar{\mathbf{U}}$ (complémentaire de la matrice de taux de densité $\mathbf{U}$).

5.3.3 Le critère de la Différence de Profils (1976)

C'est une distance entre structures de partitions, connue sous le nom de *Distance du* Φ^2 (dans le livre de Cailliez and Pagès [1976]). Il a été écrit, avec des notations relationnelles pour la première fois dans Marcotorchino [1991], et étudié par Bedecarrax and Marcotorchino [1992] et Marcotorchino and El Ayoubi [1991]. Il cherche à minimiser l'expression suivante :

$$F_{DP}(X) = ||\hat{\mathbf{A}} - \hat{\mathbf{X}}||^2 = \sum_i^N \sum_{i'}^N \left(\hat{a}_{ii'} - \hat{x}_{ii'}\right)^2. \tag{5.56}$$

Le critère de la *Différence de profils* constitue une distance euclidienne carrée entre les profils de similarité relationnels de *présence-rareté* associés aux variables $\mathbf{A}$ et $\mathbf{X}$. Minimiser ce critère revient donc à trouver une relation d'équivalence dont le profil "similaritaire de présence-rareté" est le plus proche possible, au sens de la distance euclidienne, de celui relatif à la matrice d'adjacence du graphe. Il s'agit d'un critère à "Éloignement Minimal".

Il y a deux versions[21] de la matrice $\hat{\mathbf{A}}$. Soit on la calcule de façon analogue à $\hat{\mathbf{X}}$: $\hat{a}_{ii'} = \frac{a_{ii'}}{\frac{a_{i.}+a_{.i'}}{2}}$, soit si est seulement si le graphe est non pondéré et non réflexif on calcule le terme général de $\hat{\mathbf{A}}$ comme : $\hat{a}_{ii'} = \frac{a_{ii'}}{\frac{a_{i.}+a_{.i'}+2}{2}}$. La deuxième version est une conséquence du fait que la Différence de Profils compare $\hat{\mathbf{A}}$ à $\hat{\mathbf{X}}$. Le terme général de $\hat{\mathbf{X}}$ vaut le terme général de $\mathbf{X}$, $x_{ii'}$, divisé par la moyenne des tailles des classes $\frac{x_{i.}+x_{.i'}}{2}$. Si le graphe est non réflexif et non pondéré le degré du sommet i est le nombre de sommets adjacents à i. Dans un cas idéal avec zéro coupures et sans les contraintes de transitivité les $a_{i.}$ voisins de i seraient classé ensemble dans la classe de i et l'effectif de cette classe serait, donc, $(a_{i.}+1)$. Le but de cette deuxième version est qu'il existe un certain parallélisme dans la construction de $\hat{\mathbf{A}}$ et de $\hat{\mathbf{X}}$, puisque la Différence de Profils cherche à les comparer.

En développant l'expression (5.56) et en tenant compte de la définition de $\hat{\mathbf{A}}$ et $\hat{\mathbf{X}}$ (équation (2.15)), le critère (5.56) s'écrit :

$$F_{DP}(X) = K - 2\sum_{i}^{N}\sum_{i'}^{N}\hat{a}_{ii'}\hat{x}_{ii'} + \kappa, \tag{5.57}$$

ce qui revient à maximiser l'expression suivante :

$$F_{DP}(X) = 2\sum_{i}^{N}\sum_{i'}^{N}\hat{a}_{ii'}\hat{x}_{ii'} - \kappa + K \tag{5.58}$$

ou encore maximiser la quantité :

$$\sum_{i}^{N}\sum_{i'}^{N}\left(2\hat{a}_{ii'} - \frac{1}{x_{i.}}\right)\hat{x}_{ii'}, \tag{5.59}$$

où $\mathbf{X}$ doit vérifier les contraintes d'une relation d'équivalence, voir l'équation (5.2).

L'écriture relationnelle de la "Différence de Profils", équations (5.58) et (5.59), permet de constater que ce critère est non-linéaire en $\mathbf{X}$, non-équilibré mais il est séparable. La solution optimale de ce critère n'est pas du tout triviale (cas correspondant au fait que tous les individus sont isolés les uns des autres) comme cela se produit inévitablement dans de nombreux critères inertiels aboutissant de facto à l'usage des algorithmes de type κ-means. Ce critère donne souvent des résultats intéressants comme nous allons le voir dans l'exemple d'application.

C'est l'expression (5.58) que nous allons fournir et préconditionner pour le puissant algorithme heuristique de modularisation "générique" dit (Algorithme de "Louvain") dont une version générique a été développée au sein de l'Equipe Complex networks du LIP6. Nous reviendrons sur ce point au Chapitre 7.

5.3.4 Le critère de Michalski-Goldberg (2012)

Ce critère cherche à maximiser la densité de chaque communauté telle que définie dans Goldberg [1984]. Selon Goldberg la densité d'un graphe est le rapport entre le nombre d'arêtes et le nombre de sommets. A partir de cette définition, Darlay et al. [2012] ont

21. C'est la première version qui sera considéré pour l'algorithme de calculs (voir chapitre 7.)

proposé un critère de modularisation visant à maximiser la densité d'une partition d'un graphe, définie comme la somme des densités des sous-graphes induits par chaque classe de la partition. La densité d'une classe correspond alors au rapport entre le nombre d'arêtes intra-classe et la taille de la classe. En notation relationnelle :

$$F_G(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \frac{a_{ii'}x_{ii'}}{x_{i.}} = \sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}\hat{x}_{ii'}. \tag{5.60}$$

Il s'agit du terme d'accords positifs du critère de Condorcet pondéré en $\mathbf{X}$ ou critère de Michalski-Decaestecker introduit dans Michalski and Stepp [1983] et dont l'écriture relationnelle a été donnée par Decaestecker [1992] dans sa thèse en apprentissage statistique.

Critère	**Écriture Relationnelle**
Mancoridis-Gansner (1998)	$F_{MG}(X) = \frac{1}{\kappa}\sum_{i=1}^{N}\sum_{i'=1}^{N}\frac{a_{ii'}x_{ii'}}{x_{i.}x_{.i'}} + \frac{1}{\kappa(\kappa-1)}\sum_{i}^{N}\sum_{i'}^{N}\frac{\bar{a}_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}}$ avec $\kappa > 1$
Wei-Cheng (1989) (Ratio-Cuts)	$F_{Rcut}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\frac{a_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}}$
Différence de profils (1976)	$F_{DP}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(2\hat{a}_{ii'} - \frac{1}{x_{i.}}\right)\hat{x}_{ii'}$
Michalski-Goldberg (2012)	$F_G(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}\hat{x}_{ii'}$

Table 5.3 – Critères séparables fonctions non-linéaires de $\mathbf{X}$.

Tous les critères du tableau 5.3 sont à maximiser sauf celui de *Ratio cuts*.
Le Tableau 5.3 montre des critères non-linéaires en $\mathbf{X}$. La partie variable de ces critères représente une pondération simple ou une double pondération par rapport à la taille des classes. Pour le critère de Mancoridis-Gansner, si l'on omet les termes de pondération en $\frac{1}{\kappa}$ et $\frac{1}{\kappa(\kappa-1)}$ on obtient le critère de Condorcet deux fois pondéré par la taille des classes.

5.4 Autres critères

5.4.1 Critère dit "Normalised cuts" de Shi-Malik (2000)

Étant donnée un graphe $G = (V, E)$ partitionné en κ classes disjointes, le *cut* (coupures en français) est le poids total de toutes les arêtes inter-classes, i.e. toutes les arêtes qui ont été *coupées* pour modulariser le graphe. Dans Shi and Malik [2000], les auteurs ont abordé ce problème pour $\kappa = 2$ classes : $\mathcal{C}_1$ et $\mathcal{C}_2$ et ont défini la coupure comme :

$$cut(\mathcal{C}_1, \mathcal{C}_2) = \sum_{i\in\mathcal{C}_1, i'\in\mathcal{C}_2} w_{ii'}.$$

Ce qui peut être étendu à une partition $\mathcal{C}$ en κ classes :

$$cut(\mathcal{C}) = cut(\mathcal{C}_1, V - \mathcal{C}_1) + cut(\mathcal{C}_2, V - \mathcal{C}_2) + ... + cut(\mathcal{C}_\kappa, V - \mathcal{C}_\kappa). \tag{5.61}$$

Cependant minimiser la quantité *cut* avec κ fixé[22] comme critère de partitionnement favorisait les coupures en petites classes, la plupart d'un seul sommet isolé, car lesdites coupures ont une faible valeur de *cut*. Pour éviter ce problème J. Shi et J. Malik dans Shi and Malik [2000] ont proposé de normaliser le critère *cut*. Ce qu'ils ont nommé le critère de *Normalized cut : Ncut* (coupure normalisée).

Au lieu de considérer la valeur globale de poids inter-classes, le *Ncut* considère la fraction de poids des arêtes coupées par rapport aux poids des connexions de chaque classe à tous les sommets du graphe. Ainsi, le critère *Ncut* s'écrit comme :

$$F_{Ncut}(\mathcal{C}) = \frac{cut(\mathcal{C}_1, V - \mathcal{C}_1)}{assoc(\mathcal{C}_1, V)} + \frac{cut(\mathcal{C}_2, V - \mathcal{C}_2)}{assoc(C_2, V)} + ... + \frac{cut(\mathcal{C}_\kappa, V - \mathcal{C}_\kappa)}{assoc(\mathcal{C}_\kappa, V)}, \tag{5.62}$$

où $assoc(\mathcal{C}_j, V) = \sum_{i \in \mathcal{C}_j, i' \in V} w_{ii'}$ est le poids total des arêtes connectant les sommets de la classe $\mathcal{C}_j$ à tous les sommets du graphe. Ainsi la coupure normalisée des classes à un seul sommet n'a plus un coût faible car dans ce cas la proportion de poids des arêtes coupés est proche de 100% (S'il n'y a pas de boucle est 100%).

En notation relationnelle, le poids total des arêtes coupées (c.f. (5.61) est donné par la formule :

$$F_{Ncut}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} w_{ii'} \bar{x}_{ii'}. \tag{5.63}$$

Pour le calcul de *Ncut* il s'agit tout simplement de diviser le poids de chaque coupure du sommet i par le poids total des arêtes reliant tous les sommets de la classe contenant i, notée $C(i)$, à tous les sommets du graphe, i.e. : $\sum_{i''} w_{i''.} X_{ii''}$, en effet, $w_{i''.}$ représente la somme des poids des arêtes connectées au sommet i'' et $X_{ii''}$ assure que la somme soit effectuée seulement pour les sommets contenus dans $C(i)$.

Ainsi en notation relationnelle le critère de coupures normalisées s'écrit :

$$F_{Ncut}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \frac{w_{ii'} \bar{x}_{ii'}}{\sum_{i''} w_{i''.} x_{ii''}}. \tag{5.64}$$

En 2008, L. Labiod Labiod [2008] avait déjà écrit le critère de *coupures normalisées* en notations relationnelles :

$$F_{Ncut}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \frac{L_{ii'} x_{ii'}}{\sum_{i''} d_{i''} x_{ii''}} \tag{5.65}$$

Où $L_{ii'}$ est le terme général de la matrice Laplacienne du graphe. Les équations (5.64) et (5.65) sont équivalentes.

22. Si κ n'est pas fixé la solution optimale est la partition grossière avec *cut*=0 coupures où tous les sommets sont classés ensemble et $\kappa = 1$.

5.4.2 Critère de Zhou-Dillon (1991)

Ce critère fut proposé en 1991 par Zhou et Dillon Zhou and Dillion [1991] dans le domaine d'apprentissage automatique et fouille de données, en effet, il intervient dans la phase de prétraitement des données. Lorsque l'on dispose d'un jeu de données, en fonction de la problématique considérée certaines variables deviennent plus pertinentes que d'autres. Le but du prétraitement des données est de filtrer les variables inutiles et redondantes avant la phase d'apprentissage. Le critère se traduit par une fonction d'évaluation, calculée pour chaque variable, qui juge sa pertinence.

Ces critères peuvent être des mesures d'écart à l'indépendance (de corrélation ou d'association) comme la mesure d'association du χ^2 ; ou de consistance : deux objets sont inconsistants si leurs modalités sont identiques et s'ils appartiennent à deux classes différentes.

Le tau de Zhou est une mesure de consistance servant à détecter les variables redondantes. Il mesure la dépendance entre deux variables. En apprentissage statistique lorsque l'on veut éliminer la redondance on garde les variables dont le tau de Zhou par rapport aux autres variables est faible ; il s'agit, donc, d'un critère à minimiser. Son calcul se fait à partir d'un tableau de contingence de deux variables catégorielles : A et X à P et κ modalités respectivement :

$$F_\tau(C,X) = \frac{\sum_{u=1}^{P}\sum_{v=1}^{\kappa}\left[\frac{(\frac{n_{uv}}{N})^2}{\frac{n_{u.}}{N}}\right] + \sum_{u=1}^{P}\sum_{v=1}^{\kappa}\left[\frac{(\frac{n_{uv}}{N})^2}{\frac{n_{.v}}{N}}\right] - \sum_{v=1}^{\kappa}(\frac{n_{.v}}{N})^2 - \sum_{u=1}^{P}(\frac{n_{u.}}{N})^2}{2 - \sum_{v=1}^{\kappa}(\frac{n_{.v}}{N})^2 - \sum_{u=1}^{P}(\frac{n_{u.}}{N})^2}, \tag{5.66}$$

où :

- Le terme n_{uv} est le terme général du tableau de contingence : nombre d'objets possédant les modalités u et v des variables **A** et **X** respectivement.
- $n_{u.}$: nombre d'objets possédant la modalité u de la variable **A**.
- $n_{.v}$: nombre d'objets possédant la modalité v de la variable **X**.

L'écriture de chaque terme du tau de Zhou τ en notation relationnelle s'obtient grâce aux formules de passage contingence-paires [23] (voir annexe A pour une liste non-exhaustive des formules de transfert, pour plus de détails sur leur obtention et démonstrations voir Kendall and Stuart [1961] et Marcotorchino [1984a]) :

Ainsi l'écriture relationnelle du tau de Zhou sera :

$$F_\tau(C,X) = \frac{\frac{1}{N}\sum_{i=1}^{N}\sum_{i'=1}^{N}[\hat{a}_{ii'}x_{ii'} + a_{ii'}\hat{x}_{ii'}] - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}[a_{ii'} + x_{ii'}]}{2 - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i'=1}^{N}[x_{ii'} + a_{ii'}]}. \tag{5.67}$$

5.5 Théorie spectrale du "clustering" de graphes comme outil de modularisation

La théorie spectrale du *clustering* de graphes remonte à Donath and Hoffman [1973], qui pour la première fois en 1973 ont suggéré de partitionner un graphe à partir des vecteurs

23. Ces formules permettent de lier l'Analyse Factorielle et l'Analyse Relationnelle.

et valeurs propres de la matrice d'admittance ou laplacien du graphe. Dans Donath and Hoffman [1973] les auteurs ont donné une borne inférieure au nombre d'arêtes coupées lors du partitionnement. Cette borne, exprimée en fonction des valeurs propres de la matrice laplacienne du graphe et de la matrice de la relation cherchée X, est une conséquence du théorème de Hoffman and Wielandt [1952], dont l'énoncé original est :

Théorème 5.3 (Lower Bounds for the Partitioning of Graphs). *Let a κ-partition of a graph be a division of the vertices into κ disjoint subsets containing $n_1 \leq n_2 \leq ... \leq n_\kappa$ vertices. Let E_{cut} be the number of edges whose two vertices belong to different subsets. Let $\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_\kappa$ be the κ largest eigenvalues of a matrix, which is the sum of the adjacency matrix of the graph plus any diagonal matrix* D *such that the sum of all the elements of the sum matrix is zero. Then :*

$$E_{cut} \geq \frac{1}{2} \sum_{r=1}^{\kappa} -n_r \lambda_r .$$

Clairement le membre gauche de cette inégalité E_{cut} est le nombre de coupures résultant du partitionnement. En prenant la matrice D comme l'opposée de la matrice des degrés du graphe i.e. $-\mathbf{D}$ les valeurs λ_r seraient en fait les valeurs propres de la matrice d'admittance $\mathbf{L} = \mathbf{A} - \mathbf{D}$. Cette quantité permet d'obtenir une borne supérieure pour le nombre d'arêtes inter-classes en fonction des valeurs propres de la matrice Laplacienne et des tailles des classes, qui, comme nous le verrons par la suite sont les valeurs propres de la matrice de la relation d'équivalence X cherchée.

Durant la période où M. Fiedler poursuivait ses travaux sur la connectivité des graphes (cf. Fiedler [1973]), il découvrit qu'il y avait un lien étroit entre la bi-partition d'un graphe (partition du graphe en deux ensembles) et le vecteur propre associé à la seconde plus petite valeur propre de la matrice laplacienne. Il suggéra ainsi d'utiliser ce vecteur pour partitionner un graphe.

A partir des travaux de Donath, Hoffman et Fiedler la théorie spectrale comme outil de partitionnement d'un graphe a été reprise et étudiée par plusieurs auteurs. Un bon résumé autour de l'étude du spectre de la matrice laplacienne et de la seconde plus petite valeur propre peut être trouvé dans le célèbre article : Mohar [1991]. Dans Luxburg [2007] l'auteur a décrit les principales propriétés de la matrice Laplacienne.

Étant donné un graphe $G = (V, E)$ non orienté et sans boucles ; sa matrice Laplacienne **L=D-A** possède les propriétés suivantes :

- L est symétrique et semi-définie positive, comme conséquence toutes ses valeurs propres sont réelles.
- La plus petite valeur propre de L notée λ_1, est nulle et elle est associée au vecteur propre $\mathbb{1} = (1, 1, ..., 1)^t$.
- La multiplicité de la valeur propre λ_1 est égale au nombre de composantes connexes du graphe [24]. Cette propriété est très intéressante car si le graphe est connexe la deuxième plus petite valeur propre doit être positive.
- Tout vecteur $f \in \mathbb{R}^N$ vérifie :

24. Ceci est une conséquence du théorème de Perron-Frobenius.

$$f^t L f = \frac{1}{2}\sum_{i,i'=1}^{N} w_{ii'}(f_i - f_{i'})^2,$$

où $w_{ii'}$ est l'élément ii' de la matrice de poids (cas d'un graphe pondéré) ou de la matrice d'adjacence (cas d'un graphe non pondéré).

.

- La somme de chaque ligne et colonne de L est nulle.
- Selon le *Théorème Minimax* ou Principe de Courant-Fischer, étant donné que L est semi-définie positive, sa seconde plus petite valeur propre est donnée par (cf. Fiedler [1973] et Hagen and Kahng [1992]) :

$$\lambda_2 = \min_{z\perp\mathbb{1}, z\neq\mathbf{0}} \frac{z^t L z}{|z|^2}. \tag{5.68}$$

Dans Mohar [1991] d'autres bornes relatives aux valeurs propres de L sont données.

Tous les travaux existants sur la théorie spectrale des graphes reflètent un lien étroit entre cette dernière et le partitionnement d'un graphe : Le théorème 5.3 fait le lien entre le nombre de coupures et les valeurs propres du laplacien et la taille des classes. En s'inspirant des travaux de Pothen et al. [1990] Newman, dans Newman [2006a] et Newman [2006b], propose une réécriture de son critère de modularité en fonction des vecteurs et valeurs propres d'une matrice symétrique nommée par lui "*modularity matrix*".

Soit $\mathbf{K} \in \mathcal{M}_{N,\kappa}$ le tableau disjonctif complet de la relation d'équivalence cherchée $\mathbf{X}$ à κ classes, donc chaque élément de $\mathbf{K}$ est défini de la façon suivante :

$$k_{ij} = \begin{cases} 1 & \text{si le sommet } i \text{ appartient a la classe } j, \\ 0 & \text{sinon.} \end{cases} \tag{5.69}$$

Ainsi le vecteur colonne j, noté $\mathbf{k}_j$, de $\mathbf{K} \in \mathbb{R}^N$ et son *i-ème* élément est soit nul si i n'appartient pas à la classe j, soit 1 si i est dans la classe j. La matrice $\mathbf{K}$ peut être représentée aussi comme : $\mathbf{K} = (\mathbf{k}_1|\mathbf{k}_2|...|\mathbf{k}_\kappa)$.

L'élément $x_{ii'}$ de la relation $\mathbf{X}$ peut s'écrire en fonction du produit scalaire entre la ligne (vecteur) i et la ligne (vecteur) i' de la matrice $\mathbf{K}$:

$$x_{ii'} = \delta_{ii'} = \sum_{j=1}^{\kappa} k_{ij}k_{i'j}. \tag{5.70}$$

En tenant compte de la formule précédente l'équation (5.10) peut s'écrire :

$$F_{NG}(X) = \frac{1}{2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\sum_{j=1}^{\kappa}(a_{ii'} - \frac{a_{i.}a_{.i'}}{2M})k_{ij}k_{i'j}. \tag{5.71}$$

En notant $\mathbf{B} \in \mathcal{M}_N$ la matrice carrée d'ordre N dont le terme général est : $b_{ii'} = (a_{ii'} - \frac{a_{i.}a_{.i'}}{2M})$. L'équation (5.71) s'écrit :

$$F_{NG}(K) = \frac{1}{2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\sum_{j=1}^{\kappa} b_{ii'}k_{ij}k_{i'j} = \frac{1}{2M}Tr(\mathbf{K^T B K}). \tag{5.72}$$

Comme la matrice $\mathbf{B}$ est symétrique, elle est diagonalisable dans une base orthonormée. Si $\mathbf{U} = (\mathbf{U}_1|\mathbf{U}_2|...|\mathbf{U}_N)$ est la matrice de vecteurs propres de $\mathbf{B}$ et $\mathbf{D}_B$ est la matrice diagonale d'ordre N dont l'élément ii est la *i-ème* valeur propre de $\mathbf{B}$ alors $\mathbf{B} = \mathbf{U}\mathbf{D}_\mathbf{B}\mathbf{U}^\mathbf{T}$. Donc, l'équation (5.72) peut s'écrire ainsi :

$$F_{NG}(K) = Tr(\mathbf{K}^\mathbf{T}\mathbf{B}\mathbf{K}) = Tr(\mathbf{K}^\mathbf{T}\mathbf{U}\mathbf{D}_\mathbf{B}\mathbf{U}^\mathbf{T}\mathbf{K}) = Tr((\mathbf{U}^\mathbf{T}\mathbf{K})^\mathbf{T}\mathbf{D}_\mathbf{B}(\mathbf{U}^\mathbf{T}\mathbf{K})),$$

où terme $\frac{1}{2M}$ a été omis car il constitue une normalisation et ne modifie pas la fonction à maximiser).

Si l'on note β_i la *i-ème* valeur propre de la matrice $\mathbf{B}$, l'expression précédente peut s'écrire comme :

$$F_{NG}(K) = \sum_{i}^{N}\sum_{j=1}^{\kappa}\beta_i(\mathbf{U}_\mathbf{i}^\mathbf{T}\mathbf{k}_\mathbf{j})^2. \tag{5.73}$$

Cette réécriture de la modularité (cf. Newman [2006b]) montre que, les vecteurs $\mathbf{U_i}$ et les valeurs β_i étant connus il faut choisir les vecteurs $\mathbf{k_j}$ de la matrice $\mathbf{K}$ de façon à maximiser l'expression entre parenthèses de la formule (5.73). Cette expression n'est autre que le produit scalaire entre le *i-ème* vecteur propre de $\mathbf{B}$ et le *j-ième* vecteur colonne de $\mathbf{K}$. Le produit scalaire entre deux vecteurs est maximal si ces deux vecteurs sont parallèles. S'il n'y avait pas d'autre contrainte sur les valeurs des vecteurs $\mathbf{k_j}$ le problème consistant à maximiser (5.73) reviendrait à choisir les $\mathbf{k_j}$ proportionnels aux vecteurs propres $\mathbf{U_i}$ associés aux plus grandes valeurs propres de $\mathbf{B}$. Seulement les valeurs propres positives de $\mathbf{B}$ peuvent augmenter la modularité, donc il faudrait choisir κ inférieur ou égal au nombre de valeurs propres positives de $\mathbf{B}$. cela donne une borne supérieure pour le nombre de classes de la partition optimale $\mathbf{X}$. Malheureusement les vecteurs $\mathbf{k_j}$ ne peuvent pas être proportionnels aux vecteurs $\mathbf{U_i}$ car ils contiennent des variables binaires, i.e. $u_{ij} \in \{0, 1\}$ et cette contrainte incontournable fait que le problème ne se résoud pas de façon triviale, et que tout au plus l'approche par recherche de valeurs propres donnera des bornes ou des approximations de la solution optimale.

Pour réécrire la fonction d'optimisation de Newman (équation (5.10)) en fonction du spectre de la matrice de modularité (équation (5.73)) nous avons seulement utilisé la propriété de symétrie de cette matrice. Cela signifie que tout critère consistant à chercher une relation d'équivalence $\mathbf{X}$ "correspondant au mieux" à une **matrice de données ou d'information** *symétrique*, donc tous les critères linéaires listés au tableau 5.2, peut s'écrire en fonction des vecteurs et valeurs propres de cette matrice.

Par exemple le critère de Zahn-Condorcet (5.5) revient à maximiser :

$$F_{ZC}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(a_{ii'} - \bar{a}_{ii'})x_{ii'}. \tag{5.74}$$

La matrice $(\mathbf{A} - \bar{\mathbf{A}})$ étant la différence de deux matrices symétriques est aussi symétrique d'ordre N. Elle est alors diagonalisable dans une base orthonormée. Si l'on note les vecteurs propres de $(\mathbf{A} - \bar{\mathbf{A}})$: $(\mathbf{V_i})$ et ses valeurs propres α_i, l'équation (5.74) s'écrit :

$$F_{ZC}(K) = \sum_{i=1}^{N} \sum_{j=1}^{\kappa} \alpha_i (\mathbf{V_i^T k_j})^2. \tag{5.75}$$

Le terme entre parenthèses étant le produit scalaire entre deux vecteurs devient :

$$(\mathbf{V_i^T k_j})^2 = ||\mathbf{V_i}||^2 ||\mathbf{k_j}||^2 cos\theta_{ij}^2,$$

d'où l'on tire les propriétés suivantes :

- $||\mathbf{V_i}|| = 1$ car les vecteurs $\mathbf{V_i}$ constituent une base orthonormée de $\mathbb{R}^N$, ils sont, donc unitaires.
- $||\mathbf{k_j}||^2 = n_j$, en effet les vecteurs $\mathbf{k_j}$ contiennent autant d'éléments égaux à 1 que la taille de la classe j, les autres éléments étant nuls.
- θ_{ij} est l'angle entre les vecteurs $\mathbf{V_i}$ et $\mathbf{k_j}$.

Optimiser le critère de Zahn-Condorcet en notations matricielles revient alors à maximiser :

$$F_{ZC}(K) = \sum_{i}^{N} \sum_{j=1}^{\kappa} \alpha_i n_j (cos\theta_{ij})^2. \tag{5.76}$$

Encore une fois l'équation (5.76) montre que pour maximiser le critère de Zahn-Condorcet il faut maximiser le cosinus entre les vecteurs $\mathbf{V_i}$ et $\mathbf{k_j}$. Cela signifie que les vecteurs $\mathbf{k_j}$ doivent être le plus parallèles possible aux vecteurs propres associés aux plus grandes valeurs propres de la matrice $(\mathbf{A} - \bar{\mathbf{A}})$.

D'autre part les vecteurs $\mathbf{V_i}$ étant unitaires les valeurs de leurs composantes sont comprises dans l'intervalle $[-1, 1]$, quant aux vecteurs $\mathbf{k_j}$ leurs composantes ne peuvent contenir que des valeurs $\{0, 1\}$. Une première idée pour définir le vecteur $\mathbf{k_1}$ de la classe 1 serait de mettre dans la même classe les sommets dans la composante du vecteur $\mathbf{V_N}$ (associé à la plus grande valeur propre α_N) est positive ou proche de 1.

Chapitre 6

Comparaison des critères de modularisation

6.1 Introduction

Nous verrons au chapitre suivant, qu'il existe des différences importantes dans les partitions trouvées avec les différents critères de modularisation présentés au chapitre précédent. Comme mentionné au chapitre 3, chaque critère a sa propre définition de la notion de *communauté*, et en fonction de cette définition ils optimisent différentes expressions fonctions des données d'entrée et de l'inconnue $\mathbf{X}$. À titre d'exemple, le tableau suivant montre le nombre de classes trouvées κ avec les critères linéaires pour deux réseaux réels connus dans la littérature de réseaux sociaux [1] :

Réseau	Jazz $N = 198$ $M = 2\ 742$	Internet $N = 69\ 949$ $M = 351\ 380$
Critère	κ	κ
Zahn-Condorcet	38	40 123
Owsiński-Zadrożny	6 $\alpha = 2\%$	456 $\alpha < 1\%$
Écart à l'Uniformité	20	173
Newman-Girvan	4	46
Écart à l'Indétermination	6	39
Modularité Équilibrée	5	41

TABLE 6.1 – Nombre de classes trouvées avec différents critères de modularisation.

Le premier réseau "jazz" est un graphe de collaboration entre musiciens de jazz qui jouaient entre 1912 et 1940. Chaque sommet correspond à un groupe ou orchestre de jazz. Deux sommets sont connectés s'ils ont un musicien en commun (voir Gleiser and Danon [2003]). Le réseau internet est un sous-graphe d'internet (voir Hoerdt and Magoni [2003]).

Le tableau 6.1 montre des différences importantes en ce qui concerne le nombre optimal de classes obtenu via l'optimisation de chaque critère. Les critères de Zahn-Condorcet,

1. Le processus d'optimisation a été effectué avec l'algorithme de Louvain (voir Blondel et al. [2008] et Campigotto et al. [2013]) dont nous verrons une description au chapitre suivant.

Owsiński-Zadrożny et Écart à l'Uniformité génèrent plus de classes que les critères de Newman-Girvan, Écart à l'Indétermination et Modularité Équilibrée.

Dans ce chapitre nous allons nous servir de l'écriture relationnelle des critères présentée dans les chapitres précédents pour comprendre ces différences entre les partitions trouvées. Le fait d'utiliser les mêmes notations pour tous les critères nous permettra de les comparer.

6.2 Comparaison des critères linéaires

Dans cette section nous écrirons six critères présentés dans le tableau 5.2 sous la forme générale d'un critère ayant la propriété d'équilibre linéaire présentée dans l'équation (4.6) :

$$F(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} (\phi_{ii'} - \bar{\phi}_{ii'}) x_{ii'}. \tag{6.1}$$

Nous omettrons la constante K car sa valeur n'intervient pas dans le processus d'optimisation. D'autre part, nous omettrons aussi toute constante multipliant la valeur de chaque critère.

Comme mentionné au chapitre 4, pour qu'un critère possède la propriété d'équilibre général la seule contrainte sur les fonctions $\phi_{ii'}$ et $\bar{\phi}_{ii'}$ est qu'elles soient positives et que leur somme ne soit pas nulle. Tous les critères linéaires énoncés dans le tableau 5.2 vérifient cette propriété qui garantit que la solution optimale ne soit pas triviale ni grossière et par conséquent, que l'on ne soit pas obligé de fixer le nombre de classes à l'avance. Nous allons voir maintenant les rôles de ces deux fonctions sur la partition optimale.

De la même façon que les critères de Newman-Girvan, d'Écart à l'Indétermination et d'Écart à l'Uniformité, d'autres critères du tableau 5.2 peuvent s'exprimer aussi sous forme de fonctions cherchant à maximiser l'écart entre le graphe réel et une version particulière de celui-ci, comme le montre le tableau 6.2.

Les expressions présentées dans le tableau 6.2 ont été obtenues à partir des expressions des critères (voir chapitre 5) et des définitions des matrices $\bar{\mathbf{A}}$ et $\bar{\mathbf{X}}$. Le critère de *correlation clustering* et le critère de Condorcet pondéré en $\mathbf{A}$ ne feront pas l'objet de notre étude. En ce qui concerne le premier, il possède des propriétés très voisines de celles du critère de Zahn-Condorcet et pour le deuxième le lecteur intéressé pourra se référer à Marcotorchino [1991].

Critère	**Écriture** $F(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(\phi_{ii'} - \bar{\phi}_{ii'})x_{ii'}$
Zahn-Condorcet	$F_{ZC}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{1}{2}\right)x_{ii'}$
Owsiński-Zadrożny	$F_{OZ}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(a_{ii'} - \alpha)x_{ii'}$
Écart à l'Uniformité	$F_{\text{UNIF}}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{2M}{N^2}\right)x_{ii'}$
Newman-Girvan	$F_{NG}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{a_{i.}a_{.i'}}{2M}\right)x_{ii'}$
Écart à l'Indétermination	$F_{DI}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \left(\frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2}\right)\right)x_{ii'}$
Modularité Équilibrée	$F_{BM}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} + \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M} - \left(\bar{a}_{ii'} + \frac{a_{i.}a_{.i'}}{2M}\right)\right)x_{ii'}$

Table 6.2 – Critères linéaires de modularisation écrits sous forme d'écart à une structure particulière.

Le tableau 6.2 montre que cinq critères : Zahn-Condorcet, Owsiński-Zadrożny, Écart à l'Uniformité, Newman-Girvan et Écart à l'Indétermination possèdent la même fonction ϕ, qui est égale au terme général de matrice d'adjacence, i.e. :

$$\phi_{ii'}^{ZC} = \phi_{ii'}^{OZ} = \phi_{ii'}^{NG} = \phi_{ii'}^{DI} = \phi_{ii'}^{\text{UNIF}} = a_{ii'} \,\forall\,(i,i'). \tag{6.2}$$

Par conséquent, le terme d'accords positifs que ces critères cherchent à maximiser est le même. En revanche, les cinq critères diffèrent dans le terme d'accords négatifs, caractérisé par la fonction $\bar{\phi}$. En effet, c'est le terme d'accords négatifs qui justifie les différences trouvées entre les partitions obtenues via l'optimisation des cinq critères. Nous distinguons deux types de fonctions $\bar{\phi}$:

1. Le cas où $\bar{\phi}$ est une constante ne dépendant pas de chaque paire de sommets : En effet, c'est le cas de l'expression des trois critères : Zahn-Condorcet, Owsiński-Zadrożny et Écart à l'Uniformité où la fonction $\bar{\phi}$ est constant pour toute paire de sommets du graphe i et i'.

2. Le cas où $\bar{\phi}$ est variable : C'est le cas des critères de Newman-Girvan et d'Écart à l'Indétermination pour lesquels la fonction $\bar{\phi}$ dépend de la distribution des degrés.

Nous caractériserons dans un premier temps les fonctions $\bar{\phi}$ des deux cas mentionnés. Ensuite, nous verrons le lien entre les cinq critères et la Modularité Équilibrée.

6.2.1 Lien entre les critères de Zahn-Condorcet, d'Owsiński-Zadrożny et l'Écart à l'Uniformité

À partir du tableau 6.2 nous pouvons déduire que les critères de Zahn-Condorcet et l'Écart à l'Uniformité sont des cas particuliers du critère d'Owsiński-Zadrożny. Le critère d'Owsiński-Zadrożny maximise l'écart à une constante α définie par l'utilisateur.

Le critère de Zahn-Condorcet maximise l'écart à la constante [2] $\alpha = \frac{1}{2}$. Cette valeur peut avoir plusieurs interprétations. D'une part, elle représente l'espérance d'une variable aléatoire binaire prenant les valeurs : 0 et 1. D'autre part, il s'agit du point moyen de l'intervalle $[0, 1]$, le point d'équilibre entre les deux valeurs possibles de la variable relationnelle $x_{ii'}$. Ce point d'équilibre peut s'interpréter comme une situation d'indécision entre les deux valeurs extrêmes duquel le critère cherche à écarter les variables $x_{ii'}$.

Le critère d'Écart à l'Uniformité maximise l'écart à la densité d'arêtes du graphe δ. Comme nous avons évoqué au chapitre 1, tableau 1.1, $\delta << \frac{1}{2}$ pour divers graphes réels issus de différents domaines. Par conséquent, nous aurons assez souvent pour des graphes réels :

$$\bar{\phi}^{\text{UNIF}} << \bar{\phi}^{ZC}$$

.

Ces résultats peuvent se visualiser dans le schéma 6.1.

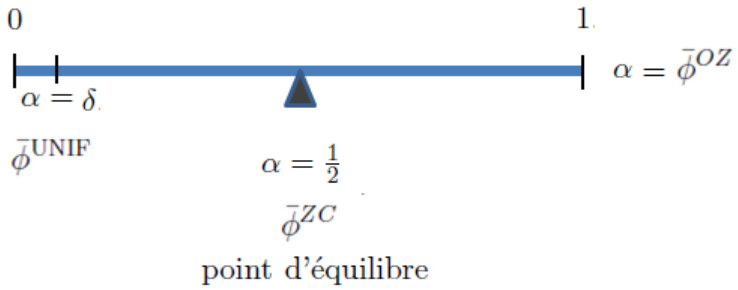

FIGURE 6.1 – Comparaison de la fonction $\bar{\phi}$ pour les critères de Zahn-Condorcet, d'Owsiński-Zadrożny et l'Écart à l'Uniformité.

6.2.2 Comparaison des critères dépendant de la distribution des degrés du graphe

Nous avons vu que le terme d'accord négatifs du critère de Newman-Girvan et celui du critère d'Écart à l'Indétermination dépendent de la distribution des degrés du graphe. Nous avons vu au chapitre 1 que le degré d'un sommet constitue une mesure de centralité qui identifie les sommets qui occupent une position centrale dans le réseau. Dans cette sous-section nous allons analyser les critères dépendant de la distribution des degrés du

2. La valeur $\frac{1}{2}$ vient de l'origine de la formulation du critère en théorie des votes : la recherche de la *Majorité absolue* : 50% (voir théorème 5.1).

graphe. Il a été prouvé que la distribution des degrés de la plupart des graphes réels (internet, réseaux biologiques, réseaux sociaux, etc.) suit une loi de puissance (voir Barabasi and Albert [1999]). Cela signifie que peu de sommets, appelés "hubs", sont connectés à un grand nombre de sommets, qui possèdent à leur tour beaucoup moins de connexions. Dans le langage de la théorie des réseaux, ces réseaux sont appelés *réseaux sans échelle*.

Comportement de la fonction $\bar{\phi}^{NG}$ du critère de Newman-Girvan

Pour le critère de Newman-Girvan la fonction caractérisant le terme d'accords négatifs est : $\bar{\phi}^{NG}_{ii'} = \frac{a_{i.}a_{.i'}}{2M}$. Nous verrons que dans la plupart des graphes réels cette fonction est comprise entre 0 et 1 (surtout dans les grands graphes où $N \to \infty$, si le graphe est connecté on a $M \geq (N-1)$, donc $M \to \infty$ et le produit $(a_{i.}a_{.i'})$ devient petit par rapport à $2M$). Lorsque $0 < \frac{a_{i.}a_{.i'}}{2M} \leq 1$ est vérifié le domaine de définition de $\phi^{NG}_{ii'} = a_{ii'}$ correspondrait à la borne inférieure et supérieure de l'intervalle de définition de $\bar{\phi}^{NG}$ car $a_{ii'} \in \{0,1\}$. D'autre part, cette fonction dépend de la structure globale du graphe car elle dépend du nombre total d'arêtes.

Voyons maintenant dans quels cas on a $\frac{a_{i.}a_{.i'}}{2M} \geq 1$. Cela implique d'avoir $a_{i.}a_{.i'} \geq 2M$, donc c'est le comportement du produit $a_{i.}a_{.i'}$ qui nous intéresse, et surtout les valeurs extrêmes de ce produit [3].

La quantité $\frac{a_{i.}a_{.i'}}{2M}$ est maximale lorsque le produit $a_{i.}a_{.i'}$ est maximal et M vaut le minimum possible. D'autre part pour tout graphe non pondéré et non réflexif nous avons $a_{i.} \leq (N-1)$ et $a_{.i'} \leq (N-1)$. Le maximum de $\bar{\phi}^{NG}_{ii'}$ a lieu lorsque ces deux dernières inégalités deviennent égalités et lorsque $M = 2(N-1)-1$ comme pour les sommets 1 et 2 du graphe montré dans la figure 6.2. Nous déduisons alors la borne suivante :

$$\bar{\phi}^{NG}_{ii'} \leq \frac{(N-1)^2}{2(2(N-1)-1)} = \frac{(N-1)^2}{2(2N-3)}. \tag{6.3}$$

Lorsque $N \to \infty$ $\bar{\phi}^{NG}_{ii'}$ peut prendre des valeurs très importantes par rapport à $\phi^{NG}_{ii'} = a_{ii'}$ qui est borné par 1.

Dans un réseau les sommets ayant pour degré $(N-1)$ possèdent une centralité d'intermédiarité (betweenness centrality) élevée et ils jouent le rôle d'intermédiaires dans le réseau.

Comportement de la fonction $\bar{\phi}^{DI}$ du critère d'Écart à l'Indétermination

Analysons maintenant la fonction caractérisant le terme d'accords négatifs du critère d'Écart à l'Indétermination : $\bar{\phi}^{DI}_{ii'} = \left(\frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2}\right)$. Ainsi, $\bar{\phi}^{DI}$ est une fonction linéaire de $a_{i.}$ et $a_{.i'}$, géométriquement il s'agit d'un plan dans l'espace, donc elle croît moins vite

3. Le seul graphe qui peut avoir un sommet ayant pour degré le nombre total d'arêtes, M, est le graphe étoile. Ainsi, si nous notons u le sommet central d'un graphe étoile, son degré vaut $a_{u.} = M$. Cependant dans ce cas-là le degré de tout autre sommet dans le graphe vaut $a_{.i} = 1 \quad \forall i \neq u$ et le produit $a_{u.}a_{.i} = M < 2M \Rightarrow \bar{\phi}^{NG}_{ui} = \frac{1}{2} \in [0,1] \forall i$ et pour toute autre paire de sommets nous avons $\bar{\phi}^{NG}_{ii'} = \frac{1}{2M} < 1 \forall i \neq u,\ i' \neq u$. Par exemple, le degré du sommet central du graphe étoile de la figure 1.5 est $a_{1.} = 7$ et $\bar{\phi}^{NG}_{1i} = \frac{7}{14} < 1 \forall i \neq 1$ et $\bar{\phi}^{NG}_{ii'} = \frac{1}{14} < 1 \forall i \neq 1,\ i' \neq 1$.

que $\bar{\phi}^{NG}_{ii'}$. Pour la plupart des graphes réels les valeurs prises par cette fonction sont comprises dans l'intervalle $[0,1]$. Lorsque $0 \leq \bar{\phi}^{DI}_{ii'} \leq 1$ est vérifié le domaine de définition de $\phi^{DI} = a_{ii'}$ correspondrait à la borne inférieure et supérieure de l'intervalle de définition de $\bar{\phi}^{DI}$. D'autre part, la fonction $\bar{\phi}^{DI}$ dépend de la structure globale du graphe car elle dépend du nombre total d'arêtes M et du nombre total de sommets N.

La condition de positivité (5.19) définit la borne inférieure de la fonction $\bar{\phi}^{DI} \geq 0$. Voyons maintenant quelle est sa borne supérieure [4]. La quantité $\left(\frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2}\right)$ est maximale lorsque les inégalités $a_{i.} \leq (N-1)$ et $a_{.i'} \leq (N-1)$ deviennent égalités, nous obtenons ainsi la borne supérieure suivante :

$$\bar{\phi}^{DI}_{ii'} \leq \frac{2(N-1)}{N} - \frac{2M}{N^2} = 2 - \frac{2}{N} - \delta. \tag{6.4}$$

Pour des grands graphes réels, $N \to \infty$, le terme $\frac{1}{N}$ tend vers zéro et la densité $\delta << 1$ devient petite (voir section 1.5). Sous ces conditions et la condition de positivité (5.19), nous obtenons les bornes suivantes pour $\bar{\phi}^{DI}_{ii'}$ pour un graphe non pondéré, non réflexif et non orienté :

$$0 \leq \left(\frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2}\right) < 2. \tag{6.5}$$

Les sommets 1 et 2 du graphe de la figure 6.2 possèdent la valeur maximale possible des fonctions $\bar{\phi}^{NG}$ et de la fonction $\bar{\phi}^{DI}$. Il est intéressant de remarquer la position centrale de ces deux sommets dans le réseau.

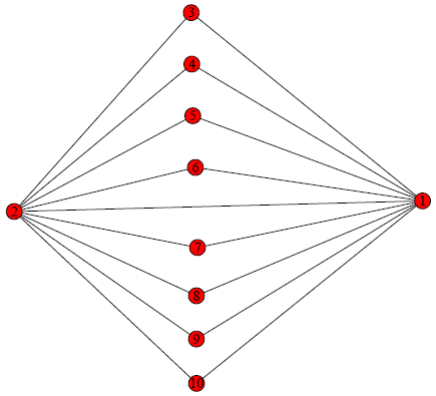

FIGURE 6.2 – Graphe possédant deux sommets, 1 et 2, de degré maximal et égal à $(N-1)$.

Comparaison entre le critère de Newman-Girvan et l'Écart à l'Indétermination

Au chapitre 5 nous avons vu que les critères de Newman-Girvan et le critère d'Écart à l'Indétermination comparaient le graphe à modulariser avec un graphe aléatoire dont le

4. Dans le cas d'un graphe étoile de sommet central u nous avons $\bar{\phi}^{DI}_{ui} = \left(\frac{M}{N} + \frac{a_{.i}}{N} - \frac{2M}{N^2}\right) = \left(\frac{(N-1)}{N} + \frac{1}{N} - \frac{2(N-1)}{N^2}\right) = 1 - \frac{2(N-1)}{N^2} < 1 \in [0,1]$ (car $a_{u.} = M$, $a_{.i} = 1 \, \forall \, i \neq u$ et $M = (N-1)$). Pour toute autre paire de sommets ne contenant pas u nous avons $\bar{\phi}^{DI}_{ii'} = \left(\frac{2}{N} - \frac{2(N-1)}{N^2}\right) = \frac{2}{N^2} < 1 \quad \forall i \neq u, \, i' \neq u$. Donc, pour un graphe étoile nous avons : $\bar{\phi}^{DI}_{ii'} < 1 \, \forall i, i'$

terme général de la matrice d'adjacence est donné par $\bar{\phi}_{ii'}^{NG}$ et $\bar{\phi}_{ii'}^{DI}$ respectivement. Nous avons vu aussi que cela conduit aux calculs simplifiés suivants (voir expressions (5.23) et (5.20)) :

$$\bar{\phi}_{ii'}^{NG} = \sqrt{\bar{\phi}_{ii}^{NG}\bar{\phi}_{i'i'}^{NG}} \qquad \bar{\phi}_{ii'}^{DI} = \frac{\bar{\phi}_{ii}^{DI} + \bar{\phi}_{i'i'}^{DI}}{2}.$$

Ainsi dans le cas d'indépendance, le poids de chaque arête correspond à la moyenne géométrique des poids de boucles des sommets se trouvant à chaque extrémité. Dans le cas d'indétermination, le poids de chaque arête correspond à la moyenne arithmétique des poids de boucles des sommets se trouvant à chaque extrémité. D'autre part la moyenne géométrique de deux nombres étant toujours inférieure à leur moyenne arithmétique (avec égalité si et seulement si les deux nombres sont égaux) nous retrouvons la propriété suivante :

$$\text{Si } \bar{\phi}_{ii}^{NG} = \bar{\phi}_{ii}^{DI} \text{ et } \bar{\phi}_{i'i'}^{NG} = \bar{\phi}_{ii'}^{DI} \Longrightarrow \bar{\phi}_{ii'}^{DI} \geq \bar{\phi}_{ii'}^{NG}.$$

Cependant nous avons toujours $\bar{\phi}_{ii}^{NG} \geq \bar{\phi}_{ii}^{DI} \quad \forall i.$

Démonstration. A partir de $\bar{\phi}_{ii}^{NG} = \frac{a_{i.}^2}{2M}$ et $\bar{\phi}_{ii}^{DI} = \frac{2a_{i.}}{N} - \frac{2M}{N^2}$ nous obtenons :
$\left(\bar{\phi}_{ii}^{NG} - \bar{\phi}_{ii}^{DI}\right) = \left(\frac{a_{i.}^2}{2M} - \frac{2a_{i.}}{N} - \frac{2M}{N^2}\right) = \left(\frac{a_{i.}}{\sqrt{2M}} - \frac{\sqrt{2M}}{N}\right)^2 \geq 0.$ □

Voyons maintenant dans quel cas nous avons $\bar{\phi}_{ii}^{NG} > \bar{\phi}_{ii}^{DI}$ et dans quel cas $\bar{\phi}_{ii}^{NG} < \bar{\phi}_{ii}^{DI}$. Calculons la différence entre ces deux quantités :

$$\left(\bar{\phi}_{ii}^{NG} - \bar{\phi}_{ii}^{DI}\right) = \left(\frac{a_{i.}a_{.i'}}{2M} - \frac{a_{i.}}{N} - \frac{a_{.i'}}{N} + \frac{2M}{N^2}\right) = \left(\frac{a_{i.}}{\sqrt{2M}} - \frac{\sqrt{2M}}{N}\right)\left(\frac{a_{.i'}}{\sqrt{2M}} - \frac{\sqrt{2M}}{N}\right). \tag{6.6}$$

A partir de cette expression et en notant $d_{av} = \frac{2M}{N}$ le degré moyen du graphe nous voyons que selon les valeurs prises par $a_{i.}$ et $a_{.i'}$ 5 cas possibles peuvent se présenter qui divisent le plan cartésien en 4 zones comme le montre le schéma 6.3 :

1. Zone 0 (en rouge) : $(a_{i.} = d_{av}) \cup (a_{.i'} = d_{av})$.
2. Zone I (couleur jaune) : $(a_{i.} < d_{av}) \bigcap (a_{.i'} < d_{av})$.
3. Zone II (couleur bleu) : $(a_{i.} > d_{av}) \bigcap (a_{.i'} < d_{av})$.
4. Zone III (couleur jaune) : $(a_{i.} > d_{av}) \bigcap (a_{.i'} > d_{av})$.
5. Zone IV (couleur bleu) : $(a_{i.} < d_{av}) \bigcap (a_{.i'} > d_{av})$.

La fonction $\bar{\phi}_{ii'}^{NG}$, caractérisant le terme d'accords négatifs du critère de Newman-Girvan, prend des valeurs plus élevées que celle du critère d'Écart à l'Indétermination $\bar{\phi}_{ii'}^{DI}$ lorsque les degrés des sommets i et i' sont soit simultanément supérieurs au degré moyen soit simultanément inférieurs au degré moyen (zones I et III du schéma 6.3). En revanche, la fonction $\bar{\phi}_{ii'}^{DI}$, caractérisant le terme d'accords négatifs du critère d'Écart à l'Indétermination est plus élevée que celle du critère de Newman-Girvan $\bar{\phi}_{ii'}^{NG}$ lorsque i possède un degré supérieur au degré moyen alors que i' possède un degré inférieur au degré moyen ou vice-versa (zones II et IV du schéma 6.3). Les deux fonctions $\bar{\phi}_{ii'}^{NG}$ et $\bar{\phi}_{ii'}^{DI}$ prennent la même valeur lorsqu'au moins un des deux sommets, i ou i', possède le degré

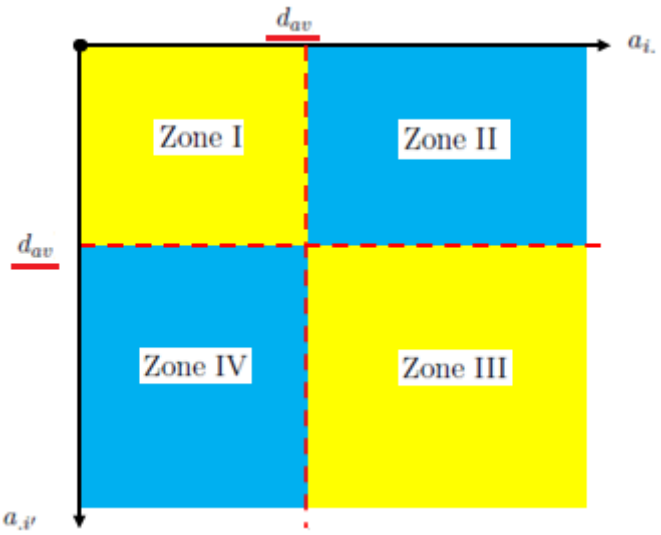

FIGURE 6.3 – Les 4 zones délimitées par le degré moyen.

moyen d_{av} (zone 0).

À ce stade il est important de rappeler, le rôle du terme d'accords négatifs. Si un critère est à maximiser, les fonctions $\bar{\phi}_{ii'}^{NG}$ et $\bar{\phi}_{ii'}^{DI}$ doivent s'accorder avec le terme général de la relation $\bar{\mathbf{X}}$, autrement dit, plus élevée est $\bar{\phi}_{ii'}$ plus $\bar{x}_{ii'}$ devra être égal à 1 pour maximiser le critère, or $\bar{x}_{ii'} = 1$ implique que i et i' seront séparés. C'est-à-dire le critère de Newman-Girvan a tendance à séparer (ne pas classer ensemble) les paires de sommets dont les degrés sont soit simultanément supérieurs au degré moyen soit simultanément inférieurs au degré moyen. En revanche, le critère d'Écart à l'Indétermination a tendance à séparer les paires de sommets contenant un sommet de degré inférieur au degré moyen et un autre sommet de degré supérieur au degré moyen. Par conséquent, les partitions obtenues via l'optimisation du critère de l'Écart à l'Indétermination seront plus homogènes quant à la distribution des degrés intra-classe que celles obtenues via l'optimisation du critère de Newman-Girvan, qui présenteront plus d'hétérogénéité dans la distribution des degrés intra-classe.

6.2.3 Lien entre le critère de Newman-Girvan, l'Écart à l'Indétermination et la Modularité Équilibrée.

Nous avons vu que $\phi_{ii'}^{BM} \neq \phi_{ii'}^{NG}$ et $\phi_{ii'}^{BM} \neq \phi_{ii'}^{DI}$ (voir tableau 6.2), donc pour pouvoir comparer le comportement de ce critère à celui du critère de Newman-Girvan et à celui du critère d'Écart à l'Indétermination, il est nécessaire de prendre en compte l'expression globale du critère (au lieu d'analyser séparément les fonctions $\phi_{ii'}^{BM}$ et $\bar{\phi}_{ii'}^{BM}$). À cet effet, nous exprimerons la Modularité Équilibrée en fonction des deux autres critères mentionnés.

Commençons par le critère de Newman-Girvan. D'abord remarquons que les termes $\bar{\phi}_{ii'}^{NG} = \frac{a_{i.}a_{.i'}}{2M}$ et $\bar{p}_{ii'} = \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M}$ vérifient le lemme suivant :

Lemme 6.1. *Si $a_{i.} = d_{av}$ ou $a_{.i'} = d_{av}$ alors $p_{ii'} + \bar{p}_{ii'} = 1$.*

Démonstration. Si $a_{i.} = d_{av}$ nous avons :

$$\bar{\phi}^{NG}_{ii'} + \bar{p}_{ii'} = \frac{a_{i.}a_{.i'}}{2M} + \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M} = \frac{\frac{2M}{N}a_{.i'}}{2M} + \frac{(N-\frac{2M}{N})(N-a_{.i'})}{N^2-2M} = \frac{a_{i.}}{N} + \frac{(N-a_{.i'})}{N} = 1.$$

La démonstration est analogue si $a_{.i'} = d_{av}$.

□

Le lemme 6.1 implique que les quantités $\bar{\phi}^{NG}_{ii'}$ et $\bar{p}_{ii'}$ sont complémentaires lorsque soit le degré du sommet i soit le degré du sommet d'_i est proche du degré moyen. Cela implique aussi que $a_{i.} = d_{av}$ et $a_{.i'} = d_{av}$ sont solutions du polynôme :

$$\bar{\phi}^{NG}_{ii'} + \bar{p}_{ii'} - 1 = 0.$$

Si nous développons ce polynôme nous obtenons :

$$\begin{aligned}
&\bar{\phi}^{NG}_{ii'} + \bar{p}_{ii'} - 1 = \frac{a_{i.}a_{.i'}}{2M} + \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M} \\
&= \frac{a_{i.}a_{.i'}N^2 - a_{i.}a_{.i'}2M + 2M(N^2 - a_{i.}N - a_{.i'}N + a_{i.}a_{.i'}) - 2MN^2 - 4M^2}{2M(N^2-2M)} \\
&= \frac{a_{i.}a_{.i'}N^2 - a_{i.}a_{.i'}2M + 2MN^2 - 2Ma_{i.}N - 2Ma_{.i'}N + 2Ma_{i.}a_{.i'} - 2MN^2 - 4M^2}{2M(N^2-2M)} \\
&= \frac{a_{i.}a_{.i'}N^2 - 2Ma_{i.}N - 2Ma_{.i'}N - 4M^2}{2M(N^2-2M)} = \frac{(a_{i.}N - 2M)(a_{.i'}N - 2M)}{2M(N^2-2M)}.
\end{aligned}$$

En divisant le numérateur et le dénominateur par N^2 nous obtenons l'égalité suivante en fonction du degré moyen et de la densité d'arêtes :

$$\bar{\phi}^{NG}_{ii'} + \bar{p}_{ii'} = \frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{2M(1-\delta)} + 1. \tag{6.7}$$

À partir de l'expression (6.7) nous avons : $\bar{p}_{ii'} = \frac{(a_{i.}-d_{av})(a_{.i'}-d_{av})}{2M(1-\delta)} + 1 - \bar{\phi}^{NG}_{ii'}$. Si nous remplaçons ce résultat dans l'expression de la Modularité Équilibrée (voir tableau 6.2) nous obtenons :

$$\begin{aligned}
&F_{BM}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(a_{ii'} - \bar{\phi}^{NG}_{ii'} - \bar{a}_{ii'} + \bar{p}_{ii'})x_{ii'} \\
&= \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(2a_{ii'} - 1 - \bar{\phi}^{NG}_{ii'} + \frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{2M(1-\delta)} + 1 - \bar{\phi}^{NG}_{ii'}\right)x_{ii'} \\
&= 2\sum_{i=1}^{N}\sum_{i'=1}^{N}(a_{ii'} - \bar{\phi}^{NG}_{ii'})x_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{2M(1-\delta)}\right)x_{ii'}.
\end{aligned}$$

Par conséquent, maximiser la Modularité Équilibrée revient à maximiser l'expression suivante en fonction du critère de Newman-Girvan :

$$F_{BM} = 2F_{NG} + \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{2M(1-\delta)}\right)x_{ii'}. \tag{6.8}$$

Cette dernière expression montre que maximiser la Modularité Équilibrée revient à maximiser deux fois le critère de Newman-Girvan plus un terme *régulateur* qui dépend de

la distribution des degrés, du degré moyen et de la densité d'arêtes. Le dénominateur de ce terme étant toujours positif, son signe est déterminé par le numérateur. Pour chaque paire de sommets i et i' deux cas possibles peuvent se présenter :

1. $\left(\frac{(a_{i.}-d_{av})(a_{.i'}-d_{av})}{2M(1-\delta)}\right) > 0$ (zones I et III du schéma 6.3). Cela peut s'interpréter ainsi : si le terme régulateur est positif, la Modularité Équilibrée modifiera le critère de Newman-Girvan de sorte à favoriser le fait de classer dans la même communauté les paires de sommets dont les degrés sont soit supérieurs au degré moyen, soit inférieurs au degré moyen simultanément.
2. $\left(\frac{(a_{i.}-d_{av})(a_{.i'}-d_{av})}{2M(1-\delta)}\right) < 0$ (zones II et IV du schéma 6.3). Donc, si le terme régulateur est négatif la Modularité Équilibrée modifie le critère de Newman-Girvan de sorte qu'il favorise la séparation des paires de sommets composées d'un sommet avec degré inférieur au degré moyen et un autre sommet avec degré supérieur au degré moyen.

À partir de ces résultats nous pouvons déduire que la Modularité Équilibrée modifie le critère de Newman-Girvan de façon à homogénéiser la distribution des degrés intra-classe. Autrement dit, si nous distinguons deux catégories de sommets : ceux dont le degré est supérieur au degré moyen et ceux dont le degré est inférieur au degré moyen ; le critère de la Modularité Équilibrée favorise le fait de classer ensemble les sommets se trouvant dans la même catégorie. Cependant, ces résultats dépendent de la contrainte de transitivité sur la partition cherchée $\mathbf{X}$ comme nous le verrons au chapitre 7.

Comparons maintenant l'Écart à l'Indétermination et la Modularité Équilibrée. Les termes $\bar{\phi}^{DI}_{ii'} = \frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2}$ et $\bar{p}_{ii'} = \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M}$ vérifient le lemme suivant :

Lemme 6.2. *Si $a_{i.} = d_{av}$ ou $a_{.i'} = d_{av}$ alors $\bar{\phi}^{DI}_{ii'} + \bar{p}_{ii'} = 1$.*

Démonstration. Si $a_{i.} = d_{av}$ nous avons :

$$\bar{\phi}^{DI}_{ii'} + \bar{p}_{ii'} = \frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2} + \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M} = \frac{2M}{N^2} + \frac{a_{.i'}}{N} - \frac{2M}{N^2} + \frac{(N-\frac{2M}{N})(N-a_{.i'})}{N^2-2M} = \frac{a_{.i'}}{N} + \frac{(N-a_{.i'})}{N} = 1$$

La démonstration est analogue si $a_{.i'} = d_{av}$.

□

Le lemme 6.2 implique que les quantités $\bar{\phi}^{DI}_{ii'}$ et $\bar{p}_{ii'}$ sont complémentaires lorsque d_i ou $d_{i'}$ sont proches du degré moyen. Cela implique aussi que $a_{i.} = d_{av}$ et $a_{.i'} = d_{av}$ sont solutions du polynôme :

$$\bar{\phi}^{DI}_{ii'} + \bar{p}_{ii'} - 1 = 0.$$

Si nous développons ce polynôme nous obtenons :

$$\bar{\phi}^{DI}_{ii'} + \bar{p}_{ii'} - 1 = \frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2} + \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M}$$

$$= \frac{N(N^2-2M)a_{i.}+N(N^2-2M)a_{.i'}-2M(N^2-2M)+N^2(N^2-a_{i.}N-a_{.i'}N+a_{i.}a_{.i'})-N^2(N^2-2M)}{N^2(N^2-2M)}$$

$$= \frac{N^3 a_{i.} - N2Ma_{i.} + N^3 a_{.i'} - N2Ma_{.i'} - 2MN^2 + 4M^2 + N^4 - a_{i.}N^3 - a_{.i'}N^3 + a_{i.}a_{.i'}N^2 - N^4 + N^2 2M}{N^2(N^2 - 2M)}$$

$$= \frac{a_{i.}a_{.i'}N^2 - N2Ma_{i.} - N2Ma_{.i'} + 4M^2}{2M(N^2 - 2M)} = \frac{(a_{i.}N - 2M)(a_{.i'}N - 2M)}{N^2(N^2 - 2M)}.$$

En divisant le numérateur et dénominateur par N^2 nous obtenons l'expression suivante qui dépend du degré moyen et de la densité d'arêtes :

$$\bar{\phi}_{ii'}^{DI} + \bar{p}_{ii'} = \frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{N^2(1-\delta)} + 1. \tag{6.9}$$

À partir des expressions (6.7) et (6.9) nous pouvons exprimer $\bar{\phi}_{ii'}^{NG}$ et $\bar{p}_{ii'}$ en fonction de $\bar{\phi}_{ii'}^{DI}$:

$$\bar{\phi}_{ii'}^{NG} = \frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{(1-\delta)} \left(\frac{1}{2M} - \frac{1}{N^2} \right) + \bar{\phi}_{ii'}^{DI}, \tag{6.10}$$

$$\bar{p}_{ii'} = \frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{N^2(1-\delta)} - \bar{\phi}_{ii'}^{DI} + 1. \tag{6.11}$$

Si nous remplaçons ces résultats dans l'expression de la Modularité Équilibrée (voir tableau 6.2), nous obtenons :

$$F_{BM}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} (a_{ii'} - \bar{\phi}_{ii'}^{NG} - \bar{a}_{ii'} + \bar{p}_{ii'}) x_{ii'}$$

$$= \sum_{i=1}^{N}\sum_{i'=1}^{N} \left(2a_{ii'} - \frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{(1-\delta)} \left(\frac{1}{2M} - \frac{1}{N^2} \right) - 2\bar{\phi}_{ii'}^{DI} + \frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{N^2(1-\delta)} \right) x_{ii'}$$

$$= 2\sum_{i=1}^{N}\sum_{i'=1}^{N} (a_{ii'} - \bar{\phi}_{ii'}^{DI}) x_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N} \left(\frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{(1-\delta)} \left(\frac{2}{N^2} - \frac{1}{2M} \right) \right) x_{ii'}.$$

Par conséquent, maximiser la Modularité Équilibrée revient à maximiser l'expression suivante en fonction du critère d'Écart à l'Indétermination :

$$F_{BM} = 2F_{DI} + \left(2 - \frac{1}{\delta} \right) \sum_{i=1}^{N}\sum_{i'=1}^{N} \left(\frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{N^2(1-\delta)} \right) x_{ii'}. \tag{6.12}$$

La dernière expression montre que maximiser la Modularité Équilibrée revient à maximiser deux fois le critère d'Écart à l'Indétermination plus un terme *régulateur* qui dépend de la distribution des degrés, du degré moyen et de la densité d'arêtes. Le dénominateur de ce terme est toujours positif, cependant ce terme est pré-multiplié par la quantité $\left(2 - \frac{1}{\delta}\right)$ qui s'annule si et seulement si la densité d'arêtes $\alpha = 0,5$, comme nous avons vu lors de la définition de densité, (voir chapitre 1, section 1.5) pour des graphes réels $\delta << 0.5$ comme le montre le tableau 1.1. Cela implique que pour des réseaux réels nous aurons dans la plupart de cas $\left(2 - \frac{1}{\delta}\right) < 0$. Désormais nous supposant que le terme $\left(2 - \frac{1}{\delta}\right)$ est négatif. Donc, pour chaque paire de sommet i et i' nous avons les deux cas suivants :

1. Le terme régulateur est négatif lorsque deux sommets ont des degrés simultanément supérieurs au degré moyen ou simultanément inférieurs au degré moyen (zones I et III du schéma 6.3) :

$\left(2-\frac{1}{\delta}\right)\left(\frac{(a_{i.}-d_{av})(a_{.i'}-d_{av})}{N^2(1-\delta)}\right) < 0$. Par conséquent, la Modularité Équilibrée modifiera le critère d'Écart à l'Indétermination de façon à défavoriser le fait de classer ces sommets dans la même communauté.

2. Le terme régulateur est positif pour toute paire de sommets composée d'un sommet avec degré inférieur au degré moyen et d'un autre sommet avec degré supérieur au degré moyen (zones II et IV du schéma 6.3) :
$\left(2-\frac{1}{\delta}\right)\left(\frac{(a_{i.}-d_{av})(a_{.i'}-d_{av})}{N^2(1-\delta)}\right) > 0$. Donc, la Modularité Équilibrée, modifiera le critère d'Écart à l'Indétermination de façon à favoriser le fait de les mettre dans la même classe).

À partir de ces résultats nous déduisons que la Modularité Équilibrée modifie le critère d'Écart à l'Indétermination de façon à hétérogénéiser la distribution des degrés intra-classe. Cependant, ces résultats dépendent de la contrainte de transitivité sur la partition cherchée $\mathbf{X}$.

Nous avons vu que les partitions obtenues via l'optimisation du critère de Newman-Girvan présentent plus d'hétérogénéité dans la distribution des degrés intra-classe que celles obtenues via l'optimisation du critère d'Écart à l'Indétermination. Nous avons vu aussi que la Modularité Équilibrée modifiait le critère de Newman-Girvan avec un terme régulateur qui permettait d'homogénéiser la distribution des degrés intra-classe. Nous venons de voir que la Modularité Équilibrée modifie le critère d'Écart à l'Indétermination avec un terme régulateur qui permet d'hétérogénéiser la distribution des degrés intra-classe. Nous pouvons déduire de ces résultats que la Modularité Équilibrée se comporte comme un critère régulateur entre les critères de Girvan-Newman et celui d'Écart à l'Indétermination.

6.2.4 Lien entre les quatre critères dépendant de la distribution des degrés : Newman-Girvan, l'Écart à l'Indétermination, l'Écart à l'Uniformité et la Modularité Équilibrée.

Théorème 6.1. *Pour un graphe dont la distribution des degrés est uniforme, i.e. tous les sommets du graphe ont le même degré qui est égal au degré moyen $d_{av} = \frac{2M}{N}$, les termes d'accords négatifs des critères de Newman-Girvan, Écart à l'Indétermination et Écart à l'Uniformité coïncident :*

$$\bar{\phi}^{NG}_{ii'} = \bar{\phi}^{DI}_{ii'} = \bar{\phi}^{\text{UNIF}}_{ii'}.$$

Par conséquent, les trois critères sont équivalents.
De plus, maximiser ces critères est équivalent à maximiser la Modularité Équilibrée :

$$\underset{X}{Max} F_{NG} = \underset{X}{Max} F_{DI} = \underset{X}{Max} F_{\text{UNIF}} = \underset{X}{Max} F_{BM}.$$

Démonstration. Si tous les sommets ont le même degré $a_{i.} = d_{av} = \frac{2M}{N} \quad \forall i$ alors :

- $\bar{\phi}^{NG}_{ii'} = \dfrac{a_{i.}a_{.i'}}{2M} = \dfrac{\frac{2M}{N}\frac{2M}{N}}{2M} = \dfrac{2M}{N^2} = \delta = \bar{\phi}^{\text{UNIF}}$ et par conséquent : $F_{NG} = F_{\text{UNIF}}$.

- $\bar{\phi}^{DI}_{ii'} = \frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2} = \frac{\frac{2M}{N}}{N} + \frac{\frac{2M}{N}}{N} - \frac{2M}{N^2} = \frac{2M}{N^2} = \delta = \bar{\phi}^{\text{UNIF}}$ et par conséquent : $F_{NG} = F_{\text{UNIF}}$.

- Nous pouvons déduire du tableau 6.2 que maximiser la Modularité Équilibrée revient à maximiser :
$F_{BM}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} + \frac{(N-a_{i.})(N-a_{.i'})}{N^2-2M} - \left(\bar{a}_{ii'} + \bar{\phi}^{NG}_{ii'}\right)\right)x_{ii'}$, si tous les sommets ont le même degré :
$F_{BM}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} + \frac{(N-\frac{2M}{N})(N-\frac{2M}{N})}{N^2-2M} - (1-a_{ii'}) - \frac{2M}{N^2}\right)x_{ii'} =$
$F_{BM}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(2a_{ii'} + \frac{(N^2-2M)(N^2-2M)}{N^2(N^2-2M)} - 1 - \frac{2M}{N^2}\right)x_{ii'} =$
$F_{BM}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(2a_{ii'} + \frac{(N^2-2M)}{N^2} - 1 - \frac{2M}{N^2}\right)x_{ii'} =$
$\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(2a_{ii'} + 1 - \frac{2M}{N^2} - 1 - \frac{2M}{N^2}\right)x_{ii'} = 2\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'} - \frac{2M}{N^2}\right)x_{ii'} = 2F_{\text{UNIF}}$.
La dernière expression étant deux fois le critère d'Écart à l'Uniformité, nous pouvons conclure que pour un graphe dont les arêtes sont équi-réparties parmi les sommets, l'optimisation des quatre critères : Newman-Girvan, Écart à l'Indétermination, Modularité Équilibrée et Écart à l'Uniformité) conduit au même résultat, à la même partition optimale.

□

La figure 6.4 compare les fonctions $\bar{\phi}^{NG}$, $\bar{\phi}^{DI}$ et $\bar{\phi}^{\text{UNIF}}$ pour le réseau de musiciens de "jazz", à $N = 198$ sommets et $M = 2742$ arêtes, mentionné dans le tableau 6.1 (voir Gleiser and Danon [2003]).

La figure 6.4 montre comme la fonction $\bar{\phi}^{NG}$ croît plus rapidement que $\bar{\phi}^{DI}$ au fur et à mesure que les degrés de sommets augmentent, alors que la fonction $\bar{\phi}^{\text{UNIF}}$ reste constante pour toute paire de sommets.

Jusqu'ici nous avons étudié les comportements des fonctions $\bar{\phi}$ des critères linéaires décrits dans le tableau 6.2. Nous avons vu les liens existant entre les différentes fonctions. Dans la section suivante nous allons nous servir de ces résultats pour expliquer la variation que les différents critères présentent quant au nombre de classes de la partition optimale dont deux exemples ont été donnés dans le tableau 6.1.

6.2.5 Coût de fusion de deux classes pour les critères linéaires

Nous avons vu dans le tableau 6.1 qu'il existe des différences importantes quant au nombre optimal de classes trouvé par le biais de l'optimisation des différents critères linéaires. Par exemple, le critère de Zahn-Condorcet et celui d'Écart à l'Uniformité génèrent beaucoup plus de classes que les autres critères. Tandis que les critères de Newman-Girvan, la Modularité Équilibrée, l'Écart à l'Indétermination et l'Écart à l'Uniformité tendent à générer peu de classes et à peu près le même nombre. Cependant cela ne garantit pas

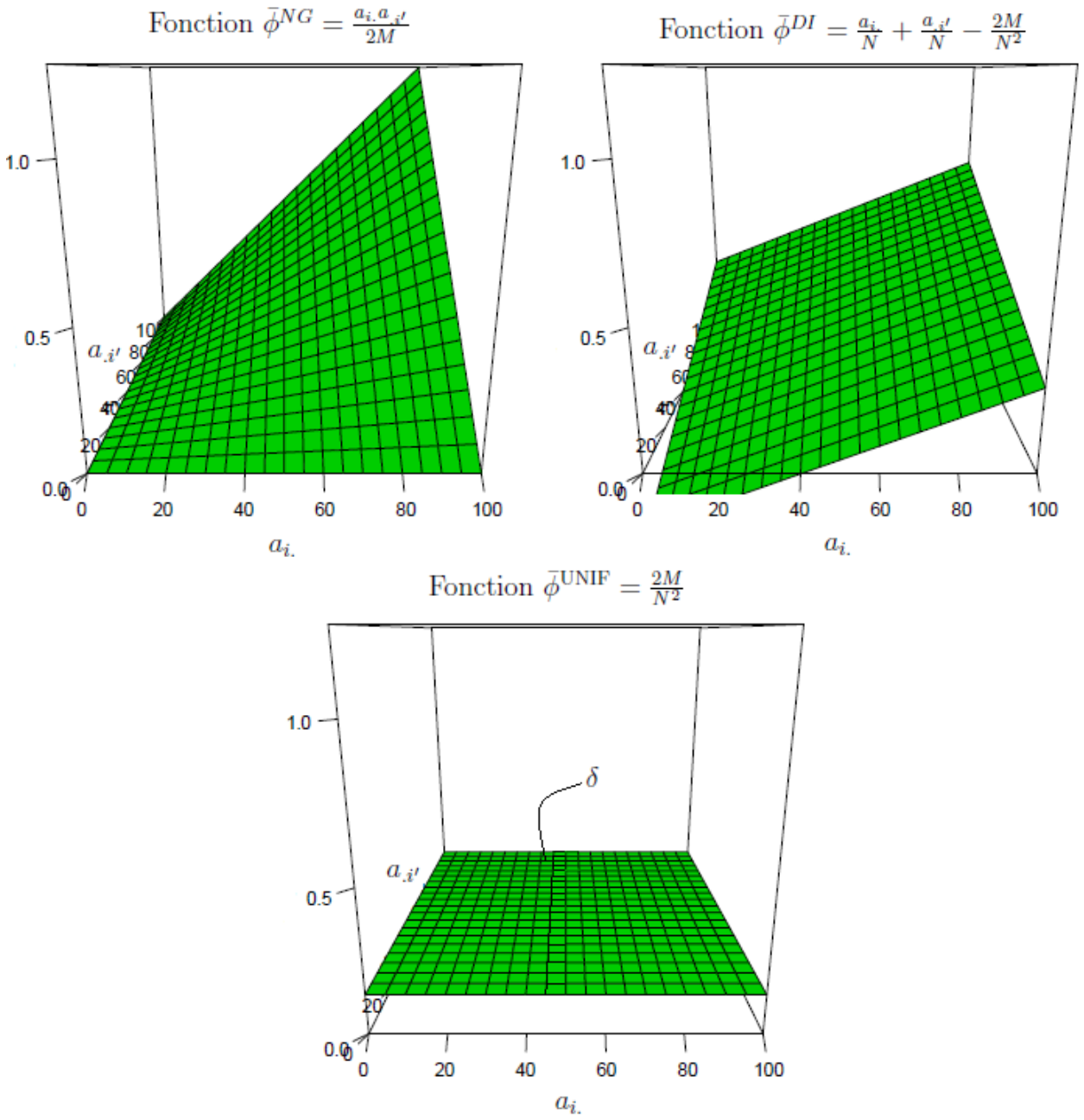

FIGURE 6.4 – Comparaison des fonctions $\bar{\phi}^{NG}$, $\bar{\phi}^{DI}$ et $\bar{\phi}^{\text{UNIF}}$ pour le réseau de musiciens de "jazz" de Gleiser and Danon [2003].

que les classes trouvées avec ces quatre critères soient identiques. Dans cette section nous allons expliquer ces différences et nous allons aussi comprendre le rôle des termes d'accords négatifs et d'accords positifs dans la détermination du nombre optimal de classes.

Pour pouvoir comprendre pourquoi certains critères génèrent plus de classes que d'autres nous allons calculer l'impact sur la valeur de chaque critère suite à la fusion de deux sous-

graphes dans le réseau. Supposons que nous voulons fusionner deux sous-graphes ou classes $\mathcal{C}_1$ et $\mathcal{C}_2$ dans le graphe de tailles respectives n_1 et n_2. Les sommets de $\mathcal{C}_1$ et de $\mathcal{C}_2$ sont connectées par l arêtes, comme le montre la figure 6.5 :

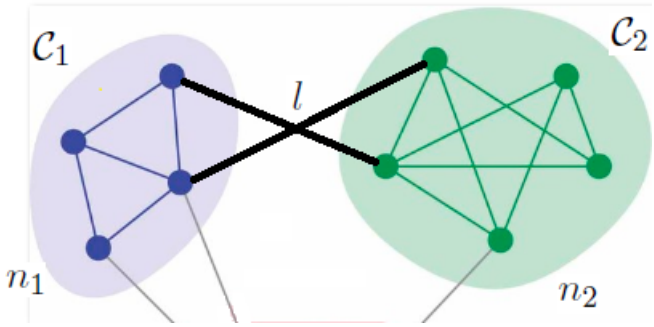

Figure 6.5 – Deux classes dans le réseau dont la fusion aura un impact sur la valeur du critère de modularisation.

L'impact de la fusion de ces deux classes sur la valeur d'un critère linéaire sera quantifié par une variable que nous nommerons *contribution* car elle représentera la contribution à la valeur du critère suite à la fusion. Ainsi, la contribution C à la valeur d'un critère linéaire F suite à la fusion de deux classes est calculée par l'expression générale suivante :

$$C_F = \sum_{i \in \mathcal{C}_1}^{n_1} \sum_{i' \in \mathcal{C}_2}^{n_2} (\phi_{ii'} - \bar{\phi}_{ii'}). \tag{6.13}$$

Deux cas possibles peuvent se présenter [5]

- Si $C_F > 0 \Longrightarrow \sum_{i \in \mathcal{C}_1}^{n_1} \sum_{i' \in \mathcal{C}_2}^{n_2} \phi_{ii'} > \sum_{i \in \mathcal{C}_1}^{n_1} \sum_{i' \in \mathcal{C}_2}^{n_2} \bar{\phi}_{ii'}$ l'optimisation du critère est pour la fusion des deux classes, donc la contribution est un gain pour la valeur du critère.

- Si $C_F < 0 \Longrightarrow \sum_{i \in \mathcal{C}_1}^{n_1} \sum_{i' \in \mathcal{C}_2}^{n_2} \phi_{ii'} < \sum_{i \in \mathcal{C}_1}^{n_1} \sum_{i' \in \mathcal{C}_2}^{n_2} \bar{\phi}_{ii'}$ l'optimisation du critère est contre la fusion des deux classes, donc la contribution est un coût, une perte pour la valeur du critère.

Démonstration. La preuve consiste à comparer la valeur du critère avant et après la fusion. La contribution est la différence entre la valeur d'*après* et la valeur d'*avant*. Étant donné une partition, la valeur de tout critère linéaire pour cette partition peut s'exprimer comme la somme des contributions de toutes ses classes. Ainsi, avant la fusion le critère vaut (nous notons cette quantité F_B pour *before* en anglais) :

$$F_B = \sum_{i=1}^{N} \sum_{i'=1}^{N} (\phi_{ii'} - \bar{\phi}_{ii'}) x_{ii'} = \sum_{i \in \mathcal{C}_1}^{n_1} \sum_{i' \in \mathcal{C}_1}^{n_1} (\phi_{ii'} - \bar{\phi}_{ii'}) + \sum_{i \in \mathcal{C}_2}^{n_2} \sum_{i' \in \mathcal{C}_2}^{n_2} (\phi_{ii'} - \bar{\phi}_{ii'}) + K,$$

5. Ce raisonnement est valable lorsque le critère est à maximiser, s'il est à minimiser il faut raisonner dans le sens inverse. Pour notre étude, tous les critères linéaires que nous avons présentés sont à maximiser.

où K représente les contributions des autres classes différentes de $\mathcal{C}_1$ et de $\mathcal{C}_2$.
Après la fusion le critère vaut (nous notons cette quantité F_A pour *after* en anglais) :

$$F_A = \sum_{i \in (\mathcal{C}_1 \cup \mathcal{C}_2)}^{n_1+n_2} \sum_{i' \in (\mathcal{C}_1 \cup \mathcal{C}_2)}^{n_1+n_2} (\phi_{ii'} - \bar{\phi}_{ii'}) + K.$$

Comme les classes $\mathcal{C}_1$ et $\mathcal{C}_2$ contiennent des éléments disjoints, la dernière expression peut se réécrire comme :

$$F_A = \sum_{i \in \mathcal{C}_1}^{n_1} \sum_{i' \in \mathcal{C}_1}^{n_1} (\phi_{ii'} - \bar{\phi}_{ii'}) + \sum_{i \in \mathcal{C}_2}^{n_2} \sum_{i' \in \mathcal{C}_2}^{n_1} (\phi_{ii'} - \bar{\phi}_{ii'}) + 2 \sum_{i \in \mathcal{C}_1}^{n_1} \sum_{i' \in \mathcal{C}_2}^{n_2} (\phi_{ii'} - \bar{\phi}_{ii'}) + K.$$

Où le coefficient 2 devant l'avant dernier terme vient de la symétrie des fonctions ϕ et $\bar{\phi}$. Maintenant la contribution $C_F = (F_A - F_B)$ s'écrit :

$$C_F = \sum_{i \in \mathcal{C}_1}^{n_1} \sum_{i' \in \mathcal{C}_2}^{n_2} (\phi_{ii'} - \bar{\phi}_{ii'})$$

.

Nous omettons la constante 2 car elle n'a aucune influence sur le signe de C_F.

Ce résultat peut être obtenu de façon visuelle en définissant une matrice dont le terme général vaut : $(\phi_{ii'} - \bar{\phi}_{ii'})x_{ii'}$. Cette matrice est diagonale par blocs, et la valeur du critère est la somme de termes de chaque bloc. Chaque bloc représente une classe. Cette matrice est symétrique aussi. La figure 6.6 montre l'évolution de la matrice "avant" et "après" la fusion.

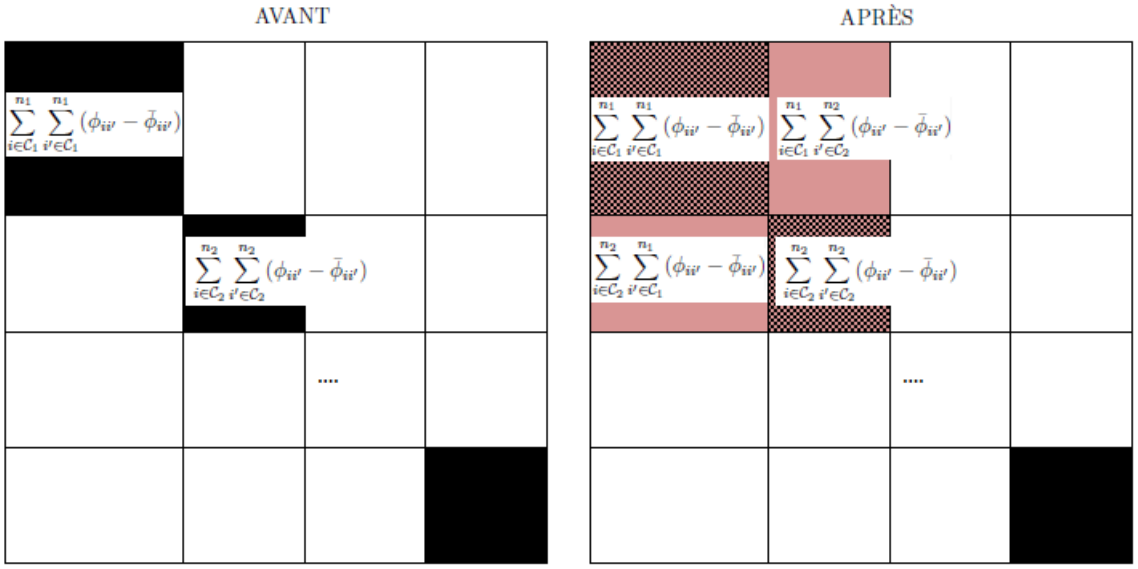

FIGURE 6.6 – Évolution de la matrice de terme général $(\phi_{ii'} - \bar{\phi}_{ii'})x_{ii'}$ avant et après la fusion.

Grâce à la symétrie de cette matrice nous obtenons :

$$\sum_{i\in\mathcal{C}_1}^{n_1}\sum_{i'\in\mathcal{C}_2}^{n_2}(\phi_{ii'}-\bar{\phi}_{ii'}) = \sum_{i\in\mathcal{C}_2}^{n_2}\sum_{i'\in\mathcal{C}_1}^{n_1}(\phi_{ii'}-\bar{\phi}_{ii'})$$

,

d'où le résultat du calcul de la contribution. □

Remarques sur l'expression de la contribution (6.13) :

- La décision de fusionner ou pas fusionner les deux classes est prise en faisant une comparaison de la quantité $\sum_{i\in\mathcal{C}_1}^{n_1}\sum_{i'\in\mathcal{C}_2}^{n_2}\phi_{ii'}$ versus la quantité $\sum_{i\in\mathcal{C}_1}^{n_1}\sum_{i'\in\mathcal{C}_2}^{n_2}\bar{\phi}_{ii'}$. La première quantité représente la contribution au terme d'accords positifs des toutes les paires de sommets ayant une composante dans $\mathcal{C}_1$ et l'autre dans $\mathcal{C}_2$. La deuxième quantité représente la contribution au terme d'accords négatifs des toutes les paires de sommets ayant une composante dans $\mathcal{C}_1$ et l'autre dans $\mathcal{C}_2$.

- Les deux quantités $\sum_{i\in\mathcal{C}_1}^{n_1}\sum_{i'\in\mathcal{C}_2}^{n_2}\phi_{ii'}$ et $\sum_{i\in\mathcal{C}_1}^{n_1}\sum_{i'\in\mathcal{C}_2}^{n_2}\bar{\phi}_{ii'}$ étant positives, c'est celle qui a la valeur la plus élevée qui décidera de fusionner ou pas fusionner. Ainsi la première quantité est "pour" la fusion tandis que l'autre est "contre" la fusion.

- C'est ici où on peut voir pourquoi les fonctions $\phi(.)$ et $\bar{\phi}(.)$ définissent la propriété d'équilibre linéaire et pourquoi elles jouent le rôle de "coûts". Elles représentent le coût qu'il faut payer pour fusionner.

- Cela explique aussi l'obtention de la solution triviale ou grossière lors de l'absence de $\phi(.)$ ou de $\bar{\phi}(.)$ respectivement. En effet, si $\phi_{ii'} = 0\,\forall(i,i') \Rightarrow C_F < 0$ et par conséquent le critère sera toujours contre la fusion et la solution optimale sera la solution triviale où tous les sommets sont isolés. En revanche, si $\bar{\phi}_{ii'} = 0\,\forall(i,i') \Rightarrow C_F > 0$ et par conséquent le critère sera toujours pour la fusion et la solution optimale sera la solution grossière où tous les sommets sont classés en une seule et classe.

L'expression (6.13), obtenue à partir de l'écriture générale de critères linéaires équilibrés (4.5), nous montre le pouvoir unificateur de la notation relationnelle. Le fait d'avoir écrit tous les critères linéaires avec les mêmes notations de base nous permet de généraliser le calcul de la contribution pour tout critère linéaire.

le tableau 6.3 montre l'expression explicite de la contribution aux critères linéaires étudiés jusqu'à présent. Toutes les expressions sont obtenues à partir de la formule (6.13).
Où :

- $d^1_{av} = \dfrac{\sum_{i\in\mathcal{C}_1}^{n_1} a_{i.}}{n_1}$ et $d^2_{av} = \dfrac{\sum_{i'\in\mathcal{C}_2}^{n_2} a_{.i'}}{n_2}$ représentent le degré moyen de $\mathcal{C}_1$ et $\mathcal{C}_2$ respectivement.

Critère : F	$\phi_{ii'}$	$\bar{\phi}_{ii'}$	$C_F = \sum_{i\in\mathcal{C}_1}^{n_1}\sum_{i'\in\mathcal{C}_2}^{n_2}(\phi_{ii'} - \bar{\phi}_{ii'})$
Zahn-Condorcet	$a_{ii'}$	$\frac{1}{2}$	$C_{ZC} = \left(l - \frac{n_1 n_2}{2}\right)$
Owsiński-Zadrożny	$a_{ii'}$	α	$C_{OZ} = (l - \alpha n_1 n_2) \quad 0 < \alpha < 1$
Écart à l'Uniformité	$a_{ii'}$	$\frac{2M}{N^2}$	$C_{\text{UNIF}} = \left(l - n_1 n_2 \frac{2M}{N^2}\right)$
Newman-Girvan	$a_{ii'}$	$\frac{a_{i.}a_{.i'}}{2M}$	$C_{NG} = \left(l - n_1 n_2 \frac{d^1_{av} d^2_{av}}{2M}\right)$
Écart à l'Indétermination	$a_{ii'}$	$\frac{a_{i.}}{N} + \frac{a_{.i'}}{N} - \frac{2M}{N^2}$	$C_{DI} = \left(l - n_1 n_2 \left(\frac{d^1_{av}}{N} + \frac{d^2_{av}}{N} - \frac{2M}{N^2}\right)\right)$

TABLE 6.3 – Contribution suite à la fusion de deux sous-graphes pour les critères linéaires.

- Pour les cinq critères nous avons $\phi_{ii'} = a_{ii'}$ et par conséquent $\sum_{i\in\mathcal{C}_1}^{n_1}\sum_{i\in\mathcal{C}_2}^{n_2}\phi_{ii'} = l$, soit le nombre d'arêtes existant entre $\mathcal{C}_1$ et $\mathcal{C}_2$.

Nous allons maintenant étudier plus en détail et analyser les conséquences des expressions montrées dans le tableau 6.3. L'annexe B montre le détail d'obtention de résultats et le tableau 6.4 présente un résumé.

Conséquences sur le nombre optimal de classes

Si nous avons toujours $\delta << 0,5$ (comme le montre le tableau 1.1 pour des graphes réels), la contribution associée au critère de Zahn-Condorcet sera toujours inférieure à celle associée au critère d'Écart à l'Uniformité, en effet,

$$C_{ZC} - C_{\text{UNIF}} = \left(l - \frac{n_1 n_2}{2}\right) - (l - n_1 n_2 \delta) = n_1 n_2 \left(\delta - \frac{1}{2}\right) < 0.$$

Ce résultat implique que quelles que soient les caractéristiques des classes à fusionner, si la contribution au critère de Zahn-Condorcet est positive, la contribution au critère d'Écart à l'Uniformité le sera aussi. En revanche, le fait que la contribution au critère d'Écart à l'Uniformité soit positive n'implique nullement que celle associée au critère de Zahn-Condorcet le soit aussi. Par conséquent, le nombre de classes obtenu via l'optimisation du critère de Zahn-Condorcet sera toujours supérieur à celui obtenu via l'optimisation du critère d'Écart à l'Uniformité :

$$\kappa_{ZC} > \kappa_{\text{UNIF}}, \tag{6.14}$$

où κ_{ZC} et κ_{UNIF} dénotent le nombre optimal de classes obtenu via l'optimisation du critère de Zahn-Condorcet et du critère d'Écart à l'Uniformité respectivement.

Par analogie, nous pouvons établir les résultats suivants concernant le nombre de classes obtenu via l'optimisation du critère d'Owsiński-Zadrożny

$$\begin{aligned} \kappa_{ZC} > \kappa_{OZ} \quad &\text{si} \quad \alpha < 0,5, \\ \kappa_{ZC} < \kappa_{OZ} \quad &\text{si} \quad \alpha > 0,5, \end{aligned} \tag{6.15}$$

où κ_{OZ} dénote le nombre optimal de classes obtenu via l'optimisation du critère d'Owsiński-Zadrożny.

Et de façon analogue, le nombre optimal de classes obtenu via l'optimisation du critère d'Owsiński-Zadrożny et celui obtenu via l'optimisation de l'Écart à l'Uniformité vérifient :

$$\begin{aligned} \kappa_{\text{UNIF}} > \kappa_{OZ} \quad &\text{si} \quad \alpha < \delta \\ &\text{et} \\ \kappa_{\text{UNIF}} < \kappa_{OZ} \quad &\text{si} \quad \alpha > \delta. \end{aligned} \tag{6.16}$$

Ces résultats peuvent être constatés facilement avec les données du tableau 6.1. La densité d'arêtes du réseau *Jazz* étant $\delta \simeq 0,14$ et $\alpha = 0,02$ nous avons $\kappa_{ZC} > \kappa_{\text{UNIF}} > \kappa_{OZ}$. En ce qui concerne le réseau *Internet*, la densité d'arêtes est $\delta \simeq 0,00014$ et $\delta < \alpha < 0,01$ l'optimisation des trois critères donne : $\kappa_{ZC} > \kappa_{OZ} > \kappa_{\text{UNIF}}$.

Comportement du critère de Newman-Girvan et du critère de Zahn-Condorcet avec les sommets à degré 1

Le fait de placer le sommet à degré 1 dans la classe de son seul voisin augmente toujours le critère de Newman-Girvan. Nous allons voir que ce comportement du critère de Newman-Girvan est contraire à celui du critère de Zahn-Condorcet et il a déjà été démontré par Brandes et al. [2008] :

Démonstration. La preuve consiste à comparer la contribution à la modularité de Newman-Girvan d'un partitionnement où le sommet de degré 1 est seul dans sa classe avec la contribution d'un partitionnement où le sommet de degré 1 est dans la classe de son voisin.

Soit v le sommet de degré 1, donc $a_{v.} = a_{.v} = 1$ et u son voisin. Si v est tout seul dans sa classe $\mathcal{C}(v)$ et u est dans une autre classe $\mathcal{C}(u)$ la contribution de ces deux classes à la modularité de Newman-Girvan, notée $NG[\mathcal{C}(u)] + NG[\mathcal{C}(v)]$ sera :

$$NG[\mathcal{C}(u)] + NG[\mathcal{C}(v)] = \frac{1}{2M} \sum_{(i,i')\in\mathcal{C}(u)} (a_{ii'} - \frac{a_{i.}a_{.i'}}{2M}) - \frac{1}{4M^2}.$$

En fusionnant les classes $\mathcal{C}(v)$ et $\mathcal{C}(u)$, la contribution de cette classe résultante à la modularité, notée $NG[\mathcal{C}(u) \cup \mathcal{C}(v)]$ sera :

$$NG[\mathcal{C}(u) \cup \mathcal{C}(v)] = \frac{1}{2M} \sum_{(i,i')\in\mathcal{C}(u)} (a_{ii'} - \frac{a_{i.}a_{.i'}}{2M}) - \frac{1}{4M^2} + \frac{2}{2M} \sum_{i\in C(u)} (a_{iv} - \frac{a_{i.}}{2M}).$$

La différence entre la contribution à la modularité après et avant la fusion sera :
$NG[\mathcal{C}(u) \cup \mathcal{C}(v)] - (NG[\mathcal{C}(u)] + NG[\mathcal{C}(v)]) = \frac{1}{M} \sum_{i\in\mathcal{C}(u)} (a_{iv} - \frac{a_{i.}}{2M})$

$a_{iv} = 1$ seulement dans le cas où $i = v$, autrement il est nul :

$$NG[\mathcal{C}(u) \cup \mathcal{C}(v)] - (NG[\mathcal{C}(u)] + NG[\mathcal{C}(v)]) = \frac{1}{M} - \frac{1}{M} \sum_{i\in\mathcal{C}(u)} \frac{a_{i.}}{2M}.$$

Le membre droit de cette dernière équation est strictement supérieur à zéro car $\sum_{i\in\mathcal{C}(u)} a_{i.} < 2M$ donc suite à la fusion des deux classes la modularité ne peut qu'augmenter sa valeur. □

Ce comportement du critère de Newman-Girvan est contraire à celui du critère de Zahn-Condorcet sous certaines conditions. Comme nous l'avons mentionné lors de l'analyse de la contribution au critère de Zahn-Condorcet, les classes obtenues via l'optimisation de celui-ci possèdent plusieurs classes à un seul sommet isolé. Nous pouvons nous inspirer de la démonstration précédente pour le démontrer :

Théorème 6.2 (Le critère de Zahn-Condorcet et les sommets de degré 1)**.** *Étant donné un graphe connecté $G = (V, E)$, le critère de Zahn-Condorcet classe tout sommet de degré* 1 *soit tout seul dans une classe, soit avec son seul voisin dans une classe où il y a au plus* 1 *autre sommet à part son voisin.*

Si v est un sommet de degré 1 de G et u son seul voisin le théorème 6.2 indique que si u est classé avec son voisin il y aura au plus 3 sommets dans cette classe.

Démonstration. De façon analogue à la démonstration précédente nous allons comparer la contribution au critère de Zahn-Condorcet dans le cas où le sommet v est classé tout seul et dans le cas où il est classé avec son voisin u.

Si v est classé tout seul, soit $\mathcal{C}(v)$ sa classe et $\mathcal{C}(u)$ la classe de son voisin. La contribution de ces deux classes au critère de Zahn-Condorcet, notée $ZC[\mathcal{C}(u)] + ZC[\mathcal{C}(v)]$ sera :

$$ZC[\mathcal{C}(u)] + ZC[\mathcal{C}(v)] = \sum_{(i,i')\in\mathcal{C}(u)}(a_{ii'} - \bar{a}_{ii'}) - 1.$$

Le dernier terme est dû au fait que $a_{vv} = 0$, donc $\bar{a}_{vv} = 1$.

En fusionnant les classes $\mathcal{C}(v)$ et $\mathcal{C}(u)$, la contribution de cette classe résultante au critère, notée $ZC[\mathcal{C}(u) \cup \mathcal{C}(v)]$ sera :

$$ZC[\mathcal{C}(u) \cup \mathcal{C}(v)] = \sum_{(i,i')\in\mathcal{C}(u)}(a_{ii'} - \bar{a}_{ii'}) - 1 + 2\sum_{i\in\mathcal{C}(u)}(a_{iv} - \bar{a}_{iv}).$$

La différence entre la contribution après et avant la fusion sera :

$$ZC[\mathcal{C}(u) \cup \mathcal{C}(v)] - (ZC[\mathcal{C}(u)] + ZC[\mathcal{C}(v)]) = 2\sum_{i\in\mathcal{C}(u)}(a_{iv} - \bar{a}_{iv}).$$

$a_{iv} = 1$ seulement dans le cas où $i = u$ et dans ce cas-là $\bar{a}_{iv} = 0$, pour $i \neq u$ on a $a_{iv} = 0$ ce qui implique $\bar{a}_{iv} = 1$:

$$ZC[\mathcal{C}(u) \cup \mathcal{C}(v)] - (ZC[\mathcal{C}(u)] + ZC[\mathcal{C}(v)]) = 2 - 2(|\mathcal{C}(u)| - 1) = 4 - 2|\mathcal{C}(u)|,$$

où $|\mathcal{C}(u)|$ est l'effectif de la classe contenant u avant la fusion, donc $|\mathcal{C}(u)| \geq 1$. On a les cas suivants :

1. Si $|\mathcal{C}(u)| = 1$ donc u est tout seul dans sa classe et la contribution au critère suite à la fusion des classes est positive, cela implique que l'optimisation du critère de Zahn-Condorcet placera u et v dans la même classe.
2. Si $|\mathcal{C}(u)| = 2$ il y a un autre sommet dans la classe de u à part u et dans ce cas-là le fait de fusionner les classes n'augmente ni diminue la valeur du critère.

3. Si $|\mathcal{C}(u)| > 2$ il y a au moins deux sommets à part u dans $\mathcal{C}(u)$ et dans ce cas l'optimisation du critère placera v seul dans sa classe.

Ces résultats sont une conséquence de la vérification de la règle de majorité absolue d'arêtes intra-classe imposée par le critère de Zahn-Condorcet. □

Nous allons illustrer ce résultat avec le petit graphe à 6 sommets de la figure 5.3. Après avoir modularisé ce graphe via l'optimisation du critère de Newman-Girvan et du critère de Zahn-Condorcet nous avons trouvé que chaque critère fournit une partition différente. Une partition à 2 classes pour le premier et à 3 classes pour le second.

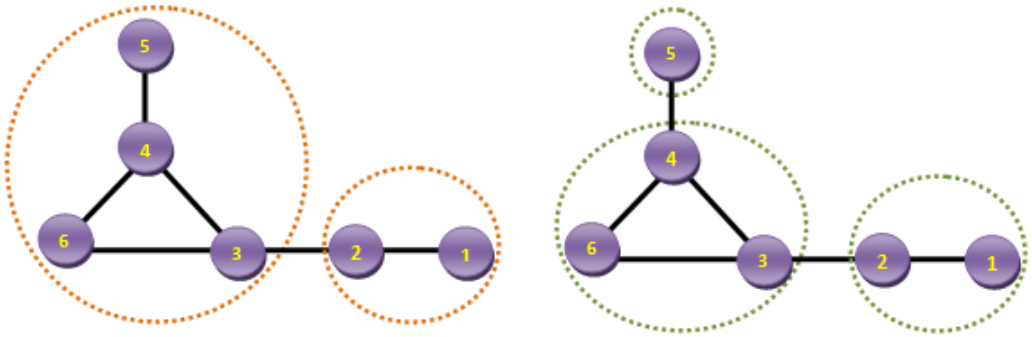

Figure 6.7 – Résultat de la modularisation du graphe de la figure 5.3. A gauche le résultat obtenu avec le critère de Newman-Girvan et à droite celui obtenu avec le critère de Zahn-Condorcet.

La figure 6.7 montre que le sommet 5 a été classé dans la classe de son voisin (le sommet 4), par le critère de Newman-Girvan tandis que le critère de Zahn-Condorcet a l'a isolé. Le sommet 5 de la figure 6.7 n'étant pas en relation avec les sommets 3 et 6 Zahn-Condorcet le isole. Quant au critère de Newman-Girvan, il met le sommet 5 dans la même classe que son voisin même s'il est connecté à un seul sommet dans cette classe car ce critère n'isole pas les sommets de degré 1. Concernant le sommet 1, il a été classé avec son seul voisin, le sommet 2, par les deux méthodes. Cela est dû principalement au fait que le sommet 2 n'a que deux voisins : le sommet 1 et le sommet 3.

Il est intéressant de remarquer que, pour cet exemple, toutes les classes obtenues par le critère de Zahn-Condorcet sont des graphes complets, des cliques à 1, 2 ou 3 sommets.

Comparaison entre les contributions au critère de Newman-Girvan, au critère d'Écart à l'Indétermination et au critère de la Modularité Équilibrée

Nous pouvons mener une analyse de comparaison entre les contributions au critère de Newman-Girvan et au critère d'Écart à l'Indétermination de façon similaire à la section précédente. Si nous comparons les contributions aux deux critères :

$$C_{NG} = \left(l - n_1 n_2 \frac{d^1_{av} d^2_{av}}{2M} \right)$$

et

$$C_{DI} = \left(l - n_1 n_2 \left(\frac{d_{av}^1}{N} + \frac{d_{av}^2}{N} - \frac{2M}{N^2} \right) \right)$$

.

La différence entre ces deux expressions réside dans les valeurs prises par les quantités : $\frac{d_{av}^1 d_{av}^2}{2M}$ et $\left(\frac{d_{av}^1}{N} + \frac{d_{av}^2}{N} - \frac{2M}{N^2} \right)$. Nous pouvons étudier la variation entre ces deux quantités :

$$\left(\frac{d_{av}^1 d_{av}^2}{2M} - \left(\frac{d_{av}^1}{N} + \frac{d_{av}^2}{N} - \frac{2M}{N^2} \right) \right) = \frac{1}{2M} \left(d_{av}^1 - \frac{2M}{N} \right) \left(d_{av}^2 - \frac{2M}{N} \right)$$

.

Donc, c'est la valeur du produit $(d_{av}^1 - d_{av})(d_{av}^2 - d_{av})$ qui nous intéresse. Selon la valeur que cette quantité prend, il peut exister les différences suivantes entre la contribution au critère de Newman-Girvan C_{NG} et la contribution au critère de l'Écart à l'Indétermination C_{DI} :

1. Si le degré moyen d'au moins une des classes $\mathcal{C}_1$ ou $\mathcal{C}_2$ est proche du degré moyen d_{av}, alors $(d_{av}^1 - d_{av})(d_{av}^2 - d_{av}) = 0$ et par conséquent $C_{NG} = C_{DI}$ ce qui implique que la décision de fusionner ou pas les deux classes est la même pour les deux critères.

2. Si $(d_{av}^1 - d_{av})(d_{av}^2 - d_{av}) > 0 \Longrightarrow C_{NG} < C_{DI}$ (zones I et III du schéma 6.3). Cela signifie que le critère d'Écart à l'Indétermination sera plus facilement en faveur de la fusion des deux classes que le critère de Newman-Girvan lorsque les deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$ possèdent simultanément soit un degré moyen supérieur au degré moyen global soit un degré moyen inférieur au degré moyen global.

3. Si $(d_{av}^1 - d_{av})(d_{av}^2 - d_{av}) < 0 \Longrightarrow C_{NG} > C_{DI}$. Cela peut être interprété comme suit : si le degré moyen d'une des deux classes ($\mathcal{C}_1$ ou $\mathcal{C}_2$) est supérieur au degré moyen du graphe tandis que le degré moyen de l'autre classe a pour degré moyen une valeur inférieure au degré moyen global le critère de Newman-Girvan soutiendra plus en faveur de la fusion des deux classes que le critère d'Écart à l'Indétermination.

À partir des expressions (6.8) et (6.12) nous pouvons exprimer la contribution de la Modularité Équilibrée en fonction des expressions de la contribution aux critères de Newman-Girvan et d'Écart à l'Indétermination respectivement. De plus, en nous basant sur la définition de contribution (6.13) nous obtenons :

$$C_{BM} = 2 \sum_{i \in \mathcal{C}_1}^{n_1} \sum_{i' \in \mathcal{C}_2}^{n_2} \left(a_{ii'} - \frac{a_{i.} a_{.i'}}{2M} \right) + \sum_{i \in \mathcal{C}_1}^{n_1} \sum_{i' \in \mathcal{C}_2}^{n_2} \frac{(a_{i.} - d_{av})(a_{.i'} - d_{av})}{2M(1-\delta)}$$

$$C_{BM} = 2l - 2n_1 n_2 \frac{d_{av}^1 d_{av}^2}{2M} + n_1 n_2 \frac{(d_{av}^1 - d_{av})(d_{av}^2 - d_{av})}{2M(1-\delta)}$$

$$\boxed{C_{BM} = 2C_{NG} + n_1 n_2 \frac{(d_{av}^1 - d_{av})(d_{av}^2 - d_{av})}{2M(1-\delta)}.} \quad (6.17)$$

En suivant les mêmes étapes que pour le calcul de la contribution à la Modularité Équilibrée en fonction de la contribution au critère d'Écart à l'Indétermination nous trouvons :

$$C_{BM} = 2C_{DI} + n_1 n_2 \left(2 - \frac{1}{\delta}\right) \frac{(d^1_{av} - d_{av})(d^2_{av} - d_{av})}{N^2(1-\delta)}. \tag{6.18}$$

Ces résultats montrent ce qui a déjà été constaté dans la section précédente.

La contribution à la Modularité Équilibrée est deux fois la contribution au critère de Newman-Girvan plus un terme *régulateur* : $\left(n_1 n_2 \left(2 - \frac{1}{\delta}\right) \frac{(d^1_{av} - d_{av})(d^2_{av} - d_{av})}{N^2(1-\delta)}\right)$ qui dépend de la distribution des degrés, du degré moyen et de la densité d'arêtes. Ce terme régulateur sera en faveur de la fusion si les degrés moyens des deux classes sont soit simultanément supérieurs soit simultanément inférieurs au degré moyen global (zones I et III du schéma 6.3) . En revanche, il sera en défaveur de la fusion si le degré moyen d'une des classes est inférieur au degré moyen global et le degré moyen de l'autre classe est supérieur au degré moyen (zones II et IV du schéma 6.3).

La contribution à la Modularité Équilibrée est deux fois la contribution au critère d'Écart à l'Indétermination plus un terme *régulateur* qui vaut $\left(n_1 n_2 \left(2 - \frac{1}{\delta}\right) \frac{(d^1_{av} - d_{av})(d^2_{av} - d_{av})}{N^2(1-\delta)}\right)$. Ce terme dépend de la distribution des degrés, du degré moyen et de la densité d'arêtes. La quantité $\left(2 - \frac{1}{\delta}\right)$ étant la plupart de temps négative pour des graphes réels (voir le tableau 1.1), le terme régulateur sera en faveur de la fusion si le degré moyen d'une des classes est inférieur au degré moyen global et le degré moyen de l'autre classe est supérieur au degré moyen (zones II et IV du schéma 6.3). En revanche, il défavorisera la fusion si les degrés moyens des deux classes sont soit simultanément supérieurs soit simultanément inférieurs au degré moyen global (zones I et III du schéma 6.3).

Cependant pour des grands graphes, lorsque $N \longrightarrow \infty$ et $M \longrightarrow \infty$ les termes régulateurs tendent vers zéro.

L'effet de l'ajout d'un sommet

Une application importante du coût de fusion de deux classes est l'étude de l'ajout d'un sommet au graphe et le suivi de l'évolution du partitionnement optimale suite à cet ajout. À titre d'exemple, nous allons étudier comment l'arrivée d'un nouveau sommet modifie la classification optimale obtenue avant la modification par les critères de Newman-Girvan et celui de Zahn-Condorcet.

Les graphes modélisent phénomènes non statiques, c'est-à-dire qui évoluent avec le temps. Ainsi avec le temps certains sommets abandonnent le réseau, d'autres y adhérent, les liens entre sommets changent aussi, etc. Par exemple, dans les réseaux sociaux il y a de nouveaux amis qui arrivent, les amitiés se créent, certains liens s'affaiblissent, etc.

Théorème 6.3. *Étant donné un graphe $G = (V, E)$, une modularisation optimale de G ne contient pas de classes à sommets dis-connectés.*

Le théorème 6.3 est une conséquence de la définition de la notion de modularisation et de celle de communauté. En effet, si dans une partition il existe une classe composée de groupes de sommets dis-connectés, le fait de placer chaque sous-groupe dans une classe différente n'entraîne aucune coupure car il n'existe aucune arête reliant les sous-groupes de sommets. Donc, un bon critère de modularisation placera chaque sous-groupe dans une classe différente.

Quelques conséquences du théorème 6.3 pour une partition optimale :

- Ayant obtenu une partition à modularisation optimale ; deux sommets d'une même classe peuvent être reliés par un chemin passant seulement par les sommets de cette classe.
- Chaque sommet du graphe soit il est classé tout seul dans une classe, soit il a au moins un voisin dans sa classe.
- S'il existe des classes à 2 sommets, ces deux sommets sont forcément adjacents.

La figure 6.8 illustre un exemple d'application du théorème 6.3. Un bon critère de modularité aura une valeur plus importante pour la partition de droite que pour celle de gauche.

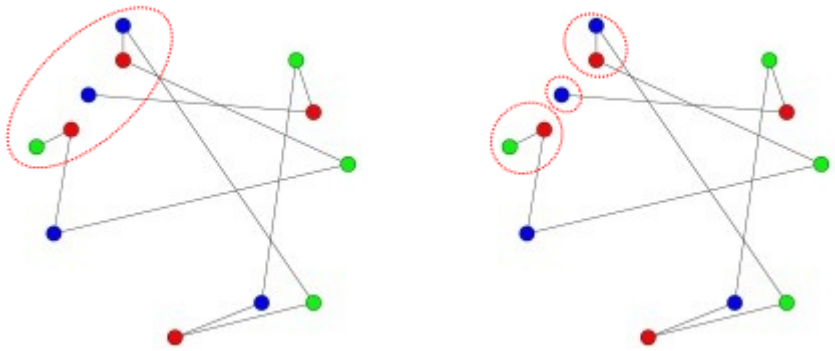

FIGURE 6.8 – À gauche : Partitionnement non optimal avec une classe à sous-groupes de sommets non connectés. À droite : La classe à sous-groupes dis-connectés a été séparée en 3 classes.

Par la suite, nous allons analyser l'effet sur le critère de modularité de Newman-Girvan et celui de Zahn-Condorcet suite à l'ajout d'un sommet au graphe.

Soit un graphe $G = (V, E)$ à $N = |V|$ sommets, Soit X^* la matrice de la partition optimale obtenue avec le critère de Zahn-Condorcet. Supposons que l'on ajoute un nouveau sommet v à G adjacent à d_v sommets et nous obtenons un nouveau graphe G' à $(N + 1)$ sommets.

La valeur optimale du critère de Zahn-Condorcet dans sa version à maximiser pour le graphe G s'écrit :

$$F_{ZC}(X^*) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(2a_{ii'} - 1)x^*_{ii'} + \sum_{i=1}^{N}\sum_{i'=1}^{N}\bar{a}_{ii'}.$$

Suite à l'ajout du sommet v nous devons maximiser la fonction :

$$F_{ZC}(X) = \sum_{i=1}^{N+1}\sum_{i'=1}^{N+1}(2a_{ii'}-1)x_{ii'} + \sum_{i=1}^{(N+1)}\sum_{i'=1}^{(N+1)}\bar{a}_{ii'}$$
$$= \sum_{i=1}^{N}\sum_{i'=1}^{N}(2a_{ii'}-1)x_{ii'} + 2\sum_{i=1}^{N}(2a_{iv}-1)x_{iv} - 1 + \sum_{i=1}^{(N+1)}\sum_{i'=1}^{(N+1)}\bar{a}_{ii'}.$$

Comme v est en relation avec d_v sommets et il n'est pas en relation avec $N-d_v$ sommets, et si on ordonne les sommets de façon que les d_v premiers soient les voisins de v, nous pouvons écrire sans perte de généralité :

$$F_{ZC}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(2a_{ii'}-1)x_{ii'} + 2(\sum_{i=1}^{d_v}x_{iv} - \sum_{i=d_v+1}^{N}x_{iv}) - 1 + \sum_{i=1}^{(N+1)}\sum_{i'=1}^{(N+1)}\bar{a}_{ii'}.$$

Cela revient à maximiser :

$$F_{ZC}(X) = 2(\sum_{i'>i}^{N}(2a_{ii'}-1)x_{ii'} + \sum_{i=1}^{d_v}x_{iv} - \sum_{i=d_v+1}^{N}x_{iv}) + K, \tag{6.19}$$

où le dernier terme est une constante égale à $K = -(N+1) + \sum_{i=1}^{(N+1)}\sum_{i'=1}^{(N+1)}\bar{a}_{ii'}$.

Maximiser l'équation (6.19) revient à maximiser l'expression entre parenthèses. On sait que $\mathbf{X}^*$ maximise le premier terme. S'il n'y avait pas la contrainte de transitivité en choisissant : $x_{iv} = 1$ si $i \in [1, d_v]$ (si i est voisin de v), 0 sinon, on maximiserait les autres deux termes. Malheureusement ce choix n'est pas possible car les voisins de v peuvent être classés en différentes classes par la partition $\mathbf{X}^*$.

Nous allons trouver la valeur optimale du critère pour le cas $d_v = 1$, c'est-à-dire le cas où le nouveau sommet est adjacent à un seul sommet que nous nommerons u. Dans ce cas l'équation (6.19) devient :

$$F_{ZC}(X) = 2(\sum_{i'>i}^{N}(2a_{ii'}-1)x_{ii'} + x_{uv} - \sum_{i=2}^{N}x_{iv}) + K. \tag{6.20}$$

Selon le théorème 6.3 dans une partition optimale soit v est classé tout seul dans une classe, soit il est classé avec son voisin u. :

- Si v est classé tout seul le critère de Zahn-Condorcet vaut
  $F_{ZC}(X) = 2\sum_{i'>i}^{N}(2a_{ii'}-1)x^*_{ii'} + K$
- si u selon la partition $\mathbf{X}^*$ était classé seul dans une classe, l'ajout de v dans cette classe augmente la valeur du critère :
  $F_{ZC}(X) = 2(\sum_{i'>i}^{N}(2a_{ii'}-1)x^*_{ii'} + 1) + K$
- Si la classe contenant u selon la partition X^* contenait un autre sommet w à part u suite à l'ajout de v dans cette classe le critère vaudrait :
  $F_{ZC}(X) = 2(\sum_{i'>i}^{N}(2a_{ii'}-1)x^*_{ii'} + 1 - 1) + K$ car $x_{vw} = 1$

- Si la classe contenant u selon la partition $\mathbf{X}^*$ contenait $|\mathcal{C}(u)| \geq 3$ sommets au total, l'ajout de v dans cette classe ne peut que diminuer la valeur du critère : $F_{ZC}(X) = 2(\sum_{i'>i}^{N}(2a_{ii'} - 1)x^*_{ii'} + 2 - |\mathcal{C}(u)|) + K$ étant donné que $|\mathcal{C}(u)| \geq 3$ il est préférable ne pas classer v avec u sinon l'isoler.

Nous avons montré que l'ajout d'un nouveau sommet de degré 1 au graphe G ne modifie pas complètement la partition obtenue avec le critère de Zahn-Condorcet avant l'ajout, le sommet ne perturbe que la classe de son voisin.

Analysons maintenant l'impact de l'ajout d'un sommet sur la valeur du critère de Newman-Girvan. Il a été montré dans Brandes et al. [2008] que suite à l'ajout d'un seul sommet de degré 1 la partition obtenue avec le critère de Newman-Girvan peut changer radicalement et même avoir un effet sur d'autres sommets éloignés du nouveau sommet v et de son voisin u. Cela est dû principalement au fait que le terme $p_{ii'} = \frac{a_{i.}a_{.i'}}{2M}$ change pour toutes les paires de sommets car M change.

Étant donné une communauté $\mathcal{C}$ et un sommet u voisin à quelques sommets dans $\mathcal{C}$, le critère de Newman-Girvan impose comme la condition suivante au nouveau sommet pour faire partie de la communauté :

$$a^{in}_{u.} > a^{out}_{u.} \frac{\sum_{i\in\mathcal{C}} a_{.i}}{2M - \sum_{i\in\mathcal{C}} a_{.i}}, \tag{6.21}$$

où :

- $a^{in}_{u.}$ représente le nombre d'arêtes entre u et les sommets appartenant à $\mathcal{C}$
- $a^{out}_{u.}$ représente le nombre d'arêtes entre u et les sommets appartenant au reste su réseau.

Ainsi au fur et à mesure que $M \to \infty$ la fraction du terme droit de l'inégalité (6.21) devient de plus en plus petit par rapport au terme de gauche, donc le nouveau sommet sera plus facilement accepté dans $\mathcal{C}$. C'est une conséquence de la limite de résolution de ce critère, donc, il ne possède pas la propriété d'**invariance d'échelle**.

Démonstration. Nous pouvons démontrer l'inégalité (6.21) à partir du calcul de la contribution à la valeur du critère suite à la fusion de deux classes (voir tableau 6.3). Ici $\mathcal{C}_1$ est le sommet u de degré d_u. Pour que le critère soit en faveur de la fusion il faut que la contribution soit positive :

$$C_{NG} = l - \frac{|\mathcal{C}|d_u d^{\mathcal{C}}_{av}}{2M} > 0$$

, où $d^{\mathcal{C}}_{av}$ est le degré moyen des sommets dans $\mathcal{C}$, ici $l = a^{in}_{u.}$, $|\mathcal{C}|d^{\mathcal{C}}_{av} = \sum_{i\in\mathcal{C}} a_{.i}$ et $d_u = a^{in}_{u.} + a^{out}_{u.}$. En remplaçons ces quantités dans le calcul de la contribution nous obtenons :

$$C_{NG} = a^{in}_{u.} - \frac{(a^{in}_{u.} + a^{out}_{u.})\sum_{i\in\mathcal{C}} a_{.i}}{2M} > 0 \Leftrightarrow a^{in}_{u.} > a^{out}_{u.} \frac{\sum_{i\in\mathcal{C}} a_{.i}}{2M - \sum_{i\in\mathcal{C}} a_{.i}}$$

.

□

6.2.6 Conclusion comparaison des critères linéaires

Le tableau 6.4 présente un résumé des caractéristiques principales des partitionnements obtenus via l'optimisation des six critères étudiés : Zahn-Condorcet, Owsiński-Zadrożny, Écart à l'Uniformité, Newman-Girvan, Ecart à l'Indétermination et Modularité Equilibrée. Le détail d'obtention de ces résultats sont montrés dans l'annexe B.

À partir du tableau 6.4 et des résultats obtenus précédemment nous déduisons les liens suivants entre les critères linéaires :

Pour $\alpha < 0,5$ l'optimisation du critère d'Owsiński-Zadrożny rendra moins de petites cliques que le critère de Zahn-Condorcet. En effet, il est plus flexible que le critère de Zahn-Condorcet. Si $\delta << 0,5$ (comme pour la plupart des graphes réels) le nombre de classes obtenu via l'optimisation du critère de Zahn-Condorcet sera toujours supérieur à celui obtenu via l'optimisation du critère d'Écart à l'Uniformité.

Les partitions obtenues via l'optimisation du critère de Newman-Girvan présentent plus d'hétérogénéité quant à la distribution des degrés intra-classe que celles obtenues via l'optimisation du critère d'Écart à l'Indétermination. En effet, l'Écart à l'Indétermination a toujours tendance à favoriser les grandes classes ayant un degré moyen élevé et les petites classes ayant un degré moyen faible.

La Modularité Équilibrée se comporte comme un critère régulateur entre les critères de Newman-Girvan et celui d'Écart à l'Indétermination. D'une part, la Modularité Équilibrée modifie le critère de Newman-Girvan de façon à homogénéiser la distribution des degrés intra-classe. D'autre part, la Modularité Équilibrée modifie le critère d'Écart à l'Indétermination de sorte à hétérogénéiser la distribution des degrés intra-classe. Cependant les partitions trouvées dépendent fortement de la distribution des arêtes et des degrés, donc, en pratique nous ne trouverons pas forcément plus ou moins d'hétérogénéité dans la distribution des degrés intra-classe des classes trouvées avec les critères mentionnés. De plus, la partition cherchée $\mathbf{X}$ dépend de la contrainte de transitivité que toute relation d'équivalence doit vérifier.

Il est aussi important de mentionner que pour des grands graphes, lorsque $N \longrightarrow \infty$ et $M \longrightarrow \infty$ les termes régulateurs de la Modularité Équilibrée deviennent négligeables. Cela explique le fait que le nombre optimal de classes obtenu via l'optimisation des 3 critères : Newman-Girvan, Écart à l'Indétermination et Modularité Équilibrée soit proche (comme pour les exemples du tableau 6.1).

On trouvera à l'annexe B une analyse de comparaison des critères non linéaires.

Critère	**Caractéristiques de la partition optimale**
Zahn-Condorcet	◦ Chaque sommet est connecté au moins à la moitié des sommets dans sa classe. Donc, la densité d'arêtes intra-classe est supérieure à 0,5 pour toutes les classes. ◦ Ce critère ne possède pas de limite de résolution. Ce qui lui confère la propriété d'invariance d'échelle. ◦ Inconvénient : parfois la solution optimale contient des petites classes, cliques à 2 ou 3 sommets ou sommets isolés.
Owsiński-Zadrożny	◦ Généralisation du critère de Zahn-Condorcet où α définit la densité minimale d'arêtes intra-classe. ◦ Chaque sommet est connecté au moins à $\alpha\%$ de sommets dans sa classe. ◦ Il ne possède pas de limite de résolution donc il possède la propriété d'invariance d'échelle. ◦ Inconvénient : le choix de la valeur α qui dépend de l'utilisateur et revient indirectement à fixer le nombre de classes.
Écart à l'Uniformité	◦ C'est un cas particulier du critère d'Owsiński-Zadrożny avec $\alpha = \delta = \frac{2M}{N^2}$ (la densité d'arêtes du graphe). ◦ Les classes obtenues possèdent une densité d'arêtes intra-classe supérieure ou égale à la densité d'arêtes globale δ. ◦ Ce critère possède une limite de résolution.
Newman-Girvan	◦ Le critère dépend de la distribution des degrés. ◦ Ce critère possède une limite de résolution. ◦ La partition optimale ne possède pas de classes à un seul sommet de degré 1.
Écart à l'Indétermination	◦ Il dépend de la distribution des degrés. ◦ Il possède une limite de résolution. ◦ Les données doivent vérifier une condition de positivité [6]. Le non-respect de celle-ci peut entrainer que la solution optimale contienne des classes à composantes non connexes.
Modularité Équilibrée	◦ Le critère dépend de la distribution des degrés. ◦ Ce critère possède une limite de résolution.

TABLE 6.4 – Partitions obtenues via l'optimisation des critères linéaires.

Chapitre 7

Applications

7.1 Introduction

Dans ce chapitre nous allons voir l'application pratique des résultats présentés dans les chapitres précédents. Pour cela, nous allons nous utiliser de graphes réels de taille différente (comme données d'entrée) ainsi que de l'algorithme de Louvain générique pour approcher la solution optimale de chaque critère. Plus précisément, nous allons chercher la solution optimale de huit critères : Zhan-Condorcet, Owsiński-Zadrożny, Écart à l'Uniformité, Newman-Girvan, écart à l'Indétermination, la Modularité Équilibrée, la Différence de Profils et Michalski-Goldberg.

Dans un premier temps nous allons parler de l'importance de l'utilisation des heuristiques [1] ad-hoc pour résoudre le problème (5.1) sous les contraintes (5.2). Ensuite, nous allons décrire quelques algorithmes connus de clustering de graphes et nous allons présenter l'algorithme de Louvain générique (voir Blondel et al. [2008] et Campigotto et al. [2013]). Finalement, nous allons présenter et comparer les partitions obtenues via l'optimisation des huit critères de modularisation cités précédemment.

7.2 Le nombre de partitions d'un ensemble fini : le nombre de Bell

Le nombre de Bell (du mathématicien Éric Temple Bell), noté B_N, est le nombre de façons de découper un ensemble V à N objets en partitions distinctes. Il s'agit du nombre de Relations d'équivalence ou Partitions constructibles à partir d'un ensemble à N éléments.

Le nombre de Bell, B_N, peut être calculé à partir du nombre de Stirling de deuxième espèce. Ce dernier n'est autre que le nombre de partitions de N éléments en κ classes et il est calculé par l'expression suivante :

$$S_{(N,\kappa)} = \frac{1}{\kappa!}\sum_{j=1}^{\kappa}(-1)^{\kappa-j}\binom{\kappa}{j}j^N. \tag{7.1}$$

1. Une heuristique est une méthode de calcul qui fournit en temps polynomial ou raisonnable une solution réalisable, pas nécessairement optimale, pour un problème d'optimisation **NP**-difficile.

Une approximation des nombres de Stirling lorsque $N \to \infty$ est donnée par $S_{(N,\kappa)} \cong \frac{\kappa}{\kappa!}$ (cette formule est assez peu précise pour des petites valeurs de N sauf en début de séquence, c'est-à-dire pour κ relativement faible comparativement à N). Bien évidemment, le nombre de Bell sera calculé à partir des nombres de Stirling de deuxième espèce selon l'expression :

$$B_N = \sum_{i=1}^{N} S_{(N,i)}, \quad N \geq 1. \tag{7.2}$$

Le nombre de Bell peut être calculé avec la formule de récurrence suivante :

$$B_{N+1} = \sum_{i=0}^{N} \binom{N}{k} B_i. \tag{7.3}$$

D'autre part, les nombres $S_{(N,\kappa)}$ satisfont l'équation de récurrence suivante :

$$S_{(N,\kappa)} = S_{(N-1,\kappa-1)} + \kappa S_{(N-1,\kappa)}. \tag{7.4}$$

Voici quelques valeurs intéressantes des nombres de Bell : $B_1 = 1$, $B_2 = 2$, $B_3 = 5$, $B_4 = 15$, $B_5 = 52$, $B_6 = 203$, $B_7 = 877$, $B_8 = 4\ 140, \ldots, B_{7116} \cong 4{,}08 \cdot 10^{74}$.

Étant donné un critère de modularisation on pourrait penser que la recherche de la partition optimale est un choix facile : il suffirait de considérer toutes les partitions possibles et de choisir celle qui optimise le critère. Cependant, cette tâche est insurmontable car le nombre de partitions devient vite astronomique. Par exemple, pour $N = 10000$ sommets (ce qui n'est pas vraiment une grosse taille, dès lors que l'on s'intéresse à des applications pratiques du Business Intelligence, ou du "Customer Relationship Management (CRM)"), le nombre de Bell nous donne quand même le résultat suivant :

$$1{,}987 \cdot 10^{26\ 014} < B_{10000} < 2{,}80 \cdot 10^{29\ 344}$$

.

Un ordinateur pouvant traiter un million de partitions par seconde mettrait plus de 10^{26014} années pour trouver la solution optimale.

Il devient, alors nécessaire d'approcher la solution optimale au moyen d'heuristiques qui donnent une approximation de bonne qualité en des temps de calcul raisonnables.

Rappelons à ce propos que la solution au problème d'optimisation linéaire à variables booléennes (2.6) caractérisé par des inégalités linéaires (5.2) :

$$F_C(X) = \sum_{k=1}^{M} \left(\sum_{i=1}^{N} \sum_{i'=1}^{N} (c^k_{ii'} x_{ii'} + \bar{c}^k_{ii'} \bar{x}_{ii'}) \right),$$

sous les contraintes linéaires sur $\mathbf{X}$ caractérisant une Relation d'équivalence, à savoir

$x_{ii'} \in \{0,1\}$		Binarité
$x_{ii} = 1$	$\forall i$	Réflexivité
$x_{ii'} - x_{i'i} = 0$	$\forall (i,i')$	Symétrie
$x_{ii'} + x_{i'i''} - x_{ii''} \leq 1$	$\forall (i,i',i'')$	Transitivité

a été trouvée pour la première fois par Marcotorchino and Michaud [1981] dans le cadre de la résolution du "Problème des Partitions Centrales" (voir chapitre 2). Le principe de résolution consiste à relâcher, dans un premier temps, la contrainte de binarité pour la mettre sous la forme $0 \leq x_{ii'} \leq 1$, des méthodes de coupe pouvant être utilisées ultérieurement pour obtenir des solutions bivalentes. Ensuite, la résolution se fait à partir du problème dual (voir Marcotorchino and Michaud [1981] et Brenac [2002]). Cependant, ce problème sous contraintes linéaires est **NP**-complet. Par conséquent, il s'avère qu'au-delà de quelques petites centaines d'objets, sa résolution exacte devient impraticable et seules des heuristiques "ad-hoc" permettent d'en donner une approximation de bonne qualité en temps raisonnable.

7.3 Algorithmes existants

Il existe plusieurs types d'algorithmes de clustering de graphes.

1. Les **algorithmes de division** cherchent à détecter les liens inter-communautaires pour les retirer du réseau au fur et à mesure. Un exemple très connu de ce type d'algorithmes est l'algorithme de Girvan-Newman (voir Girvan and Newman [2002]). Cet algorithme cherche à enlever itérativement les arêtes possédant une mesure de centralité d'intermédiarité (*betweenness*) élevée. Bien que les résultats à la sortie de cet algorithme soient de bonne qualité, il est lent et peu pratique pour des réseaux de plusieurs milliers de sommets. D'autres algorithmes de division sont décrits dans Newman and Girvan [2004] et Radicchi et al. [2004]. C'est le cas par exemple des algorithmes de type *Minimum-cut* (comme les Ratio-cut et Normalized-cut mentionnés au chapitre 5). On peut dire que ces algorithmes utilisent le principe de la classification descendante hiérarchique puisque, au début, tous les objets sont dans une seule communauté, puis ils sont séparés en classes de plus en plus petites.

2. Les **algorithmes d'agglomération** fusionnent récursivement les sommets en fonction d'une mesure de similarité entre chaque paire de sommets (voir par exemple l'algorithme de Pons and Latapy [2006]). Les méthodes de classification ascendante hiérarchique ("hierarchical clustering") appartiennent à cette catégorie. Il existe plusieurs règles pour effectuer le regroupement, les deux plus simples étant *single-linkage clustering* (regroupement simple), dans laquelle deux groupes sont considérés comme des communautés distinctes si et seulement si toutes les paires de sommets dans différents groupes ont une similarité inférieure à un seuil donné, et *complete linkage*

clustering (regroupement complet), dans laquelle toutes les paires de sommets au sein de chaque groupe ont une similarité supérieure au seuil.

3. Les **algorithmes cherchant à optimiser une fonction qualité prédéfinie**. Dans la plupart des cas, cette fonction qualité est la modularité de Newman-Girvan. Il a été prouvé que maximiser la modularité de Newman-Girvan est un problème **NP**-complet [2]. Par conséquent, tout algorithme efficace, est uniquement heuristique et rend des partitions sous optimales dans la plupart de cas. Il existe néanmoins plusieurs algorithmes capables de trouver de bonnes approximations de ce critère en temps raisonnable.

   - Les **algorithmes gloutons** (greedy algorithms). Le premier algorithme conçu pour maximiser la modularité de Newman-Girvan est un algorithme glouton proposé par Newman [2004b]. Il peut être qualifié aussi comme étant un algorithme d'agglomération ou une méthode de classification ascendante hiérarchique. L'algorithme commence avec autant de classes que de sommets et zéro arêtes. Les sommets sont successivement regroupés pour former des communautés si cela augmente la modularité. Les arêtes sont ajoutées une par une au cours de la procédure. Des améliorations à cet algorithme ont été proposées dans Clauset et al. [2004] avec l'utilisation d'arbres binaires. Malheureusement, cette méthode a tendance à former rapidement de grandes communautés au détriment des petites, même pour des graphes qui ne possèdent pas une structure communautaire significative, ce qui donne souvent des valeurs plus basses de la modularité. Cet inconvénient ralentit considérablement l'algorithme et le rend inapplicable pour des réseaux de plus d'un million de sommets. Dans Wakita and Tsurumi [2007], les auteurs ont proposé quelques modifications pour surmonter ce problème. Ils permettent ainsi à l'algorithme de Clauset et al. [2004] de traiter des réseaux de quelques millions de sommets. Le très célèbre **algorithme de Louvain** rentre dans la catégorie des algorithmes gloutons. Nous le décrirons en détail dans la section suivante puisqu'il s'agit de l'algorithme que nous allons utiliser pour tester les critères listés aux chapitres précédents.

   - **Recuit simulé (Simulated annealing) :** Il s'agit d'une méthode probabiliste inspirée d'un processus utilisé en métallurgie et utilisée dans différents domaines pour l'optimisation globale d'une fonction. Cette méthode, proposée par Kirkpatrick et al. [1983], a été adaptée à l'optimisation de la modularité par [Guimera et al., 2004], et son implémentation a été réalisée par Guimera and Amaral [2005]. Cette méthode retourne des solutions très proches de la solution optimale. Cependant, elle est lente. D'autre part, sa complexité réelle ne peut être estimée, car elle dépend fortement des paramètres initiaux choisis pour l'optimisation et non uniquement de la taille du graphe.

   - On trouvera dans Fortunato [2010], d'autres méthodes et algorithmes d'optimisation de la modularité de Newman-Girvan, comme la méthode d'Optimisation

2. Ce résultat, dû à la résolution du problème de Programmation Linéaire en variables bivalentes sujet aux contraintes linéaires d'une partition, a été démontré par Grötschel and Wakabayashi [1989] et Wakabayashi [1998] dans le cas des ordres ("Acyclic Subgraph Problem"), mais c'est pratiquement la même difficulté dans le cas du "Clustering non Supervisé". Plus tard, Brandes et al. [2006] ont démontré le même résultat dans le cas où la fonction qualité à optimiser est la modularité de Newman-Girvan.

extrême (Extremal optimization) de Boettcher and Percus [2001] adaptée à l'optimisation de la modularité par Duch and Arenas [2005]. Le lecteur intéressé peut voir aussi Aynaud et al. [2010].

4. Les **algorithmes dynamiques :** Il s'agit des méthodes qui se basent sur des processus qui parcourent le graphe, les modèles spin (voir Wu [1982] et Reichardt and Bornholdt [2004]), la méthode de **Marche aléatoire**[3] (Random walk) de Hughes [1995] peut également être utilisée pour la détection de communautés. Le principe est le suivant : "Si un graphe possède une structure communautaire bien définie, un marcheur aléatoire consacre beaucoup de temps à l'intérieur d'une communauté, en raison de la forte densité d'arêtes intra-classe et du nombre conséquent de chemins qu'il pourrait prendre". La plupart des algorithmes se reposant sur les marches aléatoires se basent sur une distance entre les paires de sommets du graphe. Pour Zhou [2003], la distance d_{ij} entre les sommets i et j est le nombre moyen d'arêtes qu'un marcheur aléatoire doit traverser pour atteindre j à partir de i. Ensuite, les sommets les plus proches sont susceptibles d'appartenir à la même communauté. Une autre distance est celle définie par Pons and Latapy [2006]. Cette distance est calculée à partir de la probabilité que le marcheur se déplace au hasard d'un sommet à un autre en un nombre fixe d'étapes. Les sommets sont ensuite regroupés dans des communautés à travers d'une méthode ascendante hiérarchique basée sur la méthode de Ward (Ward [1963]). Pour finir, la modularité de Newman-Girvan permet de sélectionner la meilleure partition du dendrogramme obtenu. Des applications pratiques ont montré qu'avec la méthode de random walk on peut trouver des partitions significatives et de plus, la méthode peut être appliquée à de grands graphes.

7.4 L'algorithme de Louvain

L'algorithme de Louvain est une heuristique de détection de communautés qui cherche à maximiser la modularité de Newman-Girvan (voir Blondel et al. [2008]). À l'heure actuelle il est reconnu comme étant l'un des algorithmes les plus performants que ce soit en terme de qualité, mais aussi en terme de temps de calcul. En effet, c'est l'un de seuls algorithmes à ce jour capable de traiter efficacement de très grands graphes. De plus, cet algorithme fournit une structure communautaire hiérarchique du graphe. Le code de l'algorithme est disponible sur le site internet `http://sourceforge.net/projects/louvain/`

L'algorithme est divisé en deux étapes qui sont répétées de manière itérative :

1. La première étape commence par défaut avec la partition triviale (chaque sommet est dans sa propre communauté). Ensuite, pour chaque sommet i on considère chacun de ses voisins et pour chaque voisin j, on évalue la contribution à la valeur de la modularité qui a lieu suite à la fusion de i et j. Si pour un des voisins la contribution (ou gain) est positive le sommet i sera placé dans la même classe du voisin pour laquelle le gain est maximal. Bien évidemment si pour tous les voisins de i le gain est négatif, la fusion n'aura pas lieu. Ce procédé est appliqué à plusieurs reprises et

3. C'est la méthode utilisée par le moteur de recherche Google pour parcourir, identifier et classer les pages du réseau internet.

de façon séquentielle pour tous les sommets jusqu'à ce qu'aucune amélioration ne puisse être réalisée. C'est la fin de la première étape.

Insistons sur le fait qu'un sommet peut être, et est souvent, considéré à plusieurs reprises. Cette première phase s'arrête lorsque aucune fusion ne peut améliorer la modularité (alors un maximum local de la modularité est atteint). Il faut également noter que la sortie de l'algorithme dépend de l'ordre dans lequel les sommets sont considérés. Cependant des résultats préliminaires sur plusieurs tests semblent indiquer que l'ordre des sommets n'a pas une influence significative sur la modularité obtenue. Néanmoins, l'ordre peut influencer le temps de calcul.

Une bonne partie de l'efficacité de l'algorithme provient du calcul rapide de la contribution à la modularité suite au déplacement d'un sommet. Ce résultat est une conséquence de la propriété de linéarité du critère de Newman-Girvan. Comme nous l'avons vu au chapitre précédent, la contribution suite à une fusion pour un critère linéaire peut se calculer facilement à l'aide de la formule (6.13).

2. La deuxième étape de l'algorithme consiste à construire un nouveau graphe dont les sommets sont maintenant les communautés trouvées lors de la première étape, donc le nouveau graphe contient des méta-sommets. Pour ce faire, les poids des liens entre deux nouveaux sommets sont donnés par la somme des poids des arêtes existant entre les sommets des deux communautés correspondantes. Les liens intra-classe de l'étape précédente deviennent des boucles dans le nouveau graphe. Une fois cette deuxième étape terminée, il est alors possible de ré appliquer la première étape de l'algorithme sur le réseau pondéré obtenu.

Si nous nommons "niveau" chaque combinaison de ces deux phases, par construction, le nombre de méta-communautés diminue à chaque niveau, et par conséquent, la majeure partie des calculs sont effectués dans le premier niveau. Les étapes sont réalisées (voir 7.1) jusqu'à ce qu'il n'y ait plus de changement et un maximum de modularité est atteint (voir la figure 7.1 pour un exemple).

L'algorithme intègre une notion de hiérarchie : des communautés de communautés sont construites au cours de chaque niveau, ce qui peut se traduire par un dendrogramme. La hauteur du dendrogramme construit est déterminée par le nombre de niveaux. Ainsi, l'algorithme de Louvain peut être qualifié comme étant une méthode ascendante hiérarchique.

L'algorithme de Louvain présente plusieurs avantages. Non seulement ses étapes sont intuitives et faciles à mettre en œuvre mais en plus il est très rapide. Des simulations effectuées sur des grands graphes montrent dans la pratique que sa complexité est linéaire. Cela est dû au calcul rapide de la contribution et au fait que le nombre de communautés diminue considérablement après quelques niveaux de sorte que la plupart du temps d'exécution est concentré sur les premiers niveaux. Par ailleurs, la méthode fournit un moyen de contourner la limite de résolution grâce à sa nature intrinsèquement multiniveau. En effet, les solutions intermédiaires trouvées par l'algorithme peuvent également être significatives et la structure hiérarchique rendue par celui-ci permet de zoomer dans le réseau et d'observer sa structure avec la résolution désirée.

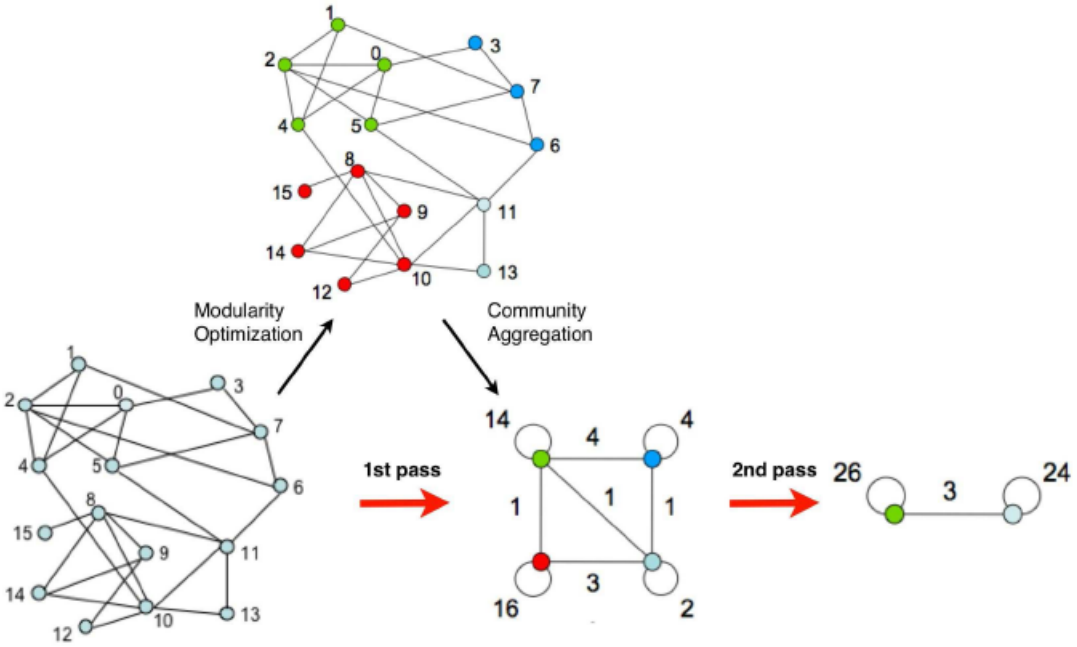

Figure 7.1 – Les deux étapes de l'algorithme de Louvain (l'image a été prise de Blondel et al. [2008]).

7.4.1 L'algorithme de Louvain générique

L'algorithme de Louvain générique est une adaptation de l'algorithme de Louvain à quelques-uns des critères de modularisation décrits au chapitre 5. Cette adaptation a été réalisée au sein de l'équipe "Complex Networks" du LIP6 (Laboratoire d'Informatique de Paris 6) par R. Campigotto et J.L. Guillaume (voir Campigotto et al. [2013]). Le principe est le même, les étapes de l'algorithme décrites précédemment restent les mêmes. C'est le calcul de la contribution et bien évidemment le calcul de la valeur de chaque critère qui changent.

Louvain générique permet d'estimer la partition optimale de dix critères :

1. Le critère de Newman-Girvan.
2. Le critère de Zahn-Condorcet.
3. Le critère d'Owsiński-Zadrożny.
4. Le critère de Condorcet pondéré en $\mathbf{A}$.
5. L'Écart à l'Indétermination.
6. L'Écart à l'Uniformité.
7. La Modularité Équilibrée.
8. La Différence de Profils.
9. La densité de Michalski-Goldberg.
10. Le critère de Shi-Malik.

Ces critères ne s'intègrent pas tous à Louvain de la même façon :

- Certains critères comme la Différence de Profils et le critère de Condorcet pondéré en $\mathbf{A}$ ont besoin d'une étape de pré traitement des données d'entrée. Bien que ces deux critères soient séparables, la fonction ϕ dépend de la matrice $\hat{\mathbf{A}}$ pour ces deux critères.

- Certains critères nécessitent des paramètres d'entrée comme le critère d'Owsiński-Zadrożny et le critère de Shi-Malik, pour lequel on doit fixer le nombre de classes minimal. Si l'on ne fixe pas le nombre de classes à l'avance, le critère de Shi-Malik rend la partition grossière (voir chapitre 6).

- Compte tenu de la première étape de l'algorithme, la fusion de deux sommets a lieu uniquement s'ils sont adjacents. L'algorithme évite ainsi l'inconvénient du critère d'Écart à l'Indétermination, qui peut rendre une partition avec des classes à composantes non connexes (voir chapitre 6).

- L'algorithme de Louvain générique permet de traiter de graphes pondérés avec tous les critères, sauf la Différence de Profils et pour le critère de Condorcet pondéré en $\mathbf{A}$. Pour un graphe pondéré si le terme général de la matrice de poids $\mathbf{W}$ vaut $w_{ii'}$ le terme général de la relation inverse $\bar{\mathbf{W}}$ vaut $\bar{w}_{ii'} = \max_{(i,i')} w_{ii'} - w_{ii'}$

- Le calcul du gain G ou contribution se fait pour tous les critères linéaires à partir de l'expression (6.13). Par exemple, pour le critère d'Owsiński-Zadrożny on trouvera dans le code de l'algorithme de Louvain générique l'expression suivante du calcul du gain suite à la fusion du sommet [4] u à la communauté de son voisin [5] c :

$$G_{OZ}(\alpha) = \mathrm{dnc} - \alpha \mathrm{w}_u \mathrm{w}_c \mathrm{max} \tag{7.5}$$

Par exemple, pour la Modularité équilibrée, on a

$$G_{BM} = \left(2\mathrm{dnc} - \frac{\mathrm{degctotc}}{2M} - \mathrm{w}_u \mathrm{w}_c \mathrm{max} + \frac{(N\mathrm{w}_u \mathrm{max} - \mathrm{degc})(N\mathrm{w}_c \mathrm{max} - \mathrm{totc})}{(N^2 \mathrm{max} - 2M)} \right), \tag{7.6}$$

où :
- dnc est le poids des arêtes reliant u à c ;
- w_u est le nombre de sommets contenus dans le méta sommet u ;
- w_c est le nombre de sommets contenus dans le méta sommet c ;
- $\mathrm{max} = \max_{(i,i')} w_{ii'}$ est le poids maximal de la matrice de poids ;s
- degc est la somme des poids des arêtes incidentes aux sommets se trouvant dans le méta sommet ou communauté u ; soit la somme de degrés des sommets dans u pour un graphe non pondéré.

4. Après la première étape de l'algorithme, ce sommet peut être un meta sommet (i.e. un groupe de sommets).

5. Après la première étape de l'algorithme, ce voisin peut être un meta sommet (i.e. un groupe de sommets).

– totc est la somme des poids des arêtes incidentes aux sommets se trouvant dans le méta sommet ou communauté c, ou la somme de degrés des sommets dans c pour un graphe non pondéré.

7.5 Exemples d'application

Dans cette section nous allons présenter les résultats obtenus suite à la modularisation des graphes réels suivants :

- Le réseau social dit "club de karaté de Zachary" de Zachary [1977]. Il s'agit d'un jeu de données fréquemment utilisé en analyse de réseaux sociaux. Le graphe de Zachary est un réseau réel composé de 34 membres d'un club de karaté d'une université américaine.

- Le réseau nommé "American College Football" de Girvan and Newman [2002]. Ce réseau représente le calendrier des matchs de football américain pour la saison 2000. Chaque sommet du graphe représente une équipe et les arêtes représentent les matchs entre les équipes. Ce réseau est intéressant car il intègre une structure communautaire connue.

- Le réseau de musiciens de "jazz" de Gleiser and Danon [2003] (introduit au chapitre précédent). Il s'agit d'un graphe de collaboration entre musiciens de jazz qui jouaient entre 1912 et 1940. Chaque sommet correspond à un groupe ou orchestre de jazz. Deux sommets sont connectés s'ils ont un musicien en commun. Les données ont été obtenues à partir de la base de données The Red Hot Jazz Archive digital.

- Le réseau "internet" : il s'agit d'un sous graphe d'internet de Hoerdt and Magoni [2003] (introduit au chapitre précédent).

- Le réseau "Amazon" recueilli sur le site internet Amazon.com. Il se base sur le principe suivant : "les clients ayant acheté un article X du site Amazon ont également acheté un article Y" (voir Yang and Leskovec [2012]). Chaque sommet représente un produit acheté sur le site Amazon. Si un produit i est fréquemment co-acheté avec un produit j, le graphe contient une arête entre i et j. Les données ont été obtenues sur le site `http://snap.stanford.edu/data/com-Amazon.html`. [6].

- Le réseau social "Youtube" (site de partage de vidéos) où chaque sommet est un utilisateur. Certains utilisateurs créent des groupes que d'autres utilisateurs peuvent rejoindre. Il existe un lien entres deux utilisateurs s'ils ont rejoint un même groupe. Les données sont fournies par Mislove et al. [2007].

6. Dans Yang and Leskovec [2012], les auteurs ont défini des communautés réelles en se basant sur la catégorie du produit fourni par Amazon qu'ils ont appelées "ground-truth communities" (vraies communautés de terrain). Le nombre de communautés réelles ou ground-truth est fourni sur ce site. Cependant, la description de chaque communauté n'est pas fournie, car la catégorie de produit n'est pas renseignée. D'autre part, ces communautés sont chevauchantes, i.e. elles permettent la multi-appartenance.

Le tableau 7.1 et la figure 7.2 comparent le nombre de classes obtenus suite à l'application de l'algorithme de Louvain pour chaque critère de modularisation.

	Karaté	**Fooball**	**Jazz**	**Internet**	**Amazon**	**Youtube**
N	34	115	198	69 949	334 863	1 134 890
M	78	613	2 742	351 380	925 872	2 987 624
Critère	κ	κ	κ	κ	κ	κ
Zahn-Condorcet	19	16	36	40 123	161 439	878 849
Ecart à l'Uniformité	6	10	20	173	265	51 584
Newman-Girvan	4	10	4	46	250	5 567
Ecart à l'Indétermination	4	10	6	39	246	13 985
Modularité Equilibrée	4	10	5	41	230	6 410
Différence de Profils	4	9	3	3 373	23 118	66 276
Michalski-Goldberg	11	27	39	21 260	109 974	248 621

TABLE 7.1 – Nombre de classes trouvées par les différents critères avec l'algorithme de Louvain générique.

Le tableau 7.1 et la figure 7.2 montrent que pour tous les jeux de données les critères de Zahn-Condorcet et Michalski-Goldberg génèrent plus de classes et cette différence s'accentue au fur et à mesure que la taille du graphe augmente. Comme attendu, les critères qui possèdent une limite de résolution : Newman-Girvan, l'Écart à l'Uniformité, l'Écart à l'Indétermination et la Modularité Équilibrée génèrent moins de classes lorsque N et M augmentent car ils ont du mal à identifier les groupes densément connectés. Quant au nombre de classes rendu par le critère d'Écart à l'Uniformité celui-ci dépend de la densité d'arêtes du graphe δ, qui est décroissant en N :

	Karaté	**Fooball**	**Jazz**	**Internet**	**Amazon**	**Youtube**
Densité δ	0,13	0,09	0,14	$1{,}44 \cdot 10^{-4}$	$1{,}65 \cdot 10^{-5}$	$4{,}64 \cdot 10^{-6}$

TABLE 7.2 – Densité d'arêtes des graphes étudiés.

Plus grand est δ plus exigeant est ce critère quant au pourcentage minimal d'arêtes intra-classe. Pour un graphe réel avec densité d'arêtes $\delta < \frac{1}{2}$ nous aurons toujours $\kappa_{ZC} >> \kappa_{\text{UNIF}}$, i.e. le nombre de classes obtenu via l'optimisation du critère d'Écart à l'Uniformité sera toujours inférieur à celui obtenu avec le critère de Zahn-Condorcet.

L'annexe C présente une analyse plus détaillée pour les réseaux "Karaté" et "Football". Nous nous intéressons à ces petits réseaux car ils possèdent une structure communautaire prédéfinie qui permet d'interpréter les partitions trouvées par chaque critère. Nous utilisons le critère de Rand pour comparer deux partitions (voir Rand [1971], Marcotorchino [1984b], Saporta [1988], Saporta and Youness [2002]). L'annexe C présente également pour le réseau "Jazz" une analyse des partitions trouvées avec les critères de Newman-Girvan, Écart à l'Indétermination et Modularité Équilibrée.

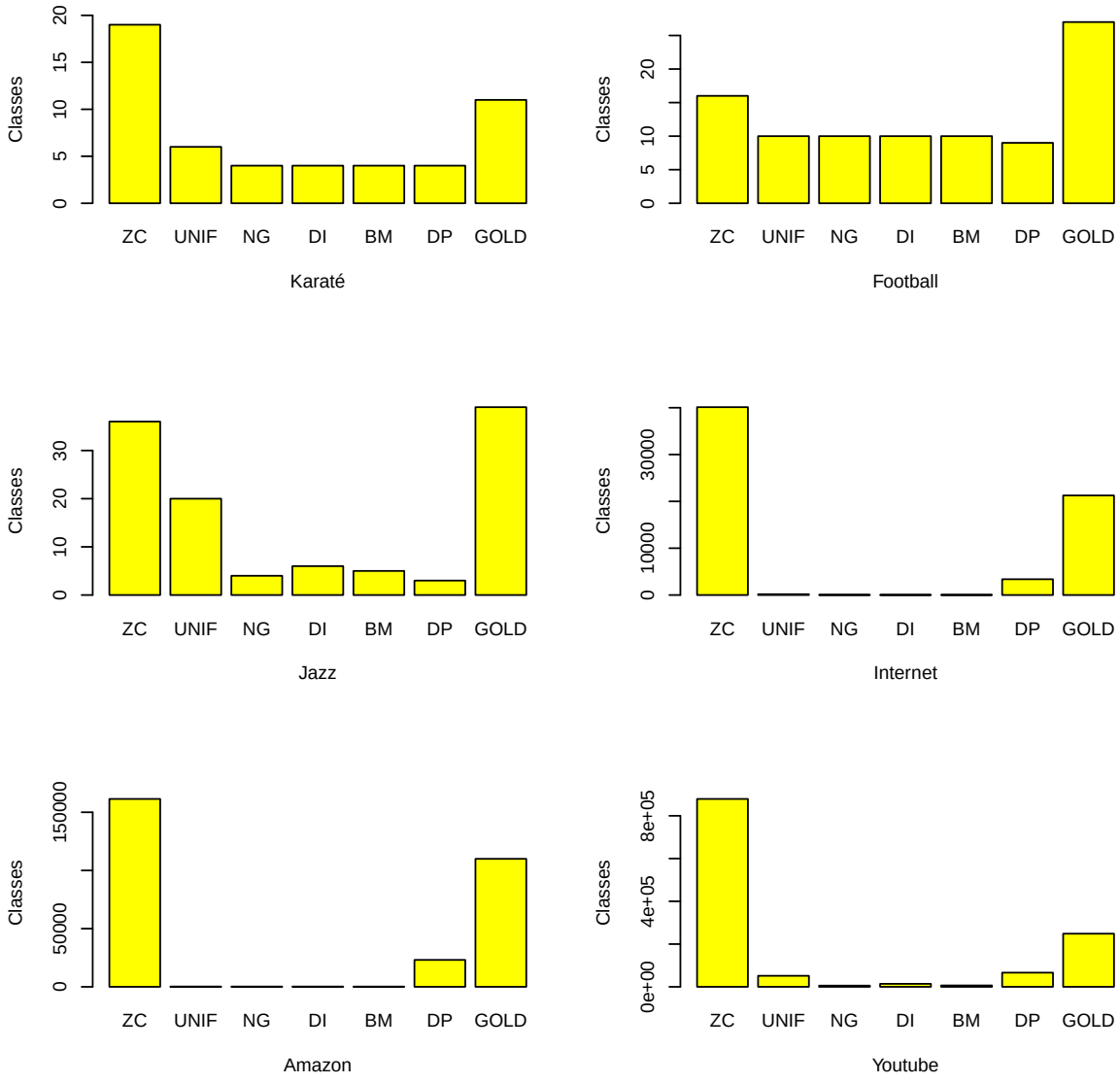

Figure 7.2 – Nombre de classes selon obtenus pour les critères : Zahn-Condorcet (ZC), Ecart à l'Uniformité (UNIF), Newman-Girvan (NG), Ecart à l'Indétermination (DI), Modularité Equilibrée (BM), Différence de Profils (DP) et Michalski-Goldberg (GOLD).

Le tableau 7.3 montre le nombre de communautés obtenu via l'optimisation des critères linéaires (Zahn-Condorcet, Écart à l'Uniformité, Newman-Girvan, Ecart à l'Indétermination et Modularité équilibrée) pour des graphes réels plus grands que ceux du tableau 7.1. Les graphes utilisés sont *Web nd.edu* (voir Albert et al. [1999]), *WebUk05* et *WebBase01* (voir Boldi et al. [2004] et Boldi and Vigna [2004]).

	Web nd.edu	**WebUk05**	**WebBase01**
N	325k	39M	118M
M	1M	783M	1B
Densité	$2{,}77 \cdot 10^{-05}$	$1{,}20 \cdot 10^{-06}$	$1{,}46 \cdot 10^{-07}$
Critère	κ	κ	κ
Zahn-Condorcet	201 647	21 738 667	71 806 729
Ecart à l'Uniformité	711	69 347	2 777 580
Newman-Girvan	511	19 819	2 759 248
Ecart à l'Indétermination	324	78 213	2 770 647
Modularité Equilibrée	333	20 788	2 736 808

TABLE 7.3 – Nombre de classes trouvées par les différents critères avec l'algorithme de Louvain générique.

Le tableau 7.3 montre, encore une fois, que pour tous les jeux de données le critère de Zahn-Condorcet génère plus de classes que les autres quatre critères, lesquels possèdent une limite de résolution.

Chapitre 8

Conclusion générale et perspectives

Dans cette thèse nous avons mené un travail de recherche concernant la problématique de la recherche des communautés ou partitionnement des grands graphes et grands réseaux. Nous avons mis l'accent sur l'importance de définir une fonction qualité à optimiser, que nous avons appelé aussi critère de modularisation, qui puisse juger la qualité de telles partitions. Nous avons rencontré dans la littérature différents critères de modularisation proposés pour faire face à des problématiques issues des différents domaines et différents contextes. L'optimisation de chaque critère fournissant une partition différente du même graphe nous avons vu le besoin de réécrire ces fonctions qualités avec les mêmes notations de base dans le but de pouvoir comparer et comprendre ces différences. C'est l'écriture relationnelle qui nous a permis d'accomplir cette tâche, d'autant plus qu'un graphe représente une relation binaire particulière. Dans ce cadre les travaux effectués et résultats obtenus sont listés ci-après :

- Dans le but de classifier les critères selon les propriétés qu'ils vérifient nous avons énoncé trois propriétés importantes que doit vérifier un *bon critère* : linéarité, séparabilité et équilibre. Nous avons défini et étendu la définition de la propriété d'équilibre pour les critères linéaires. Nous avons étudié les conséquences de la vérification ou non vérification de cette propriété. Ainsi nous avons défini différents niveaux d'équilibre linéaire :

  1. La propriété d'équilibre général : dont la violation a pour conséquence que la partition optimisant la valeur du critère soit soit la partition triviale (tous les objets sont isolés), soit la partition grossière (tous les objets sont dans la même classe) ; obligeant ainsi à l'utilisateur à fixer le nombre de classes à l'avance.
  2. La propriété d'équilibre général local : la formulation du critère se base sur une condition locale (pour chaque paire de sommets).
  3. La propriété d'équilibre général global : la formulation du critère se base sur une condition globale (pour l'ensemble du graphe).

  Tout critère équilibré localement est aussi équilibré globalement. Un sous-ensemble des critères vérifiant la propriété d'équilibre général global est l'ensemble de critères qui reposent sur la définition de **modèle nul**. Tout modèle nul possède une **limite de résolution**, une caractéristique qu'ont certains critères de modularisation qui

fait que son optimisation ne permette pas la détection de communautés en dessous d'une certaine échelle qui dépend de caractéristiques globales du réseau. La limite de résolution a pour conséquence que le critère ne soit pas **invariant d'échelle**.

- Nous avons présenté l'écriture relationnelle des différents critères de modularisation existant dans la littérature : Zahn-Condorcet, Owsiński-Zadrożny, Condorcet pondéré, Newman-Girvan, Mancoridis-Gansner, la Différence de profils, Michalski-Goldberg, etc. Cette écriture nous a permis d'étudier leurs propriétés. Nous avons introduit ou adapté à la modularisation de graphes trois critères : la Modularité Équilibrée, l'Écart à l'Indétermination et l'Écart à l'Uniformité.

- Grâce à la dualité indépendance-indétermination en statistiques de contingences et en faisant le lien avec la théorie de transport optimal nous avons donné une interprétation intéressante aux graphes suivant la structure d'indétermination.

- L'étude de la dualité indépendance-indétermination nous a permis aussi de définir une représentation des critères de modularisation dans un environnement *semi-contingentiel*, i.e. un environnement qui croise une relation binaire symétrique avec une relation d'équivalence. La première représentée par l'espace des arêtes du graphe (supposé non pondéré, non orienté et non réflexif) présentes dans la matrice d'adjacence et la deuxième représentée par la partition du graphe en classes d'équivalence que nous cherchons à obtenir.

- Nous avons caractérisé les partitions trouvées via l'optimisation de six critères linéaires : Zahn-Condorcet, Owsiński-Zadrożny, Newman-Girvan, l'Écart à l'Uniformité, l'Écart à l'Indétermination et la Modularité Équilibrée. Nous avons basé notre étude sur l'impact sur la valeur de chaque critère suite à la fusion de deux sous-graphes. Cela nous a permis de comprendre les différences trouvées quant au nombre optimal de classes fourni par chaque critère. Nous avons vu aussi que certains critères se basent sur la densité d'arêtes intra-classe (critères de Zahn-Condorcet, Owsiński-Zadrożny et l'Écart à l'Uniformité) tandis que d'autres dépendent de la distribution de degrés (Newman-Girvan, l'Écart à l'Indétermination et la Modularité Équilibrée).

- Grâce à l'algorithme de Louvain générique nous avons modularisé des graphes de tailles différentes avec les différents critères étudiés. La comparaison des partitions trouvées nous a permis de valider les résultats trouvés de façon théorique.

Nous terminons en présentant une liste de possibles travaux futurs qui donneront suite aux travaux menés :

- Bien évidemment la liste de critères présentés dans cette thèse n'est pas exhaustive et d'autres critères peuvent être étudiés comme par exemple, les critères se basant sur d'autres propriétés du graphe comme la distance entre deux sommets ou le coefficient de classification. Nous pouvons citer par exemple la "Cut Point Additive Distance" (CPAD) introduite par Chebotarev [2013]. Nous pouvons citer aussi les critères proposés ou adaptés à la recherche de communautés dans les graphes de Ah-Pine [2013].

- L'étude de l'espace semi-contingentiel que nous avons introduit peut être étendue, généralisée et développée afin d'obtenir une liste complète de formules de transfert

entre la notation relationnelle et la notation semi-contingentielle. De même, l'environnement semi-contingentiel peut être généralisé au croisement de deux relations de nature différente.

- À partir de l'étude de la théorie spectrale comme outil de modularisation que nous avons menée, il est possible de déduire un algorithme de classification. Cet algorithme se baserait sur la recherche de vecteurs du tableau disjonctif complet de la partition cherchée qui s'approche au mieux des vecteurs propres associés aux plus grandes valeurs propres de la matrice de données. De plus, ce principe peut se généraliser pour tous les critères linéaires dont la matrice de données est symétrique.

- Une étude statistique peut être réalisée sur la distribution des degrés qui, selon la littérature, correspond à une loi de puissance pour les graphes réels. Nous pouvons exploiter ces résultats pour inférer sur la partition optimale des critères dépendant de la distribution des degrés.

- L'algorithme de Louvain générique a pour principal atout son efficacité dans le traitement des grands graphes. Cet algorithme pourrait aussi s'étendre au calcul des indicateurs qui mesurent la qualité de partitions trouvées.

- Les jeux de données de grands graphes réels disponibles sur internet, comme facebook ou twitter, contiennent des communautés connues chevauchantes. Une étude peut être faite sur ces données pour définir les partitions souhaitées ou réelles à partir de ces communautés chevauchantes qui serviront à juger le bon comportement de chaque critère.

ANNEXES

Annexe A

Les formules de transfert

Formules de transfert des espaces Contingentiels aux espaces Relationnels

Etant données deux variables catégorielles X et Y à p et q modalités respectivement décrivant N individus ; les formules de passage des espaces Contingentiels aux espaces Relationnels sont données dans le tableau suivant :

Ecriture contingentielle $\leftrightarrow$ Ecriture relationnelle
$\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}y_{ij}$
$\sum_{u=1}^{p} n_{u.}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} x_{ij}$
$\sum_{v=1}^{q} n_{v.}^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij}$
$\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}n_{u.}n_{.v} = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{x_{i.}+x_{.j}}{2}\right) y_{ij}$
$\sum_{u=1}^{p}\sum_{v=1}^{q} n_{uv}n_{u.}n_{.v} = \sum_{i=1}^{N}\sum_{j=1}^{N}\left(\frac{y_{i.}+y_{.j}}{2}\right) x_{ij}$
$\sum_{u=1}^{p}\sum_{v=1}^{q} \frac{n_{uv}^2}{n_{u.}n_{.v}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \frac{x_{ij}}{x_{i.}}\frac{y_{ij}}{y_{.j}}$
$\sum_{u=1}^{p}\sum_{v=1}^{q}\left[\frac{(\frac{n_{uv}}{N})^2}{\frac{n_{u.}}{N}}\right] = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{x}_{ij}y_{ij}$
$\sum_{u=1}^{p}\sum_{v=1}^{q}\left[\frac{(\frac{n_{uv}}{N})^2}{\frac{n_{.v}}{N}}\right] = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}x_{ij}\hat{y}_{ij}$

où :

- n_{uv} est le terme général du tableau de contingence croisant X et Y dont $n_{u.} = \sum_{v=1}^{q} n_{uv}$ et $n_{.v} = \sum_{u=1}^{p} n_{uv}$ sont les distributions marginales.

- x_{ij} et y_{ij} sont les termes généraux des tableaux relationnels de comparaison par paires des variables X et Y respectivement dont $x_{i.} = \sum_{i=1}^{N} x_{ij}$ et $y_{.j} = \sum_{j=1}^{N} y_{ij}$ sont les distributions marginales.
- Où $\hat{x}_{ij} = \frac{x_{ij}}{x_{i.}} = \frac{x_{ij}}{x_{.j}}$ et $\hat{y}_{ij} = \frac{y_{ij}}{y_{i.}} = \frac{y_{ij}}{y_{.j}}$.

Annexe B

L'Impact de la fusion de deux classes

B.1 L'Impact de la fusion de deux classes pour les critères linéaires

Nous présentons ici le détail des résultats présentés au tableau 6.4. Les résultats ont été obtenus à partir du tableau 6.3.

B.1.1 L'Impact de la fusion de deux classes sur le critère de Zahn-Condorcet

Le tableau 6.3 montre que pour fusionner les deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$, i.e. $C_{ZC} > 0$, le nombre total d'arêtes inter-classes doit être au moins égal à la moitié du nombre maximal possible d'arêtes qui pourrait exister, soit $l > \frac{n_1 n_2}{2}$. Ce résultat, nullement surprenant, provient de la formulation d'origine de ce critère : "*La règle de la majorité absolue de Condorcet*" en théorie des votes (voir théorème 5.1). Cela signifie que la fusion des deux classes contribuera à l'optimisation du critère si et seulement si chaque sommet de $\mathcal{C}_1$ est connecté au moins à la moitié des sommets de $\mathcal{C}_2$ et vice-versa. Ce résultat a les conséquences suivantes :

- La densité d'arêtes intra-classe est supérieure à 0.5 pour toutes les classes identifiées par ce critère.

- les communautés obtenues via l'optimisation du critère de Zahn-Condorcet possèdent un diamètre de 2, car chaque sommet étant connecté à plus de la moitié des sommets dans la communauté, les sommets de toute paire de sommets ont au moins un voisin commun.

- Ce critère ne possède pas de limite de résolution (comme c'est le cas des critères se basant sur un modèle nul, voir Fortunato and Barthelemy [2006]), car la contribution dépend seulement des caractéristiques locales des classes à fusionner : l, n_1 et n_2. Les caractéristiques globales du graphe, comme la taille du graphe ou le nombre total d'arêtes, n'interviennent pas dans le calcul de la contribution. Ce qui confère à ce critère la propriété d'*invariance d'échelle*. Ces résultats sont cohérents avec ce

que nous avons énoncés au chapitre 5 : ce critère possède la propriété d'équilibre général local. Donc, selon la définition de cette propriété (voir chapitre 4, un critère ne peut pas être équilibré localement et être un modèle nul au même temps et c'est précisément les critères se basant sur un modèle nul qui possèdent une limite de résolution.

- Un inconvénient de ce critère, que nous constatons avec des exemples pratiques au chapitre 7, est le fait que la solution optimale contienne beaucoup de petites classes, soient des cliques à 2 ou 3 sommets soient des classes à un seul sommet isolé. Autrement dit, des sous-graphes ayant un coefficient de clustering supérieur à 50%, voire égal à l'unité. Ceci est dû au respect de la règle de majorité absolue que ce critère impose au nombre de connexions minimal que chaque sommet doit avoir vis-à-vis de ses voisins dans sa communauté.

B.1.2 L'Impact de la fusion de deux classes sur le critère d'Owsiński-Zadrożny

Le tableau 6.3 montre que pour fusionner les classes $\mathcal{C}_1$ et $\mathcal{C}_2$, i.e. $C_{OZ} > 0$, le nombre total d'arêtes inter-classes doit être au moins égal à α pourcent du nombre maximal possible d'arêtes qui pourrait exister, soit $l > \alpha n_1 n_2$ (voir théorème 5.2). Donc, pour que la contribution soit positive, chaque sommet de $\mathcal{C}_1$ doit être connecté au moins à α pourcent des sommets de $\mathcal{C}_2$ et vice-versa. Pour $\alpha = 0.5$ nous retrouvons le critère de Zahn-Condorcet. En effet, le coefficient α définit le balance entre les fonctions ϕ et $\bar{\phi}$. Ce résultat a les conséquences suivantes :

- La densité d'arêtes de intra-classe est supérieure ou égale à α pour toutes les classes identifiées par ce critère.

- Ce critère ne possède pas de limite de résolution (comme le critère de Zahn-Condorcet), car la contribution dépend seulement des caractéristiques locales des classes à fusionner et de α qui est défini par l'utilisateur. Ce qui confère à ce critère la propriété d'invariance d'échelle.

- Pour $\alpha < 0.5$ l'optimisation de ce critère rendra moins de petites cliques que le critère de Zahn-Condorcet.

- Un inconvénient de ce critère c'est le choix de la valeur α qui revient indirectement à fixer le nombre de classes.

B.1.3 Impact de la fusion de deux classes sur le critère d'Écart à l'Uniformité

Selon le tableau 6.3 le critère d'Écart à l'Uniformité requiert que la fraction d'arêtes existant entre $\mathcal{C}_1$ et $\mathcal{C}_2$ soit au moins égale à la densité d'arêtes globale du graphe. On vérifie que le critère d'Écart à l'Uniformité est un cas particulier du critère d'Owsiński-Zadrożny avec $\alpha = \delta = \frac{2M}{N^2}$. Ce résultat implique les résultats suivants :

- Ce critère possède une **limite de résolution** car la décision de fusionner les deux classes dépend de la densité globale du graphe qui dépend à son tour de la taille du graphe N et le nombre total d'arêtes M. En effet, il s'agit d'un modèle nul (voir chapitre 4).

- Les classes obtenues via l'optimisation du critère d'Écart à l'Uniformité possèdent une densité d'arêtes intra-classe supérieure ou au moins égale à la densité d'arêtes du graphe δ.

B.1.4 Impact de la fusion de deux classes sur le critère de Newman-Girvan

Le tableau 6.3 montre que la contribution au critère de Newman-Girvan dépend de la quantité $\left(n_1 n_2 \frac{d^1_{av} d^2_{av}}{2M}\right)$, qui est fonction du degré moyen des deux classes, et du nombre total d'arêtes M.

Le fait que la contribution dépende du nombre total d'arêtes M confirme un des inconvénients majeurs de ce critère : sa **limite de résolution** (un résultat très connu et énoncé pour la première fois par Fortunato and Barthelemy [2006]). C'est une conséquence du fait que la définition du critère répose sur la définition de **modèle nul** (voir chapitre 4).
En effet, supposons que la contribution est négative et que nous rajoutons quelques sommets au graphe, et par conséquet quelques arêtes pour que le graphe reste connexe (donc N et M augmentent). Supposons aussi que les nouveaux sommets ne sont pas connectés ni aux sommets de $\mathcal{C}_1$ ni aux sommets de $\mathcal{C}_2$. La quantité $n_1 n_2 \frac{d^1_{av} d^2_{av}}{2M}$ devient plus petite, car n_1, n_2, d^1_{av} et d^2_{av} n'ont pas changé alors que M a augmenté. Donc, la contribution augmente, puisque l ne change pas, et à partir d'une certaine valeur de M elle devient positive et par conséquent la fusion des deux classes favorise l'optimisation du critère, alors que les caractéristiques des deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$ n'ont pas changé. Donc, plus M augmente, plus facile devient la fusion des deux classes.

Du calcul de la contribution nous pouvons déduire aussi que la partition obtenue suite à l'optimisation du critère de Newman-Girvan ne possède pas de classes à un seul sommet de degré 1. En effet, supposons que $\mathcal{C}_1$ est une classe d' un seul sommet de degré 1 connecté par sa seule arête à $\mathcal{C}_2$, nous avons $n_1 = 1$, $d^1_{av} = 1$, $l = 1$ et par conséquent la contribution est toujours positive : $C_{NG} = \left(1 - n_2 \frac{d^2_{av}}{2M}\right) = \left(1 - \frac{\sum_{i=1}^{n_2} a_{i.}}{2M}\right) > 0$ car $\sum_{i=1}^{n_2} a_{i.} > 2M$.

Ce résultat a déjà été retrouvé et démontré dans Brandes et al. [2008] où les auteurs ont montré qu'étant donné un graphe connexe $G = (V, E)$ un partitionnement de celui-ci avec modularité maximale ne peut pas contenir une classe d'un seul sommet de degré 1 : *"A clustering with maximum modularity has no cluster that consists of a single node with degree 1"*.

B.1.5 Impact de la fusion de deux classes sur le critère d'Écart à l'Indetermination

Ce critère possède des caractéristiques similaires au critère de Newman-Girvan. La contribution à la valeur de ce critère (voir tableau 6.3) dépend du degré moyen des classes $\mathcal{C}_1$ et $\mathcal{C}_2$, et par conséquent de la distribution des degrés.

La condition de positivité (5.19) implique que la quantité $n_1 n_2 \left(\frac{d^1_{av}}{N} + \frac{d^2_{av}}{N} - \frac{2M}{N^2} \right)$ soit positive [1]. Le non-respect de cette condition pourrait entrainer que la solution optimale contienne des classes à composantes non connexes [2]. En effet, supposons que cette condition ne soit pas tout le temps vérifiée, donc il existe deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$ telles que $\left(\frac{d^1_{av}}{N} + \frac{d^2_{av}}{N} - \frac{2M}{N^2} \right) < 0$, supposons aussi qu'il n'existe aucune arête entre $\mathcal{C}_1$ et $\mathcal{C}_2$ et donc $l = 0$, dans ce cas-là la contribution vaut : $C_{DI} = -\left(\frac{d^1_{av}}{N} + \frac{d^2_{av}}{N} - \frac{2M}{N^2} \right) > 0$ et par conséquent, la fusion augmente la valeur du critère alors que les sous-graphes $\mathcal{C}_1$ et $\mathcal{C}_2$ ne sont pas connexes et on risque d'obtenir des classes comme celle montrée dans la figure 6.8.

Ce critère possède une **limite de résolution** comme la plupart des critères se basant sur un **modèle nul** (voir chapitre 4). On peut constater cela si l'on regarde l'expression de la contribution. Celle-ci dépend de propriétés globales du réseau, à savoir le nombre total de sommets N et le nombre total d'arêtes M. En effet, si N augmente vu que la quantité $\left(\frac{d^1_{av}}{N} + \frac{d^2_{av}}{N} - \frac{2M}{N^2} \right) > 0$ est décroissante en N, même si l est petit la contribution augmentera et à partir des certaines valeurs de N et M l'optimisation du critère aura du mal à identifier même les classes clairement bien définies, comme les cliques connectées par une seule arête.

B.1.6 Impact de la fusion de deux classes sur la Modularité Équilibrée

Le calcul de la contribution à la Modularité Équilibrée suite à la fusion d $\mathcal{C}_1$ et $\mathcal{C}_2$ est calculée à partir de l'expression de ce critère donnée dans le tableau 6.2 et de l'équation (6.13) et en tenant compte que $\bar{a}_{ii'} = (1 - a_{ii'})$:

$$C_{BM} = 2l + \frac{(n_1 N - \sum_{i \in \mathcal{C}_1}^{n_1} a_{i.})(n_2 N - \sum_{i' \in \mathcal{C}_2}^{n_2} a_{.i'})}{N^2 - 2M} - n_1 n_2 - \frac{\sum_{i \in \mathcal{C}_1}^{n_1} a_{i.} \sum_{i' \in \mathcal{C}_2}^{n_2} a_{.i'}}{2M},$$

ce qui peut encore s'écrire :

$$C_{BM} = \left(2l + n_1 n_2 \frac{(N - d^1_{av})(N - d^2_{av})}{N^2 - 2M} - n_1 n_2 - n_1 n_2 \frac{d^1_{av} d^2_{av}}{2M} \right). \tag{B.1}$$

À partir de cette expression nous pouvons déduire les résultats suivants :

1. La positivité de $n_1 n_2 \left(\frac{d^1_{av}}{N} + \frac{d^2_{av}}{N} - \frac{2M}{N^2} \right)$ s'obtient en appliquant la condition de positivité (5.19) à toutes les paires de sommets composées d'un sommet dans $\mathcal{C}_1$ et d'un autre dans $\mathcal{C}_2$ et en faisant ensuite la somme sur les n_1 et n_2 sommets.

2. Cette condition n'est pas vérifiée tout les temps. Cependant l'algorithme de Louvain empêche d'obtenir des composantes non connexes (voir chapitre 7).

- La contribution dépend de la distribution des degrés car son calcul fait intervenir les degrés moyens des deux classes $\mathcal{C}_1$ et $\mathcal{C}_2$.

- Ce critère possède une limite de résolution car la contribution dépend des propriétés globales : M et N. Rien de surprenant étant donné qu'il s'agit d'un modèle nul.

B.2 Comparaison des critères non linéaires

Nous allons étudier, dans un premier temps, les comportements des critères dont la formulation générale est la suivante :

$$F(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \phi_{ii'}\psi(x_{ii'}) + \sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{\phi}_{ii'}\bar{\psi}(\bar{x}_{ii'}) + K, \tag{B.2}$$

où ψ et $\bar{\psi}$ sont des fonctions non linéaires de $\mathbf{X}$.

Nous allons mener aussi une analyse de coût de fusion de deux classes, comme pour les critères linéaires. Cela nous aidera à comprendre le nombre de classes attendu via l'optimisation de chaque critère. Cependant, comme les critères ne sont pas linéaires nous ne pouvons plus utiliser la formule (6.13) pour calculer la contribution à la valeur de chaque critère. Nous allons, donc déduire l'expression de la contribution critère par critère.

Nous finirons notre étude avec l'analyse du critère de la Différence de Profils.

Dans toute cette section nous supposons que les graphes que nous voulons modulariser sont non orientés et non pondérés.

B.2.1 Le critère de Michalski-Goldberg

Si dans l'expression (B.2) nous avons $\phi_{ii'} = a_{ii'}$, $\psi_{ii'} = \hat{x}_{ii'} = \frac{x_{ii'}}{x_{i.}}$ et $\bar{\phi}_{ii'} = 0 \quad \forall\,(i,i')$ nous retrouvons le critère de Mischalski-Goldberg (voir expression (5.60)) :

$$F_G(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}\hat{x}_{ii'}.$$

Ce critère est à maximiser et sa solution optimale n'est pas triviale si le graphe n'est pas réflexif. Cependant, si le graphe est réflexif la solution optimale de ce critère est triviale comme l'énonce le lemme suivant :

Lemme B.1. *La solution optimale obtenue via la maximisation du critère de Michalski-Goldberg est la solution triviale si le graphe est réflexif.*

Démonstration. Grâce aux propriétés de la matrice $\hat{\mathbf{X}}$ (matrice bi-stochastique) nous avons $\sum_{i=1}^{N}\sum_{i'=1}^{N} \hat{x}_{ii'} = N$. Ce qui implique que $\forall\,\mathbf{X}\, F_G(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}\hat{x}_{ii'} \leq N$. Cependant si le graphe est réflexif, i.e. $a_{ii} = 1\,\forall i$ et $\mathbf{X}^{\text{triv}}$ est la relation de la partition triviale le critère atteint sa borne supérieure : $F_G(X^{\text{triv}}) = N$
Donc, pour toute autre partition la valeur du critère sera inférieure ou égale à N. □

Cependant, bien que le valeur du critère atteigne sa borne supérieure du critère pour la partition triviale, cela n'implique pas que celle-ci soit la seule. En effet, selon la topologie du graphe il peut y avoir une autre partition qui fasse que la valeur du critère éteigne la valeur N lorsque le graphe est réflexif.

Désormais nous considérons que les graphes que nous traiterons sont non réflexifs.

Si le graphe n'est pas réflexif la solution optimale du critère de Michalski-Goldberg n'est pas triviale, car la valeur du critère est nulle dans ce cas. Encore plus, si le graphe n'est pas réflexif la valeur du critère est bornée entre 0 et $N - \kappa : 0 \geq F_G \geq (N - \kappa)$.

De plus pour un graphe complet la valeur optimale de ce critère est la partition grossière. En effet si le graphe est complet le critère vaut [3] :

$$F_G(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \hat{x}_{ii'} - \sum_{i=1}^{N} \frac{1}{x_{i.}} = N - \kappa. \tag{B.3}$$

L'expression (B.3) atteint son maximum lorsque κ est minimal et égal à 1. Donc, pour la partition grossière.

L'impact de la fusion de deux classes sur la valeur du critère de Michalski-Goldberg peut être calculé en tenant compte de la contribution de chaque classe d'équivalence à la valeur du critère, soit :

$$F_G = \sum_{j=1}^{\kappa} \frac{l_j}{n_j},$$

où l_j dénote le nombre d'arêtes intra-classe de la classe j et n_j le nombre de sommets appartenant à la classe j.

Ainsi, avant la fusion le critère vaut (nous utilisons l'exposant B pour *before* en anglais) :

$$F_G^B = \frac{l_1}{n_1} + \frac{l_2}{n_2} + \sum_{j=3}^{\kappa} \frac{l_j}{n_j},$$

après la fusion le critère vaut (nous utilisons l'exposant A pour *after* en anglais) :

$$F_G^A = \frac{l + l_1 + l_2}{n_1 + n_2} + \sum_{j=3}^{\kappa} \frac{l_j}{n_j}.$$

La contribution vaut, alors :

$$C_G = \frac{l + l_1 + l_2}{n_1 + n_2} - \frac{l_1}{n_1} - \frac{l_2}{n_2} = \frac{n_1 n_2 l_1 + n_1 n_2 l_2 + n_1 n_2 l - n_2(n_1 + n_2) l_1 - n_1(n_1 + n_2) l_2}{(n_1 + n_2) n_1 n_2}$$
$$= \frac{n_1 n_2 l_1 + n_1 n_2 l_2 + n_1 n_2 l - n_2 n_1 l_1 - n_2^2 l_1 - n_1^2 l_2 - n_1 n_2 l_2}{(n_1 + n_2) n_1 n_2} = \frac{n_1 n_2 l - n_2^2 l_1 - n_1^2 l_2}{(n_1 + n_2) n_1 n_2}.$$

3. Cette expression est obtenue en se servant des propriétés de la matrice $\hat{\mathbf{X}}$ qui est une matrice bistochastique et de l'expression (2.17).

Donc, pour que la fusion puisse avoir lieu, i.e. $C_G > 0$ la condition suivante doit être vérifiée :

$$\frac{l}{n_1 n_2} > \frac{l_1}{n_1^2} + \frac{l_2}{n_2^2}. \tag{B.4}$$

Ce résultat montre que la densité d'arêtes entre les classes $\mathcal{C}_1$ et $\mathcal{C}_2$ doit être supérieure à la somme des densités d'arêtes de $\mathcal{C}_1$ et $\mathcal{C}_2$ pour que la fusion contribue à la maximisation du critère.

B.2.2 Le critère de Michalski-Goldberg pondéré

Si dans l'expression (B.2) nous prenons $\phi_{ii'} = a_{ii'}, \psi_{ii'} = \frac{x_{ii'}}{x_{i.}x_{.i'}}, \bar{\phi}_{ii'} = 0 \,\forall (i, i')$ nous obtenons le critère de Mischalski-Goldberg pondéré en $\mathbf{X}$ (MGP) qui s'écrit :

$$F_{\mathrm{MGP}}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}\hat{\hat{x}}_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N} \frac{a_{ii'}x_{ii'}}{x_{i.}x_{.i'}},$$

où $\hat{\hat{x}}_{ii'}$ est le terme général de la matrice de densité ou de taux d'occupation $\mathbf{U}$ (voir définition 5.3).

Il s'agit d'une fonction à maximiser. Si le graphe est réflexif la solution optimale obtenue via sa maximisation est la partition triviale (où tous les sommets sont isolés) :

Lemme B.2. *La solution optimale du critère de Michalski-Goldberg pondéré est la solution triviale si le graphe est réflexif.*

Démonstration. Selon l'équation (2.18) $\sum_{i=1}^{N}\sum_{i'=1}^{N} \hat{\hat{x}}_{ii'} = \kappa$. Ce qui implique que $\forall\, \mathbf{X} \quad F_{\mathrm{MGP}}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}\hat{\hat{x}}_{ii'} \leq \kappa \leq N$. Si le graphe est réflexif $a_{ii} = 1 \,\forall i$ cette borne est atteinte pour la partition triviale : $F_{\mathrm{MGP}}(X^{\mathrm{triv}}) = N$. □

Si le graphe n'est pas réflexif la valeur du critère est nul pour la partition triviale : $F_{\mathrm{MGP}}(X^{\mathrm{triv}}) = 0$, soit la borne inférieure du critère. Donc lorsque le graphe n'est pas réflexif la solution optimale de ce critère n'est pas triviale et nous ne sommes pas obligés de fixer le nombre de classes en avance. Désormais nous considérons que les graphes que nous traitons sont non réflexifs.

Le théorème suivant énonce la solution optimale rendue par ce critère pour un graphe complet :

Lemme B.3. *Pour un graphe complet non réflexif la solution optimale du critère de Michalski-Goldberg pondéré contient $\lfloor \frac{N}{2} \rfloor$ classes. Si N est pair toutes les classes ont 2 sommets. Si N est impair la solution optimale contient $\left(\frac{N-3}{2}\right)$ classes à 2 sommets et une classe à 3 sommets.*

En effet, si le graphe est complet le critère vaut :

$$F_{\mathrm{MGP}}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \hat{\hat{x}}_{ii'} - \sum_{i=1}^{N} \frac{1}{x_{i.}^2} = \kappa - \sum_{j=1}^{\kappa} \frac{1}{n_j}, \tag{B.5}$$

où n_j est l'effectif de la classe j. Cette expression est obtenue à partir de (2.18) et des propriétés de la matrice de densité ou de taux d'occupation $\mathbf{U}$ (définition 5.3).

Démonstration. Dans l'expression (B.5) le nombre de classes κ étant fixé, la quantité $\sum_{j=1}^{\kappa} \frac{1}{n_j}$ est minimale si toutes les classes sont de la même taille, autrement dit, si $n_j = \frac{N}{\kappa} \forall i$, et dans ce cas $\sum_{j=1}^{\kappa} \frac{1}{n_j} = \frac{\kappa^2}{N}$. Donc, à partir de (B.5) le critère vaut :

$$F_{\mathrm{MGP}}(X) = \kappa - \frac{\kappa^2}{N}.$$

Cette dernière expression est une fonction concave de κ. Nous obtenons son maximum en dérivant par rapport à κ, ce qui donne $\kappa^{\mathrm{opt}} = \frac{N}{2}$. Le critère vaut alors :

$$F_{\mathrm{MGP}}(X)^{\mathrm{OPT}} = \frac{N}{2} - \frac{N^2}{4N} = \frac{N}{4}.$$

Cependant le nombre optimal de classes est $\kappa^{\mathrm{opt}} = \frac{N}{2}$ lorsque N est pair. Lorsque N est impair $\frac{N}{2}$ n'est pas entier et nous avons deux choix de κ :

1. Soit $\kappa^{\mathrm{opt}} = \frac{N-1}{2}$ et la partition optimale contient $\frac{N-3}{2}$ classes à 2 sommets et 1 classe à 3 sommets. Le critère vaut dans ce cas :

$$F_{\mathrm{MGP}}(X) = \kappa - \sum_{j=1}^{\kappa} \frac{1}{n_j} = \frac{N-1}{2} - \frac{N-3}{4} - \frac{1}{3} = \frac{N}{4} - \frac{1}{12}.$$

2. Soit $\kappa^{\mathrm{opt}} = \frac{N+1}{2}$ et la partition optimale contient $\frac{N-1}{2}$ classes à 2 sommets et 1 classe à 1 sommet. Le critère vaut dans ce cas :

$$F_{\mathrm{MGP}}(X) = \kappa - \sum_{j=1}^{\kappa} \frac{1}{n_j} = \frac{N+1}{2} - \frac{N-1}{4} - 1 = \frac{N}{4} - \frac{1}{4}.$$

La valeur du critère étant toujours supérieure pour le premier cas nous obtenons que N est impair la partition optimale contiendra $\frac{N-1}{2}$ classes dont $\frac{N-3}{2}$ classes à 2 sommets et une classe à 3 sommets. □

L'impact de la fusion de deux classes au critère de Michalski-Goldberg pondéré peut être calculé de façon analogue au calcul de l'impact sur le critère de Michalski-Goldberg. Avant la fusion le critère vaut :

$$F^B_{\mathrm{MGP}}(X) = \frac{l_1}{n_1^2} + \frac{l_2}{n_2^2} + \sum_{j=3}^{\kappa} \frac{l_j}{n_j^2},$$

après la fusion le critère vaut :

$$F^A_{\mathrm{MGP}}(X) = \frac{l + l_1 + l_2}{(n_1 + n_2)^2} + \sum_{j=3}^{\kappa} \frac{l_j}{n_j^2}.$$

La contribution vaut alors :

$$C_{\mathrm{MGP}} = \frac{l+l_1+l_2}{(n_1+n_2)^2} - \frac{l_1}{n_1^2} - \frac{l_2}{n_2^2} = \frac{n_1^2 n_2^2 l_1 + n_1^2 n_2^2 l_2 + n_1^2 n_2^2 l - n_2^2 (n_1+n_2)^2 l_1 - n_1^2 (n_1+n_2)^2 l_2}{(n_1+n_2)^2 n_1^2 n_2^2}$$
$$= \frac{n_1^2 n_2^2 l_1 + n_1^2 n_2^2 l_2 + n_1^2 n_2^2 l - n_2^2 (n_1^2+2n_1n_2+n_2^2) l_1 - n_1^2 (n_1^2+2n_1n_2+n_2^2) l_2}{(n_1+n_2)^2 n_1^2 n_2^2}$$

$$= \frac{n_1^2 n_2^2 l_1 + n_1^2 n_2^2 l_2 + n_1^2 n_2^2 l - n_2^2 n_1^2 l_1 - 2 l_1 n_2^3 n_1 - n_2^4 l_1 - n_1^4 l_2 - 2 n_1^3 n_2 l_2 - n_1^2 n_2^2 l_2}{(n_1+n_2)^2 n_1^2 n_2^2}$$

$$= \frac{n_1^2 n_2^2 l - 2 l_1 n_2^3 n_1 - n_2^4 l_1 - n_1^4 l_2 - 2 n_1^3 n_2 l_2}{(n_1+n_2)^2 n_1^2 n_2^2}.$$

Pour que la fusion puisse avoir lieu, i.e. $C_{\text{MGP}} > 0$ la condition suivante doit être vérifiée :

$$l > 2 l_1 \frac{n_2}{n_1} + 2 l_2 \frac{n_1}{n_2} + l_1 \frac{n_2^2}{n_1^2} + l_2 \frac{n_1^2}{n_2^2}.$$

Cette condition implique que le nombre d'arêtes entre $\mathcal{C}_1$ et $\mathcal{C}_2$ doit vérifier :

$$l > \left(l_1 \frac{n_2}{n_1} \left(2 + \frac{n_2}{n_1} \right) + l_2 \frac{n_1}{n_2} \left(2 + \frac{n_1}{n_2} \right) \right). \tag{B.6}$$

Nous pouvons comparer la condition (B.6) à la condition obtenue dans (B.4) pour le critère de Michalski-Goldberg :

$$l > \left(l_1 \frac{n_2}{n_1} + l_2 \frac{n_1}{n_2} \right). \tag{B.7}$$

Le terme gauche de l'inégalité (B.6) étant toujours supérieure à celle de l'inégalité (B.7) la contribution au critère de Michalski-Goldberg est toujours supérieure à celle du critère de Michalski-Goldberg pondéré quelles que soient les caractéristiques des sous-graphes à fusionner. Par conséquent, le nombre de classes obtenu via l'optimisation du critère de Michalski-Goldberg pondéré sera toujours supérieur à celui obtenu via l'optimisation du critère de Michalski-Goldberg :

$$\kappa_{\text{MGP}} > \kappa_G \tag{B.8}$$

où κ_{MGP} et κ_G dénotent le nombre optimal de classes obtenu via l'optimisation du critère de Michalski-Goldberg pondéré et du critère Michalski-Goldberg respectivement.

B.2.3 Le critère de Mancoridis-Gansner

Si dans l'expression (B.2) nous choisissons : $\phi_{ii'} = a_{ii'}$, $\psi_{ii'} = \hat{\hat{x}}_{ii'} = \frac{x_{ii'}}{x_{i.} x_{.i'}}$, $\bar{\phi}_{ii'} = \bar{a}_{ii'}$ et $\bar{\psi}_{ii'} = \frac{\bar{x}_{ii'}}{x_{i.} x_{.i'}}$ nous obtenons le critère suivant :

$$F_{\text{CPP}}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \frac{a_{ii'} x_{ii'}}{x_{i.} x_{.i'}} + \sum_{i}^{N} \sum_{i'}^{N} \frac{\bar{a}_{ii'} \bar{x}_{ii'}}{x_{i.} x_{.i'}} \tag{B.9}$$

que nous appelerons *Critère de Condorcet deux fois pondéré* (CPP). Il s'agit du critère de Mancoridis-Gansner sans les coefficients $\frac{1}{\kappa}$ et $\frac{1}{\kappa(\kappa-1)}$ du terme d'accords positifs et du terme d'accords négatifs respectivement.

Ce critère possède un comportement proche de celui de Michalski-Goldberg pondéré. Si le graphe est réflexif la solution optimale qui maximise la valeur de ce critère est la partition triviale où chaque sommet est classé dans sa propre classe et $\kappa = N$ comme l'énonce le lemme suivant :

Lemme B.4. *La solution optimale du critère de Condorcet deux fois pondéré est la partition triviale si le graphe est réflexif.*

Démonstration. Le terme d'accords positifs est borné par κ : $\sum_{i=1}^{N}\sum_{i'=1}^{N} a_{ii'}\hat{\hat{x}}_{ii'} \leq \kappa \leq N$ et atteint son maximum lorsque $\kappa = N$ et cela a lieu lorsque $\mathbf{X}$ correspond à la partition triviale si le graphe est réflexif.
Le terme d'accords négatifs $\sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{a}_{ii'}\bar{\hat{\hat{x}}}_{ii'}$ atteint son maximum aussi lorsque $\mathbf{X}$ est à la partition triviale et vaut dans ce cas $\sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{a}_{ii'}\bar{\hat{\hat{x}}}_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N} \bar{a}_{ii'} = (N^2 - 2M)$ car $\bar{\hat{\hat{x}}}_{ii'} = 1\,\forall i \neq i'$. Donc le critère vaut : $F_{\mathrm{CPP}}(X^{\mathrm{triv}}) = N + (N^2 - 2M)$. □

Si le graphe n'est pas réflexif la solution optimale de ce critère n'est pas triviale et de plus $F_{CPP}(X^{\mathrm{triv}}) = 0$, soit la borne inférieure du critère car $0 \leq F_{\mathrm{CPP}}(X^{\mathrm{triv}}) \leq \kappa^2$.

Désormais nous considérons que les graphes sont non réflexifs. Nous obtenos le résultat suivant pour les graphes complets :

Lemme B.5. *Pour un graphe complet non réflexif la solution optimale du critère de Condorcet deux fois pondéré non pondéré contient $\lfloor \frac{N}{2} \rfloor$ classes. Si N est pair toutes les classes ont 2 sommets. Si N est impair la solution optimale contient $\left(\frac{N-3}{2}\right)$ classes à 2 sommets et 1 classe à 3 sommets.*

En effet, pour un graphe complet la solution optimale du critère est la même que celle du critère de Michalski-Goldberg pondéré. Si le graphe est complet $\bar{\mathbf{A}} = \mathbf{I}_N$, par conséquent le terme d'accords négatifs sera toujours nul $\left(\sum_i^N \sum_{i'}^N \frac{\bar{a}_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}} = 0\right)$ car $\bar{\hat{\hat{x}}}_{ii} = 0\,\forall i$ et nous obtenons $F_{\mathrm{CPP}} = F_{\mathrm{MGP}}$.

Comme le terme d'accords positifs est borné par κ et le terme d'accords négatifs est borné par $\kappa(\kappa - 1)$ ce critère à tendance à générer beaucoup de classes, même pour un graphe complet. Le critère de Mancoridis-Gansner évite ce problème en divisant le terme d'accords positifs par κ et le terme d'accords négatifs par $\kappa(\kappa-1)$. En effet, pour ce critère $\psi_{ii'} = \frac{\hat{\hat{x}}_{ii'}}{\kappa}$ et $\bar{\psi}_{ii'} = \frac{\bar{\hat{\hat{x}}}_{ii'}}{\kappa(\kappa-1)}$ (voir expression (5.48)) :

$$F_{MG}(X) = \frac{1}{\kappa}\sum_{i=1}^{N}\sum_{i'=1}^{N}\frac{a_{ii'}x_{ii'}}{x_{i.}x_{.i'}} + \frac{1}{\kappa(\kappa-1)}\sum_{i}^{N}\sum_{i'}^{N}\frac{\bar{a}_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}}$$

pour $\kappa > 1$.

Il s'agit d'un critère borné par $[0, 2]$ qui ne rend pas de partitions triviales et on n'est pas obligé de fixer à l'avance le nombre de classes. L'impact sur le critère suite à la fusion de deux classes est une expression longue qui ne fera pas l'objet de notre étude.

B.2.4 Le critère de Wei-Cheng (Ratio-Cuts)

Dans l'expression (B.2) en choisissant $\phi_{ii'} = 0 \quad \forall i, i'$, $\bar{\phi}_{ii'} = a_{ii'}$ et $\bar{\psi}_{ii'} = \frac{\bar{x}_{ii'}}{x_{i.}x_{.i'}}$ nous obtenons le critère de *Ratio-Cuts* (voir expression (5.55)) :

$$F_{\text{Rcut}}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \frac{a_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}}.$$

C'est un critère à minimiser. La partition qui minimise ce critère est la **partition grossière** ($\kappa = 1$ et tous les sommets sont classés dans une seule et unique classe) si l'on ne fixe pas le nombre de classes à l'avance.

En effet, la partie variable de ce critère étant égale au terme général de la matrice $\bar{\mathbf{U}}$ (complémentaire de la matrice de taux de densité $\mathbf{U}$) et les termes $a_{ii'}$ étant toujours positives, la valeur de ce critère est minimale et nulle lorsque $u_{ii'} = 0 \Leftrightarrow x_{ii'} = 0 \quad \forall(i, i')$. Ce résultat est une conséquence de l'absence du terme d'accords positifs, en effet, ce critère n'est pas équilibré, ce qui provoque que sa solution optimale soit la partition grossière.

B.2.5 Critère de la Différence de Profils

Nous avons vu que le critère de la Différence de Profils cherche à minimiser la distance euclidienne entre les matrices $\hat{\mathbf{A}}$ et $\hat{\mathbf{X}}$ (voir expression (5.56)) :

$$F_{DP}(X) = ||\hat{\mathbf{A}} - \hat{\mathbf{X}}||^2 = \sum_{i}^{N}\sum_{i'}^{N} \left(\hat{a}_{ii'} - \hat{x}_{ii'}\right)^2.$$

La solution optimale de ce critère n'est pas du tout triviale. De plus pour un graphe complet la valeur optimale de ce critère est la partition grossière. En effet, si le graphe est complet et non réflexif $\hat{a}_{ii'} = \frac{1}{(N-1)}$ (car chaque sommet est connecté aux $(N-1)$ sommets restants dans le graphe) le critère vaut :

$$F_{DP}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N} \hat{a}_{ii'}^2 - 2\sum_{i=1}^{N}\sum_{i'=1}^{N} \hat{a}_{ii'}\hat{x}_{ii'} + \kappa = -\frac{2}{(N-1)}\sum_{i=1}^{N}\sum_{i'=1}^{N} \hat{x}_{ii'} + \kappa + \frac{2}{(N-1)}\sum_{i=1}^{N}\frac{1}{x_{i.}} + K$$

(où $K = \sum_{i=1}^{N}\sum_{i'=1}^{N} \hat{a}_{ii'}^2$ est une constante qui n'intervient pas dans le processus d'optimisation),

$$= -\frac{2N}{(N-1)} + \kappa + \frac{2\kappa}{(N-1)} = -\frac{2N}{(N-1)} + \frac{(N+1)\kappa}{(N-1)}.$$

Cette dernière quantité est minimale lorsque κ est minimal et vaut 1. Donc la partition optimale pour un graphe complet non réflexif est la partition grossière.

Si le graphe est complet et réflexif $\hat{a}_{ii'} = \frac{1}{N}$ le critère vaut :

$$F_{DP}(X) = -\frac{2}{N}\sum_{i=1}^{N}\sum_{i'=1}^{N} \hat{x}_{ii'} + \kappa + \frac{2}{N}\sum_{i=1}^{N}\frac{1}{x_{i.}} + K = -2 + \frac{(N+2)\kappa}{N} + K.$$

Cette dernière expression atteint son minimum lorsque κ est minimal et égal à 1. Donc, pour un graphe complet réflexif la partition optimale qui optimise ce critère est la partition grossière.

L'impact de la fusion de deux classes au critère de la Différence de Profils peut être calculé en tenant compte que minimiser ce critère revient à maximiser l'expression suivante (voir équation (5.58)) :

$$F_{DP}(X) = 2\sum_{i}^{N}\sum_{i'}^{N}\hat{a}_{ii'}\hat{x}_{ii'} - \kappa + K.$$

En notant $\hat{l}_1 = \sum_{i\in\mathcal{C}_1}^{n_1}\sum_{i'\in\mathcal{C}_1}^{n_1}\hat{a}_{ii'}$, $\hat{l}_2 = \sum_{i\in\mathcal{C}_2}^{n_2}\sum_{i'\in\mathcal{C}_2}^{n_2}\hat{a}_{ii'}$, $\hat{l} = \sum_{i\in\mathcal{C}_1}^{n_1}\sum_{i'\in\mathcal{C}_2}^{n_2}\hat{a}_{ii'}$ et κ_0 le nombre de classes avant la fusion, le critère avant la fusion vaut :

$$F_{DP}^{B} = \frac{\hat{l}_1}{n_1} + \frac{\hat{l}_2}{n_2} - \kappa_0 + K,$$

après la fusion le nombre de classes sera $(\kappa_0 - 1)$, donc la valeur du critère sera :

$$F_{DP}^{A} = \frac{\hat{l} + \hat{l}_1 + \hat{l}_2}{n_1 + n_2} - (\kappa_0 - 1) + K.$$

La contribution aura pour valeur :

$C_{DP} = \frac{\hat{l}+\hat{l}_1+\hat{l}_2}{n_1+n_2} - \frac{\hat{l}_1}{n_1} - \frac{\hat{l}_2}{n_2} + 1 = \frac{n_1n_2\hat{l}_1+n_1n_2\hat{l}_2+n_1n_2\hat{l}-n_2(n_1+n_2)\hat{l}_1-n_1(n_1+n_2)\hat{l}_2}{(n_1+n_2)n_1n_2} + 1$
$= \frac{n_1n_2\hat{l}_1+n_1n_2\hat{l}_2+n_1n_2\hat{l}-n_2n_1\hat{l}_1-n_2^2\hat{l}_1-n_1^2\hat{l}_2-n_1n_2\hat{l}_2+(n_1+n_2)n_1n_2}{(n_1+n_2)n_1n_2} = \frac{n_1n_2\hat{l}-n_2^2\hat{l}_1-n_1^2\hat{l}_2+(n_1+n_2)n_1n_2}{(n_1+n_2)n_1n_2}.$

Donc, pour que la fusion puisse avoir lieu, i.e. $C_{DI} > 0$ la condition suivante doit être vérifiée : $(n_1n_2\hat{l} - n_2^2\hat{l}_1 - n_1^2\hat{l}_2 + (n_1 + n_2)n_1n_2) > 0$, ce qui implique :

$$\hat{l} + (n_1 + n_2) > \hat{l}_1\frac{n_2}{n_1} + \hat{l}_2\frac{n_1}{n_2} \tag{B.10}$$

La contribution dépend de la distribution des degrés du graphe car les quantités $\hat{l}_1$, $\hat{l}_2$ et $\hat{l}$ sont fonction des degrés des sommets. Le membre gauche de l'inégalité (B.10) montre que plus grandes sont les tailles des classes à fusionner n_1 et n_2 plus facilement la fusion aura lieu.

Annexe C

Résultats d'applications pratiques

C.1 Club de Karaté de Zachary

Le graphe de Zachary est un jeu de données fréquemment utilisée en analyse de réseaux sociaux. Un désaccord entre l'administrateur (sommet 1) et l'instructeur du club (sommet 34) a séparé le réseau en deux groupes de taille semblable [1]. Chaque groupe est représenté par une couleur différente.

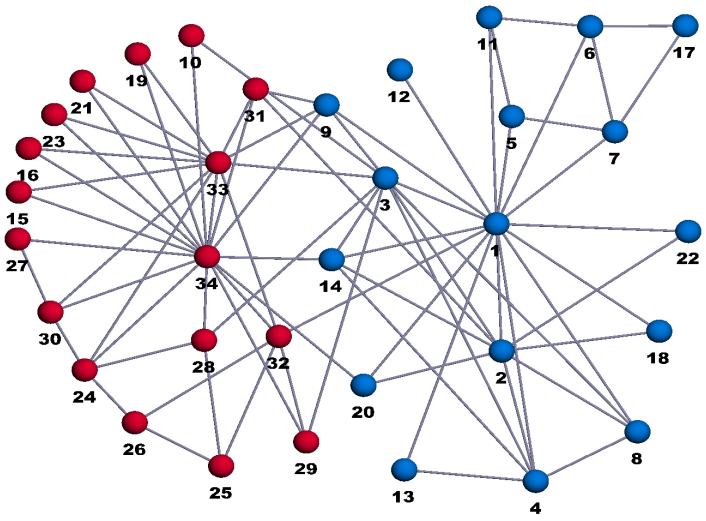

FIGURE C.1 – Le réseau "Club de karaté de Zachary" après scission.

Le tableau C.1 montre le nombre de classes obtenu pour chaque critère [2].

Quelques remarques importantes par rapport aux partitions obtenues :

- L'algorithme a rendu la même partition optimale pour les critères de Newman-Girvan, Écart à l'Indétermination et Modularité Équilibrée.

1. La figure a été prise sur le site internet http://ifisc.uib-csic.es/~jramasco/Structure.html

2. Nous avons choisi $\alpha = 0,2$ pour le critère d'Owsiński-Zadrożny car les densités intra-classe des classes bleue et rouge (voir la figure C.1) sont 0,23 et 0,25 respectivement. Donc, nous voulions connaitre la partition rendue par l'algorithme si le critère demande une densité intra-classe juste un peu au-dessous de celles des classes réelles.

Critère	**Nombre de classes**	**Commentaires**
Zahn-Condorcet	19	dont 12 classes à 1 sommet isolé
Owsiński-Zadrożny	7 ($\alpha = 0,2$)	dont 3 classes à 1 sommet isolé
Écart à l'Uniformité	6	dont 2 classes à 1 sommet isolé
Newman-Girvan	4	
Écart à l'Indétermination	4	
Modularité Équilibrée	4	
Différence de Profils	4	
Michalski-Goldberg	11	dont 1 classe à 1 sommet isolé

TABLE C.1 – Nombre de classes obtenu selon critère pour le Club de "karaté de Zachary"

- Comme $N = 34$ et $M = 78$, le degré moyen sera $d_{av} \cong 4,6$ et la densité globale d'arêtes sera $\delta \cong 0,13$. Comme $\frac{1}{2} > \alpha > \delta$ nous avons bien (voir expressions (6.14), (6.16) et (6.15)) : $\kappa_{ZC} > \kappa_{OZ} > \kappa_{\text{UNIF}}$.

- La figure C.2 montre les partitions trouvées par tous les critères [3] présentés au tableau C.1 (les partitions obtenues par les critères de Zahn-Condorcet et Michalski-Goldberg ne sont pas présentées car elles possèdent beaucoup de petites classes et de sommets isolés).

  La Figure C.2 montre que les résultats obtenus avec les six critères sont proches. Aucun critère ne trouve la partition réelle en deux classes de la figure C.1. Ils trouvent tous au moins quatre classes.

  Voici l'interprétation des partitions trouvées :
  - Les six critères sont d'accord pour séparer les sommets, 5,6,7,11 et 17 (sommets en rose en haut à droite).

  - les six critères mettent les sommets 24, 25, 26, 28 et 32 (sommets en vert en bas à gauche) dans une classe à part. Newman-Girvan, l'Écart à l'Indétermination et la Modularité Équilibrée mettent aussi le sommet 29 dans cette classe.

  - Les critères d'Owsiński-Zadrożny et l'Écart à l'Uniformité isolent le sommet 10 (en jaune) tandis que Newman-Girvan, l'Écart à l'Indétermination et la Modularité Équilibrée le classent dans la classe de l'instructeur (sommet 34) et la Différence de Profils le classe dans la classe de l'administrateur (sommet 1). Dans la construction du réseau réel Zachary avait hésité avec la classe du sommet 10.

  - Le sommet 12 (couleur blue claire) est isolé par Owsiński-Zadrożny et l'Écart à l'Uniformité. C'est souvent le cas pour les sommets à faible degré qui sont adjacents aux sommets à degré élevé. Le sommet 12 a pour seul voisin le sommet 1 (administrateur) qui occupe bien une position centrale dans le réseau. C'est aussi

3. Les graphes ont été dessinés avec le logiciel Gephi (voir http://gephi.org/).

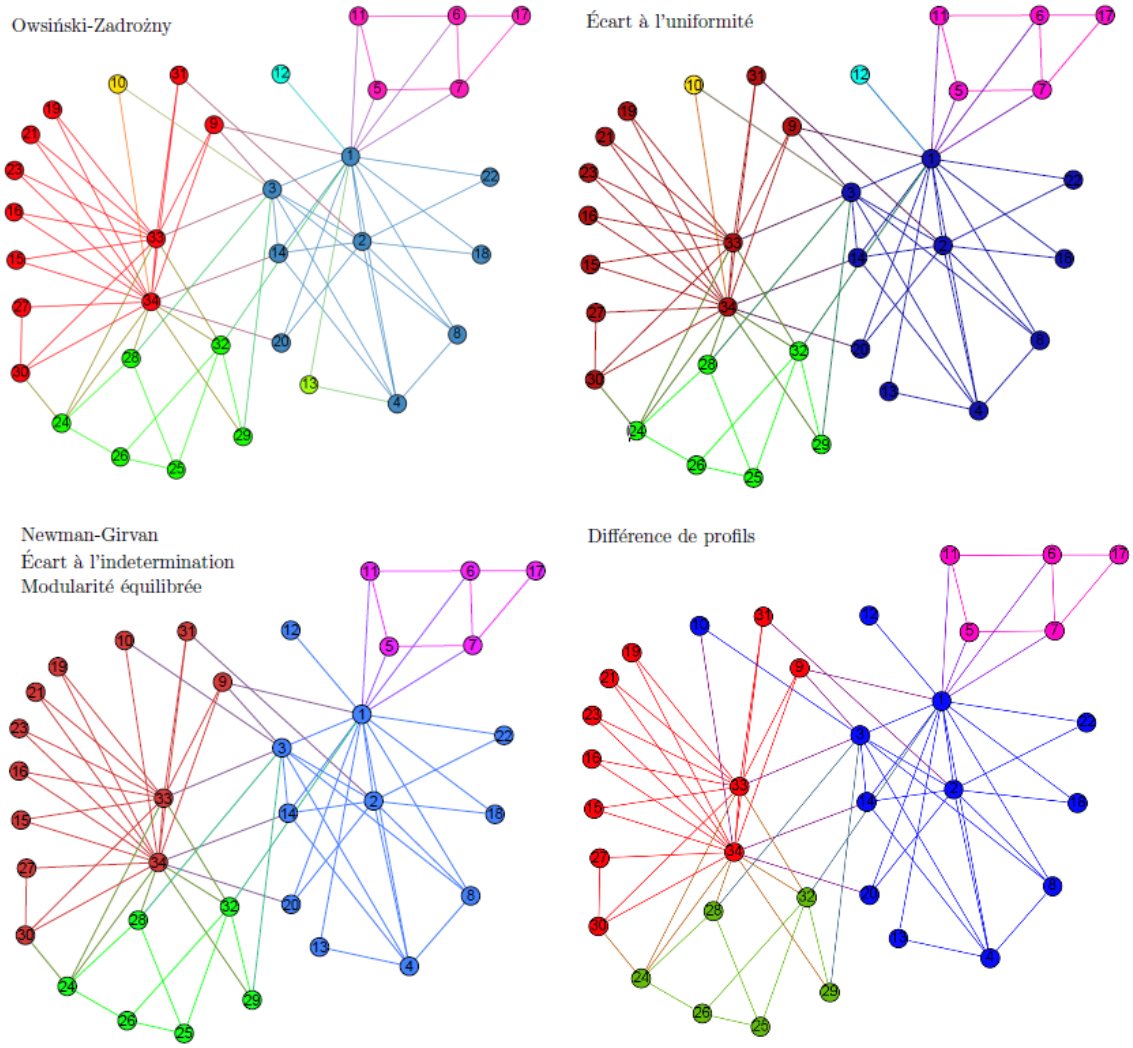

Figure C.2 – Résultat de la Modularisation du réseau "club de karaté de Zachary".

le cas du sommet 13 qui a été isolé par le critère d'Owsiński-Zadrożny.

– L'optimisation directe du critère de Condorcet-Zahn partitionnait le graphe en 19 classes. Le d'Owsinski-Zadrozny avec $\alpha = 0,2$ a permis de rendre Condorcet-Zahn plus flexible.

Les tableaux suivants montrent les tailles de classes, le nombre d'arêtes intra-classe et la densité intra-classe pour chaque critère. Le calcul de la densité intra-classe se fait comme suit : étant donné une classe à n_j sommets et $|E_j|$ arêtes intra-classe, la densité de cette classe δ_j vaut :

$$\delta_j = \frac{|E_j|}{\frac{n_j(n_j-1)}{2}}, \tag{C.1}$$

soit le ratio entre le nombre d'arêtes existant et le nombre maximal d'arêtes qui pourraient exister.

– **Zahn-Condorcet** :

Communauté	Effectif	Arêtes intra-classe	Densité
1	6	14	0,93
3	4	6	1
2	3	3	1
14	3	3	1
5	2	1	1
15	2	1	1
17	2	1	1

– **Owsinski-Zadrozny** :

Communauté	Effectif	Arêtes intra-classe	Densité
5	11	20	0,37
0	10	22	0,49
1	5	5	0,5
4	5	6	0,6

– **Ecart à l'uniformité**

Communauté	Effectif	Arêtes intra-classe	Densité
5	12	23	0,35
0	10	22	0,49
2	5	6	0,6
4	5	5	0,5

– **Newman-Girvan, Ecart à l'indétermination et Modularité Équilibrée**

Communauté	Effectif	Arêtes intra-classe	Densité
4	12	21	0,32
1	11	23	0,42
3	6	7	0,47
2	5	6	0,6

– **Différence de Profils**

Communauté	Effectif	Arêtes intra-classe	Densité
0	13	26	0,33
1	6	7	0,47
3	10	17	0,38
2	5	6	0,6

– **Michalski-Goldberg**

Communauté	Effectif	Arêtes intra-classe	Densité
11	7	11	0,52
1	6	9	0,6
2	3	2	0,67
3	3	2	0,67
4	3	3	1
9	3	3	1
5	2	1	1
6	2	1	1
8	2	1	1
10	2	1	1

La partition obtenue par le critère de Zahn-Condorcet contient des classes à plus de 90% d'aêtes intra-classe. En effet, les classes obtenues avec ce critère sont dans la plupart de cas des cliques. Quant au critère de Michalksi Goldberg, toutes les classes obtenues possèdent une densité supérieure à 50% et certaines même à 100%, en effet, ce critère a aussi tendance à générer des petites classes à haute densité d'arêtes. En ce qui concerne les autres critères les densités de leurs classes sont comprises entre 30% et 60%. On vérifie bien que pour le critère d'Écart à l'Uniformité et pour le critère d'Owsiński-Zadrożny toutes les classes possèdent une densité supérieure à $\delta = 0,13$ et $\alpha = 0,2$ respectivement.

C.2 American College football

Ce réseau contient 115 sommets (équipes) et 613 arêtes (matchs entre équipes). Les équipes sont divisées en 12 tournois contenant entre 7 à 13 équipes chacun. Les matchs sont plus fréquents entre les membres du même tournoi qu'entre les membres de tournois différents.

L'accord entre la partition connue (représentée par les tournois) et la partition trouvée par chaque critère a été calculé en comparant terme à terme les éléments de la matrice relationnelle de Condorcet associée à chaque partition obtenue avec celle de la partition connue, soit avec l'indice de Rand pour comparer deux partitions (voir Rand [1971], Marcotorchino [1984b], Saporta [1988], Saporta and Youness [2002]). Si l'on note ρ_{agree} le pourcentage d'accords positifs plus accords négatifs, cette quantité vaut

$$\rho_{\text{agree}} = \frac{\sum_{i=1}^{N}\sum_{i'=1}^{N}(y_{ii'}x_{ii'} + \bar{y}_{ii'}\bar{x}_{ii'})}{N^2}, \tag{C.2}$$

où $\mathbf{Y}$ et $\mathbf{X}$ représentent la matrice relationnelle de la partition réelle, en *tournois*, et la matrice relationnelle de la partition trouvée via l'algorithme de Louvain, pour chaque critère respectivement.

Dans le cas idéal où la matrice obtenue est identique à celle de la partition originale, tous les termes de ces deux matrices sont identiques, et on obtient un accord égal à 100%.

Le tableau C.2 montre les résultats obtenus pour tous les critères :

- Les partitions obtenues avec les critères Newman-Girvan, Ecart à l'Indétermination et Modularité Équilibrée sont identiques.

Critère	κ	NCCI	Total Accords	ρ_{agree} (%)
Graphe Réel	12		13225	100
Zahn-Condorcet	16	7	12927	97,7
Newman-Girvan, Écart à l'Indétermination, Modularité Équilibrée	10	6	12817	96,9
Écart à l'Uniformité	10	5	12777	96,6
Différence de Profils	9	4	12569	95,0
Michalski-Goldberg	27	0	12517	94,6

TABLE C.2 – Résultats trouvés avec le réseau "College football". NCCI est le nombre de classes. correctement identifiées par chaque critère

- Le tableau C.2 montre que l'on obtient un pourcentage d'accords élevé pour tous les critères. Cependant, le nombre optimal de classes varie d'un critère à l'autre. En effet, Newman-Girvan, l'Écar à l'indétermination, l'Écart à l'Uniformité et la Différence de Profils sous estiment le nombre de classes tandis que les partitions obtenues avec les critères de Zahn-Condorcet et Michalski-Goldberg contiennent un nombre de classes supérieur au nombre de classes attendu, à savoir 12. Le critère de Michalski-Goldberg n'identifie correctement aucune classe, en effet, il coupe le graphe en toutes petites classes.

C.3 Le réseau de musiciens de "Jazz"

Le réseau "jazz" contient $N = 198$ sommets et $M = 2742$ arêtes. Comme le montre le tableau 7.1 les partitions trouvées avec les critères de Newman-Girvan, Écart à l'Indétermination et Modularité Équilibrée sont de taille différente, donc nous pouvons les analyser pour comprendre ces différences. Cela nous permettra de comparer ces trois critères.

Les tableaux suivants montrent, pour les partitions trouvées avec ces trois critères et pour chaque classe j trouvée :

- La taille de la classe n_j.
- Le degré moyen des sommets d_{av}^j.
- L'écart-type σ_j des degrés des sommets.
- le coefficient de variation des degrés cv_j des sommets de la classe.

Pour la partition trouvée avec le critère de Newman-Girvan

n_j	d_{av}^j	σ_j	cv_j
62	32,3	18,5	0,57
53	30,5	16,2	0,53
61	20,3	14,1	0,69
22	28,4	20,1	0,71

Pour la partition trouvée avec la Modularité Équilibrée

n_j	d_{av}^j	σ_j	cv_j
60	33,1	18,2	0,55
53	31,3	16,3	0,52
61	20,3	14,1	0,69
23	26	19,4	0,75
1	1	0	0

Pour la partition trouvée avec l'Écart à l'Indétermination

n_j	d_{av}^j	σ_j	cv_j
63	19,8	14,2	0,71
63	33,7	16	0,48
18	13,8	5,2	0,37
51	36,4	17,7	0,49
2	2,5	2,1	0,85
1	1	0	0

La partition trouvée avec la Modularité Équilibrée contient une classe de plus que le critère de Newman-Girvan. Ce critère isole un sommet à degré 1. En effet, ce sommet il a un degré très bas pour appartenir à la classe de son seul voisin. La classe des son voisin possède un degré moyen proche de 33.

Le critère d'Écart à l'Indétermination génére six classes dont une correspond à un sommet isolé à degré 1 et l'autre contient deux sommets à degré 2 et 3 (car leur degré moyen est 2,5). Il a donc, créé deux classes à degré moyen bas.

Deux indicateurs de la dispertion sont l'écart-type et le coefficient de variation (rapport entre l'écart-type et la moyenne). La figure C.3 montre le coefficient de variation de la variable "degré" pour les classes trouvées avec les trois critères.

La figure C.3 montre clairement que le coefficient de variation pour le critère d'Écart à l'Indétermination est inférieur à ceux des critères de Newman-Girvan et la Modularité Équilibrée.

L'exemple du réseau des musiciens "Jazz" permet de montrer clairement la différence entre les partitions trouvées avec les trois critères. Cependant, les partitions trouvées avec les trois critères dépendent fortement de la distribution d'arêtes et des degrés.

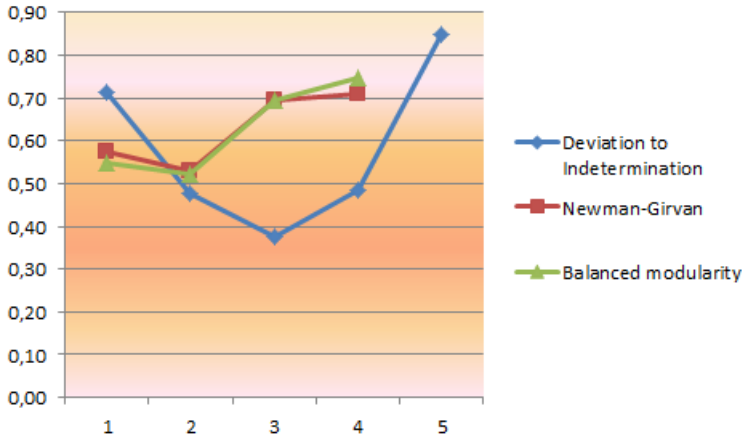

FIGURE C.3 – Coefficient de variation du degré intra-classe pour le réseau de musiciens de "Jazz"

Bibliographie

J. Ah-Pine. *Sur les Aspects Algébriques et Combinatoires de l'Analyse Relationnelle*. PhD thesis, Université Pierre et Marie Curie, LSTA, Paris, France, 2007.

J. Ah-Pine. Graph clustering by maximizing statistical association measures. In A. Tucker, editor, *IDA 2013 12th International Symposium on Intelligent Data Analysis*, pages 56–67, London, United Kingdom, 2013. Springer-Verlag.

J. Ah-Pine and F. Marcotorchino. Statistical, geometrical and logical independences between categorical variables. *Proc. of the ASMDA2007 Symposium, Chania, Greece*, 2007.

J. Ah-Pine and F. Marcotorchino. *Unifying some association criteria between partitions by using relational matrices*. Communications in Statistics - Theory and Methods, 39 (3) :531–542, 2010.

R. Albert, H. Jeong, and A.L. Barabási. Internet : Diameter of the world-wide web. *Nature*, 401(6749) :130–131, 1999.

A. Arenas, A. Fernandez, and S. Gomez. Analysis of the structure of complex networks at different resolution levels. *New Journal of Physics*, 10(5) :053039, 2008.

K.J. Arrow. *Social Choice and Individual Values*. Monograph (Yale University). Yale University Press, 1963.

K.J. Arrow and H. Raynaud. *Social Choice and Multicriterion Decision Making*. MIT Press, 1986.

T. Aynaud, V. Blondel, J.L. Guillaume, and R. Lambiotte. *chapitre 14 : Optimisation locale multi-niveaux de la modularité dans le livre "Partitionnement de graphe : optimisation et applications, Traité IC2"*. Hermes-Lavoisier, 2010.

N. Bansal, A. Blum, and S. Chawla. Correlation clustering. In *IEEE Symp. on Foundations of Computer Science*, 2002.

A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286 : 509–512, 1999.

A.L. Barabási. Dossier : La théorie de la complexité. *Journal pour la Science : La Recherche*, pages 36–49, May 2012.

A.L. Barabási and J. Frangos. *Linked : The New Science Of Networks Science Of Networks*. Perseus Publishing., 2002.

A. Bavelas. A mathematical model for group structures. *Human Organizations*, 7 :16–30, 1948.

C. Bedecarrax and F. Marcotorchino. *"La Distance de la Différence de Profils" dans le livre "Distance"*, pages 199–203. Publication Université de Haute Bretagne, 1992.

W. A. Belson. Matching and prediction on the principle of biological classification. *Survey Research Centre, London School of Economics and Political Science*, 1959.

J.P. Benzécri. *L'analyse des données : Classification Automatique*, volume 1. Dunod, 1973a.

J.P. Benzécri. *L'analyse des données, Tome II : L'analyse des correspondances*, volume 2. Dunod, Paris, 1973b.

C. Berge. *Théorie des graphes et ses applications.* Collection Universitaire des Mathématiques, Dunod, Paris, 1958.

C. Berge. *Graphes et hypergraphes.* Collection Universitaire des Mathématiques, Dunod, Paris, 1970.

V. Blondel, J.L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment P10008*, 2008.

S. Boettcher and A.G. Percus. Optimization with extremal dynamics. *Physical Review Letters*, 86 :5211–5214, 2001.

P. Boldi and S. Vigna. The WebGraph framework I : Compression techniques. In *Proceedings of the Thirteenth International World Wide Web Conference (WWW 2004)*, pages 595–601, Manhattan, USA, 2004. ACM Press.

P. Boldi, B. Codenotti, M. Santini, and S. Vigna. Ubicrawler : A scalable fully distributed web crawler. *Software : Practice & Experience*, 34(8) :711–726, 2004.

P. Bonacich. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, 2(1) :113–120, 1972.

U. Brandes, D. Delling, M. Gaertler, R. Goerke, M. Hoefer, Z. Nikoloski, and D. Wagner. Maximizing modularity is hard. 2006.

U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner. On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering*, 20 (2) :172–188, 2008.

T. Brenac. *Contribution des méthodes de partition centrale à la mise en évidence expérimentale de catégories cognitives.* Rapport outils et méthodes INRETS. INRETS, 2002.

F. Cailliez and J.P. Pagès. *Introduction à l'analyse des données.* Société de Mathématiques Appliquées et de Sciences Humaines, 1976.

R. Campigotto, P. Conde-Céspedes, and J.L. Guillaume. A generalized and adaptive method for community detection. 2013.

S. Chah. *Optimisation linéaire en classification automatique.* PhD thesis, Université Pierre et Marie Curie, LSTA, Paris, France, 1983.

P. Chebotarev. Styding new classes of graph metrics. In F. Nielsen and F. Barbaresco, editors, *Proc. First International Conference, Geometric Science of Information*, number 1, pages 207–214, Paris, France, 2013. Springer-Verlag.

A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Physical Review E*, pages 1– 6, 2004.

P. Conde-Céspedes and J.F. Marcotorchino. Comparison of linear modularization criteria of networks using relational metric. In *45èmes Journées de Statistique, SFdS*, Toulouse, France, May 2013.

P. Conde-Céspedes and F. Marcotorchino. Comparison different modularization criteria using relational metric. In F. Nielsen and F. Barbaresco, editors, *Proc. First International Conference, Geometric Science of Information*, number 1, pages 180–187, Paris, France, 2013. Springer-Verlag.

Caritat A. Marquis de Condorcet. Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix. *Journal of Mathematical Sociology*, 1(1) : 113–120, 1785.

J. Darlay, N. Brauner, and J. Moncel. Dense and sparse graph partition. *Discrete Applied Mathematics*, 160(16-17) :2389–2396, 2012.

F. De Montgolfier, M. Soto, and L. Viennot. Modularité asymptotique de quelques classes de graphes. In Fabien Mathieu and Nicolas Hanusse, editors, *14èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications (AlgoTel)*, pages 1–4, La grande motte, France, 2012.

C. Decaestecker. *Apprentissage en classification conceptuelle incrémentale.* PhD thesis, Université Libre de Bruxelles (Faculté des Sciences), 1992.

M. Delest, J.M. Fedou, and G. Melancon. A quality measure for multi-level community structure. In *Proceedings of the Eighth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, SYNASC '06, pages 63–68, Timisoara, Romania, 2006. IEEE Computer Society.

E. D. Demaine and N. Immorlica. Correlation clustering with partial information. In *In Proceedings of the 6th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX)*, pages 1–13, Princeton, 2003. Springer.

E.D. Demaine, D. Emanuel, A. Fiat, and N. Immorlica. Correlation clustering in general weighted graphs. *Theoretical Computer Science*, 361(2) :172–187, 2006.

W.E. Donath and A.J. Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17(5) :420–425, 1973.

J. Duch and A. Arenas. Community detection in complex networks using extremal optimization. *Physical Review E*, 72 :027104, 2005.

P. Erdős and A. Rényi. On random graphs. I. *Publicationes Mathematicae Debrecen*, 6 : 290–297, 1959.

M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23 : 298–305, 1973.

L. R. Ford and D.R. Fulkerson. Maximal flow through a network. *Canadian Journal of Mathematics*, 8 :399–404, 1956.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5) :75 – 174, 2010.

S. Fortunato and M. Barthelemy. Resolution limit in community detection. In *Proceedings of the National Academy of Sciences of the United States of America*, 2006.

L. C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1) : 35–41, 1977.

L.C. Freeman. Centrality in social networks : Conceptual clarification. *Social Networks*, 1(3) :215–239, 1979.

A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*, 1(1), 2007.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99 (12) :7821–7826, June 2002.

P.M. Gleiser and L. Danon. Community structure in jazz. *Advances in Complex Systems (ACS)*, 06(04) :565–573, 2003.

A.V. Goldberg. Finding a maximum density subgraph. Technical Report UCB/CSD-84-171, EECS Department, University of California, Berkeley, 1984.

M. Gondran and M. Minoux. *Graphes et algorithmes*. Eyrolles, Paris, 3e édition, 1995.

B.H. Good, Y.A. De Montjoye, and A. Clauset. Performance of modularity maximization in practical contexts. *Physical Review E*, 81(4) :046106, 2010.

M. Grötschel and Y. Wakabayashi. A cutting plane algorithm for a clustering problem. *Mathematical Programming Journal*, 45(1) :59–96, August 1989.

G.T. Guilbaud. *Les théories de l'intérêt général et le problème logique de l'agrégation.* Revue économique, 1952.

J.L. Guillaume. *Analyse statistique et modélisation des grands réseaux d'interactions.* These, Université Paris-Diderot - Paris VII, 2004.

R. Guimera and L. Amaral. Functional cartography of complex metabolic networks. *Nature*, 433 :895–900, FEB 2005.

R. Guimera, M. Sales-Pardo, and L. Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70 :art. no. 025101, AUG 2004.

L. Hagen and A.B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE Trans. on CAD of Integrated Circuits and Systems*, (9) :1074–1085, 1992.

F. Harary. *Graph Theory.* Addison-Wesley, Reading, MA, 1969.

J.B. Hiriart-Urruty. Du calcul différentiel au calcul variationnel : un aperçu de l'évolution de p. fermat à nos jours. *version écrite d'un exposé aux journées Fermat, hôtel Assézat,Toulouse,*, 2004.

M. Hoerdt and D. Magoni. *Proceedings of the 11th International Conference on Software, Telecommunications and Computer Networks 257*, 2003.

A. J. Hoffman. On simple linear programming problems. In *Proceedings of Symposia in Pure Mathematics, Vol. VII*, pages 317–327. American Mathematical Society, Providence, R.I., 1963.

A.J. Hoffman and H.W. Wielandt. The variation of the spectrum of a normal matrix. *Duke Math. Journal*, (20) :37–39, 1952.

B. Hughes. *Random Walks and Random Environments : Random Walks Vol 1*. Clarendon Press, March 1995.

N. A. Idrissi. *Contribution à l'Unification de Critères d'Association pour Variables Qualitatives*. PhD thesis, Université Pierre et Marie Curie, LSTA, Paris, France, 2000.

S. Janson and J. Vegelius. The j- index as a measure of association for nominal scale response agreement. *Applied psychological measurement*, 6(1) :111–121, 1982.

J.P. Kemeny. Mathematics without numbers. *Daedalus, The MIT Press*, (4) :577–591, 1959.

M.G. Kendall and A. Stuart. *The Advanced theory of statistics.*, volume 2. Griffin, 1961.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220 :671–680, 1983.

L. Labiod. *Contribution au Formalisme Relationnel de la Classification Croisée de deux Ensembles*. PhD thesis, Université Pierre et Marie Curie, Paris, France, 2008.

L. Labiod, N. Grozavu, and Y. Bennani. Relationship between the modularity criterion and the relational analysis. In *Advanced Information Management and Service (IMS), 2010 6th International Conference on*, pages 229–235, 2010.

A. Lee and I. Streinu. Pebble game algorithms and sparse graphs. *Discrete Mathematics*, 308(8) :1425 – 1437, 2008.

U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing, (A previous version of this paper appeared as Technical Report 149, Max Planck Institute for Biological Cybernetics, 2006)*, 17 :395–416, December 2007.

F. Malliaros and M. Vazirgiannis. Clustering and community detection in directed networks : A survey. *CoRR*, abs/1308.0971, 2013.

S. Mancoridis, B.S. Mitchell, C. Rorres, Y. Chen, and E.R. Gansner. Using automatic clustering to produce high-level system organizations of source code. In *In the IEEE Proceedings of the 1998 International Workshop on Program Understanding (IWPC'98)*, pages 45–52, Ischia, Italy, June 1998. IEEE Computer Society.

F. Marcotorchino. Utilisation des comparaisons par paires en statistique des contingences (partie i). *Publication du Centre Scientifique IBM de Paris, F057, et Cahiers du Séminaire Analyse des Données et Processus Stochastiques Université Libre de Bruxelles*, pages 1–57, 1984a.

F. Marcotorchino. Présentation des critères d'association en analyse des données qualitatives. *Publication AD0185, Université Libre de Bruxelles*, pages 1–57, 1984b.

F. Marcotorchino. Utilisation des comparaisons par paires en statistique des contingences (partie iii). *Etude F-081 du Centre Scientifique IBM de Paris*, pages 1–39, 1985.

F. Marcotorchino. *Liaison Analyse Factorielle-Analyse Relationnelle (I) : "Dualité Burt-Condorcet"*. Etude du Centre Scientifique IBM France, No F142, Paris, 1989.

F. Marcotorchino. *L'analyse Factorielle-Relationnelle (parties 1 et 2)*. Etude du Centre Scientifique IBM France, M06, Paris, 1991.

F. Marcotorchino. *"Dualité Burt-Condorcet :relation entre analyse factorielle des correspondances et analyse relationnelle", dans le livre "Analyse des Correspondances et Techniques Connexes"*. Moreau J., Doudin P.A., Cazes P. Editeurs, Springer-Verlag Berlin, Berlin, 2000.

F. Marcotorchino. *Classifications, Partitionnements, Classements et Ordonnances : Unification de Critères par Modélisation Relationnelle et Approches Spectrales Associées*. 2012.

F. Marcotorchino. Optimal transport, spatial interaction models and related problems, impacts on relational metrics, adaptation to large graphs and networks modularity. *Internal Publication of Thales*, 2013.

F. Marcotorchino and P. Conde-Céspedes. Optimal transport and minimal trade problem, impacts on relational metrics and applications to large graphs and networks modularity. In F. Nielsen and F. Barbaresco, editors, *Proc. First International Conference, Geometric Science of Information*, number 1, pages 169–179, Paris, France, 2013. Springer-Verlag.

F. Marcotorchino and N. El Ayoubi. Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association. *Revue de Statistique Appliquée*, 39(2) : 25–46, 1991.

F. Marcotorchino and P. Michaud. *Optimisation en Analyse ordinale des données*. Masson, Paris, 1979.

F. Marcotorchino and P. Michaud. Agrégation de similarités en classification automatique. *Revue de Statistique Appliquée*, 30(2) :21–44, 1981.

R.S. Michalski and R. Stepp. Learning from observation : Conceptual clustering. In R.S. Michalski, J.G. Carbonell, T. Mitchell, and M. Kaufmann, editors, *Machine Learning : An Artificial Intelligence Approach*, volume 1, chapter 11, pages 331–364. Tioga, 1983.

P. Michaud. Agrégation à la majorité : Hommage à condorcet. *Technical Report du Centre Scientifique IBM France No. F051*, (F051), 1982.

S. Milgram. The small world problem. *Psychology Today*, 1(1) :61–67, 1967.

B.G. Mirkin and L.B. Cherny. Deriving a distance measure between partitions of a finite set. *Automation and Remote Control Journal*, 31(5) :91–98, 1970.

A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proceedings of the 5th ACM/Usenix Internet Measurement Conference (IMC'07)*, San Diego, CA, October 2007.

B. Mohar. The laplacian spectrum of graphs. In *Graph Theory, Combinatorics, and Applications*, volume 2, pages 871–898, New York,USA, 1991. Wiley.

G. Monge. *Mémoire sur la théorie des déblais et des remblais.* Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année, 1781.

J. Moon and L. Moser. On cliques in graphs. *Israel Journal of Mathematics*, 3 :23–28, 1965.

M. E. J. Newman. *The New Palgrave Encyclopedia of Economics*, chapter Mathematics of networks. Palgrave Macmillan, Basingstoke, 2 edition, 2008.

M.E.J. Newman. Detecting community structure in networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2) :321–330, 2004a.

M.E.J. Newman. Fast algorithm for detecting community structure in networks. *Physical Review E*, 69, 2004b.

M.E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23) :8577–8582, June 2006a.

M.E.J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical Review E*, 74 :36104, May 2006b.

M.E.J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E.*, 69(2), 2004.

J.W. Owsiński and S. Zadrożny. Clustering for ordinal data : a linear programming formulation. *Control and Cybernetics*, 15(2) :183–193, 1986.

P. Pons and M. Latapy. Computing communities in large networks using random walks. *Journal of Graph Algortihms and Applications*, 10(2) :191–218, 2006.

A. Pothen, H. D. Simon, and K.P. Liou. Partitioning sparse matrices with eigenvectors of graphs. *SIAM Journal on Matrix Analysis and Applications*, 11 :430–452, May 1990.

F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. *Proceedings of the National Academy of Sciences*, 101(9) : 2658, 2004.

W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336) :846–850, 1971.

S. Régnier. *Sur quelques aspects mathématiques des problèmes de classification automatique*, volume 4. I.C.C. Bulletin, 1966.

J. Reichardt and S. Bornholdt. Detecting fuzzy community structures in complex networks with a potts model. *Physical Review Letters*, 93(21) :218701, November 2004.

J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E.*, 74(1), July 2006.

S.A. Rice. *Quantitative Methods in Politics.* Borzoi Books. A.A. Knopf, 1928.

G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31 :581–603, December 1966.

G. Saporta. About maximal association criteria in linear analysis and in clustering. In Hans H. Bock, editor, *Classification and Related Methods of Data Analysis : Proceedings of the First Conference of the International Federation of Classification Societies (IFCS)*, Aachen, Germany, 1988. North Holland.

G. Saporta and G. Youness. Comparing two partitions : Some proposals and experiments. In *Proc. Computational Statistics*, Berlin, Germany, 2002.

S.E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1) :27–64, 2007.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8) :888–901, 2000.

É. Stemmelen. *Tableaux d'échanges, description et prévisions.* Number 28 in Cahiers du Bureau universitaire de recherche opérationnelle. Série Recherche. Université Pierre-et-Marie-Curie, Paris, France, 1977.

I. Streinu and L. Theran. Sparse hypergraphs and pebble game algorithms. *European Journal of Combinatorics*, 30(8) :1944 – 1964, 2009.

E. Viennet. Recherche de communautés dans les grands réseaux sociaux. *Revue des Nouvelles Technologies de l'Information (RNTI-A3)*, pages 145–160, July 2009.

Y. Wakabayashi. The complexity of computing medians of relations, 1998.

K. Wakita and T. Tsurumi. Finding community structure in mega-scale social networks. *CoRR*, abs/cs/0702048, 2007.

J.H. Ward. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301) :236–244, 1963.

Y.-C. Wei and C.-K. Cheng. Ratio cut partitioning for hierarchical designs. *IEEE Trans. on CAD of Integrated Circuits and Systems*, pages 911–921, 1991.

Y.C. Wei and C.K. Cheng. Towards efficient hierarchical designs by ratio cut partitioning. *IEEE International Conference on Computer-Aided Design*, pages 298–301, 1989.

R.S. Weis and E. Jacobson. A method for the analysis of complex organizations. *American Sociological Review*, 20 :661–668, 1955.

A. G. Wilson. The Use of Entropy Maximising Models, in the Theory of Trip Distribution, Mode Split and Route Split. *Journal of Transport Economics and Policy*, 3(1) :108–126, 1969.

A.G. Wilson. *A Statistical Theory of Spatial Distribution Models*, volume 1 of *Transportation Review.* 1967.

A.G. Wilson. *Entropy in urban and regional modelling.* Monographs in spatial and environmental systems analysis. Pion, London, 1970.

F.Y. Wu. The potts model. *Reviews of Modern Physics*, 54(1) :235–268, January 1982.

J. Yang and J. Leskovec. Defining and evaluating network communities based on ground-truth. In *International Conference on Data Mining*, volume abs/1205.6233, pages 745–754. IEEE Computer Society, 2012.

G. Youness and G. Saporta. Some measures of agreement between close partitions. *Student*, 5 :1–12, 2004.

W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33 :452–473, 1977.

C.T. Zahn. Approximating symmetric relations by equivalence relations. *SIAM Journal on Applied Mathematics*, 12 :840–847, 1964.

H. Zhou. Distance, dissimilarity index, and network community structure. *Phys. Rev. E*, 67(6) :061901, June 2003.

X.J. Zhou and T.S. Dillion. A statistical-heuristic feature selection criterion for decision tree induction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13 : 834–841, August 1991.