Modularisation et Recherche de Communautés dans les réseaux complexes par Unification Relationnelle

Patricia Conde-Céspedes*, Jean-François Marcotorchino**,

*Laboratoire de Statistique théorique et Appliquée 4 place Jussieu 75252 Paris cedex 05
patricia.conde_cespedes@upmc.fr

**Thales Communications et Sécurité, TCS, 4 Avenue des Louvresses, 92230 Gennevilliers
jeanfrancois.marcotorchino@thalesgroup.com,

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 via 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, le critère de Mancoridis-Gansner, le critère de Zahn-Condorcet, etc. Cette approche permet de faciliter leur compréhension, et d'interpréter plus clairement leurs finalités en y associant la preuve de leur utilité dans certains contextes pratiques.

1 Introduction à la recherche de communautés

De nos jours, l'étude des réseaux est devenue à la mode grâce essentiellement à l'utilisation massive des "Réseaux Sociaux" comme FaceBook, Twitter, Linkedln, etc. Ces réseaux génèrent eux-mêmes des quantités considérables de données, ce qui contribue à accroître leur complexité. Cet afflux massif de données provient principalement d'Internet, des télécommunications et des technologies numériques en général. Toutes les disciplines, notamment la physique, la biologie et les sciences sociales ont été confrontées à cette arrivée d'énormes quantités de données à gérer. C'est donc la manipulation de cette masse de données latentes qui a rajouté un niveau supplémentaire de complexité à l'exploration de tels réseaux.

Les réseaux sont des objets très variés que l'on rencontre dans de nombreuses configurations, applicatives en particulier partout où l'on peut (ou doit) définir des connexions entre entités. Tous ces réseaux ont en commun de pouvoir être décrits par un graphe où les "sommets" représentent les différents éléments en interaction et les "arêtes" représentent l'influence d'un sommet sur un autre. Par exemple : en biologie moléculaire (l'interaction entre protéines), en informatique (les sites internet et les liens entre eux ou les réseaux client-serveur), en physique (les circuits électriques), etc.

La quantité de données disponibles (existantes ou collectées) et la taille des réseaux complexes rend difficile, voire impossible, leur analyse directe. Il devient alors indispensable de pouvoir les "partitionner" en sous-ensembles gérables et faciles à comprendre. Partitionner un graphe n'est autre que faire du "clustering" de graphes où chaque sous-ensemble est une classe contenant des éléments plus densément connectés entre eux qu'avec le reste du réseau. Plusieurs auteurs appellent ce processus de partitionnement d'un graphe "la modularisation" car il évoque la notion de "module". En sociologie, on parle plutôt de "recherche de communautés". Partitionner un graphe peut être utile dans de nombreux domaines : en cybersécurité, la recherche des zones de connexion les plus denses en vue d'isoler les zones vulnérables ; en biologie, la détection des individus "super propagateurs" à traiter en priorité dans le cadre d'une campagne de vaccination par exemple.

L'obtention de ce résultat nécessite que l'on dispose d'un "bon" critère qualifiant le processus de partitionnement. C'est ce choix de critères possibles qu'il importe de justifier et c'est l'ensemble des problématiques associées à leurs propriétés que nous développerons.

Or, qui dit "graphe" sous-entend "relations" ou liens entre sommets du graphe, et qui dit "relations" préfigure leur analyse au moyen de techniques appropriées comme l'est l'Analyse Relationnelle Mathématique (ARM) [1]. Dans cet article nous allons privilégier une structuration logique à l'expression d'un nombre important de critères de modularisation de graphes. L'écriture relationnelle constituera un outil de comparaison des critères, permettant ainsi de savoir quel critère utiliser selon le besoin et la définition de la notion de "communauté" choisie.

La section 2 présente un bref récapitulatif de l'ARM au niveau des définitions de base et des notations standard ; la section 3 retrace l'importance de la recherche de communautés dans les grands réseaux ; la section 4 présente les propriétés des structures relationnelles ; la section 5 expose différents critères de modularisation ; la section 6 présente deux exemples d'application connu sous l'angle de plusieurs critères. Enfin, les perspectives et conclusions sont exposées dans la section 7.

2 L'Analyse Relationnelle

2.1 Définitions et principes

L'Analyse Relationnelle et Mathématique des Données, 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 vu théorique, l'ARM étudie les relations binaires et particulièrement les problèmes qui ont trait à leurs mesures d'associations, leurs agrégations et à la

1. Dans ce document nous utiliserons l'acronyme ARM pour faire référence à l'Analyse Relationnelle Mathématique.

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 la 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 contemporain des "Réseaux Sociaux".

L'idée originale de Condorcet (voir [Caritat A. (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 [2], en fait, elle préfigurait l'approche relationnelle associée, qui, elle, a permis de reformaliser 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 [Caritat A. (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 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)] et 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.

Les bases fondamentales de l'ARM reposent sur la définition de la notion de *Relation Binaire*. Voici une définition générale d'une relation binaire croisant le même ensemble [3] :

Définition 1 : "Relation Binaire sur un ensemble" Une relation binaire $\mathcal{R}$ sur un ensemble E est un sous-ensemble du produit cartésien $E \times E$, noté $G(\mathcal{R})$.

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, 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..."

2. Voir les travaux de [Guilbaud (1952)], premier texte français où l'on introduit les travaux de Condorcet et en particulier le fameux "Effet Condorcet"

3. Cette définition peut s'étendre à une relation binaire entre deux ensembles, ce que l'on appelle *Relation Binaire bipartite.* Ainsi, une relation binaire bipartite $\mathcal{R}$ croisant deux ensembles E et F (ou de E vers F) est un sous-ensemble du produit cartésien $E \times F$.

- "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 : "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 $u \not\mathcal{R} v$.

2.2 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 **C** et de taille $(N \times N)$, est définie de la façon suivante :

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

Comme nous allons le voir par la suite, la matrice **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 : $\not\mathcal{R}$:

$$\bar{c}_{ii'} = \begin{cases} 1 & \text{si } i\not\mathcal{R}i' \forall (i, i') \in V \times V \\ 0 & \text{sinon} \end{cases} \tag{2}$$

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 plus ou moins de propriétés élémentaires). Le Tableau 1 montre une liste non exhaustive de propriétés qu'une relation binaire $\mathcal{R}$ pourrait vérifier.

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* [4].

4. 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.

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	$i\mathcal{R}i' \wedge i'\mathcal{R}i'' \Rightarrow i\mathcal{R}i'' \qquad \forall (i,i',i'')$	$c_{ii'} + c_{i'i''} - c_{ii''} \leq 1$

TAB. 1 – *Principales propriétés vérifiées par une relation binaire.*

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.

2.3 Les codages

Une variable qualitative nominale à p modalités décrivant N objets [5] d'un ensemble V n'est rien d'autre qu'une relation d'équivalence définie sur V. 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 **K**, de taille $(N \times p)$ dont l'élément k_{ij} est une variable de présence-absence et il est donné par :

$$k_{ij} = \begin{cases} 1 & \text{si l'objet } i \text{ possède la modalité } j \\ 0 & \text{sinon} \end{cases} \tag{3}$$

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 **C** associée. Ainsi, le terme général de $c_{ii'}$ du tableau de Condorcet vaut dans ce cas-là :

$$c_{ii'} = \begin{cases} 1 & \text{si les objets } i \text{ et } i' \text{ possèdent la même modalité.} \\ 0 & \text{sinon} \end{cases} \tag{4}$$

A titre d'exemple nous illustrons sur la FIG.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).

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

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

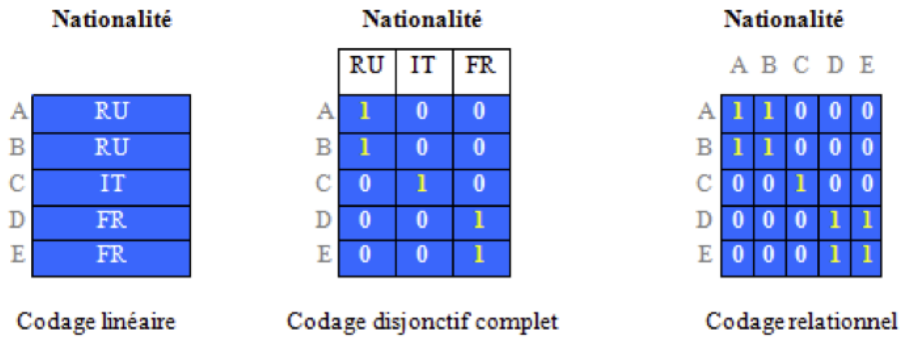

FIG. 1 – *3 codages possibles d'une variable catégorielle (relation d'équivalence).*

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.

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 suit :

$$\hat{c}_{ii'} = \frac{2c_{ii'}}{c_{i.} + c_{.i'}}, \forall i, i' \in V \times V \qquad \text{avec} \qquad c_{i.} = \sum_{i'=1}^{N} c_{ii'}, \forall i \in V \tag{5}$$

Dans le cas où $\mathcal{R}$ représente une relation d'équivalence, des simplifications importantes se produisent dans la formule (5). 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 (5) se simplifie en :

$$\hat{c}_{ii'} = \frac{c_{ii'}}{c_{i.}} \tag{6}$$

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], la somme des carrés de ses termes est égal au nombre total de modalités p (voir équation (8)), ce qui confirme bien que la connaissance du nombre de modalités est contenue implicitement dans le tableau.

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

$$\bar{\hat{c}}_{ii'} = \frac{\hat{c}_{ii} + \hat{c}_{i'i'}}{2} - \hat{c}_{ii'} \tag{7}$$

Cette matrice peut s'interpréter par un écart à la situation d'*auto similarité maximale* [7].
Autre résultat intéressant valide pour une relation d'équivalence est la propriété relationnelle suivante :

$$\sum_{i=1}^{N}\sum_{i'=1}^{N} \hat{c}^2_{ii'} = \sum_{i=1}^{N}\sum_{i'=1}^{N} \frac{c_{ii'}}{c_{i.}c_{.i'}} = p \tag{8}$$

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

Conclusion : Il existe un lien direct entre l'ARM et la théorie des graphes, selon le principe qu' "*un graphe est considéré comme une structure mathématique servant à modéliser les relations binaires entre objets d'un même ensemble*". Ainsi, la matrice d'adjacence du graphe équivaut à la matrice de Condorcet décrivant cette relation binaire.

3 Définition de la notion de *communauté*

Comme nous l'avions déjà signalé dans la section précédente, l'ARM est un outil adapté à la modélisation des graphes (par exemple, un graphe non-orienté et non-réflexif est entièrement représenté par sa matrice d'adjacence qui n'est rien d'autre qu'une matrice de Condorcet représentant une relation binaire symétrique). Ceci justifie que soient identifiées au sein de l'analyse des grands graphes et des grands réseaux, des thématiques qui sont en ligne directe avec les méthodologies et problèmes de classification déjà traités par l'ARM. Dans cet article nous allons présenter différents critères de modularisation ainsi que leurs principales propriétés. Mais auparavant, nous allons revenir sur la définition de communauté car chaque critère dépend fortement de ce que l'on sous-entend par communauté, nous utiliserons de façon indistincte les mots : classe, module, cluster pour faire référence à une communauté.

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. Benzecri en "Analyse Factorielle des Correspondances" en 1973, (voir [Benzécri et Collaborateurs (1973a)] et [Benzécri et Collaborateurs (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)].

Les réseaux réels ne sont pas des graphes aléatoires car ils présentent des hétérogénéités importantes. Ils révèlent un haut degré d'ordre et d'organisation. En plus, la distribution d'arêtes est localement hétérogène. Étant très dense à l'intérieur de certains groupes de sommets et rare entre ces groupes.

La formulation mathématique du problème de détection des communautés est apparentée au partitionnement (ou clustering) de graphes. Ce problème n'est pas bien défini car il n'existe pas une définition universelle du concept de *communauté*. En effet, différents critères de clustering de graphes ont été proposés au fil du temps pour modéliser des phénomènes issus de domaines divers chacun avec une définition légèrement différente du terme "*communauté*". Voici quelques définitions du terme "communauté" trouvées dans la littérature :

- Groupes de sommets densément connectés, avec quelques rares connexions entre eux.
- 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 communautés 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 : $M >> N$ (où M est le nombre d'arêtes et N le nombre de sommets). Ainsi, par exemple, dans un *arbre* [8] ou un graphe en forme de grille (*lattice ou grid graph en anglais*) la notion de module n'a pas de sens, bien évidement il n'existe pas dans ce cas, de groupes de sommets qui soient remarquablement plus connectés que d'autres.

La détection de communautés a de multiples utilités :

- Fréquemment les sommets densément connectés partagent une propriété commune que l'on peut mettre en évidence majoritairement a posteriori (souvent au niveau interprétatif). Par exemple, dans le cas des 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 compressée (résumée) en un seul "méta-sommet" (le meilleur représentant de la communauté [9]), permettant une analyse du réseau à un niveau plus grossier, et une focalisation sur la structure de niveau supérieur. Cette approche peut

8. Nous rappelons qu'un arbre est un graphe minimalement connecté, i.e. $M = N - 1$
9. Ceci renvoie aux notions de centralité (de différentes sortes) que nous avons explicitées précédemment

servir pour visualiser un réseau de façon simplifiée, en perdant le minimum d'information.

– Identifier les modules d'un graphe permet d'interpréter la fonction de chaque sommet dans le module. Ainsi, un sommet qui possède une position centrale (i.e. il est adjacent 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 périphérie de la classe jouent un rôle important de médiation et d'échange avec les autres communautés (voir [Barabasi et 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 [10].
– Dans un réseau informatique, par exemple, la détection de modules peut servir à savoir quelle est la meilleure façon de répartir les tâches au niveau des processeurs afin de minimiser les communications entre eux et permettre ainsi une meilleure performance de calcul.

Dans la section suivante nous allons proposer une liste non exhaustive de critères de modularisation, chacun ayant sa propre définition relative à la notion de *module*.

4 Propriétés des structures relationnelles

Modulariser ou partitionner un graphe signifie donc chercher les communautés qui le composent, quelle que 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"*.

Plusieurs critères de modularisation ont été proposés dans la littérature. Dans cette sous-section nous présenterons à la fois ces critères connus dans la littérature, mais présentés selon leur écriture relationnelle, mais également des critères nouveaux ou nettement moins connus. A cette fin, nous donnerons quelques définitions utiles en Analyse Relationnelle : la matrice d'adjacence d'un graphe et la matrice de la relation d'équivalence cherchée.

– **Matrice d'adjacence A :** Étant donné un graphe non-orienté (G, V) à $|V| = N$ sommets et M arêtes. La matrice d'adjacence **A** de V est une matrice carrée d'ordre N dont les éléments sont définis de la façon suivante :

$$a_{ii'} = \begin{cases} 1 & \text{s'il existe une arête entre } i \text{ et } i' \\ 0 & \text{sinon} \end{cases} \tag{9}$$

– **Matrice de la relation d'équivalence cherchée X :** Il s'agit d'une matrice carrée d'ordre N dont les éléments sont définis de la façon suivante :

$$x_{ii'} = \begin{cases} 1 & \text{si } i \text{ et } i' \text{ appartiennent à la même classe} \\ 0 & \text{sinon} \end{cases} \tag{10}$$

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

Modulariser un graphe revient à trouver **X** (une partition sur l'ensemble de sommets) la *plus proche possible* de **A**. Pour évaluer cette *proximité* il faut définir une fonction de **A** et de **X** qui mesure soit une distance (écart), soit une similarité ou une proximité entre ces deux matrices. Cette fonction est à minimiser dans le cas d'un écart et à maximiser dans le cas d'une proximité, elle sera notée $F(\mathbf{A}, \mathbf{X})$. Ainsi tout critère sera écrit en notation relationnelle de la façon suivante :

$$\underset{X}{Max} \text{ ou } \underset{X}{Min} F(A, X) \tag{11}$$

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

$$\begin{array}{rll} x_{ii'} \in \{0,1\} & & \text{Binarité} \\ x_{ii} = 1 & \forall i & \text{Réflexivité} \\ x_{ii'} - x_{i'i} = 0 & \forall (i,i') & \text{Symétrie} \\ x_{ii'} + x_{i'i''} - x_{ii''} \leq 1 & \forall (i,i',i'') & \text{Transitivité} \end{array} \tag{12}$$

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

Pour pouvoir comparer les différentes fonctions ou critères de partitionnement il est nécessaire de les écrire avec la même notation standard. C'est la notation relationnelle qui nous permettra d'effectuer cette comparaison. D'autre part, chaque critère sera jugé en fonction des propriétés qu'il vérifie. Le Tableau 2 montre une liste non exhaustive de quelques propriétés que doit vérifier un bon critère :

Propriété	**Écriture Relationnelle de chaque propriété**
Linéarité	$F(X) = \sum_{i=1}^{n} \sum_{i'=1}^{n} a_{ii'} x_{ii'} + constante$
Séparabilité	$F(X) = \sum_{i=1}^{n} \sum_{i'=1}^{n} a_{ii'} \psi(x_{ii'}) + Constante$

TAB. 2 – *Propriétés des critères de partitionnement.*

La **propriété de linéarité** implique que le critère est linéaire par rapport à la variable inconnue **X**. La **propriété de séparabilité** implique que le critére soit séparable au sens : "variables-données".

5 Critères de modularisation et leur écriture relationnelle

5.1 Critères linéaires en X

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

Ce critère a été introduit, quant à ses concepts axiomatiques sous-jacents, dans le cadre de la Théorie des votes, des suffrages et des élections par Antoine Caritat, Marquis de Condorcet (cf. [Caritat A. (1785)]). Sa formalisation Relationnelle Mathématique a été redéfinie dans [Marcotorchino et Michaud (1979)].

Soit J l'ensemble de $M = |J|$ variables catégorielles $J = \{V^1, V^2, ..., V^M\}$ qui décrivent un ensemble I de N objets aussi appelés éléments. Chaque variable V^k possède p_k modalités, et $P = \sum_{k=1}^{M} p_k$ est le nombre total de modalités [11]. Comme chaque objet peut appartenir à une seule catégorie de chaque descripteur, chaque variable définit une relation d'équivalence sur l'ensemble des objets.

Notons $\mathbf{C}^k$ la matrice relationnelle de Condorcet de la variable V^k. Dans le but de trouver une relation d'équivalence **X** qui résume au mieux les M classifications définies sur les N objets, la fonction à maximiser proposée par Condorcet en 1785 s'écrit :

$$F_C(X) = \sum_{k=1}^{M} \left[\sum_{i=1}^{N} \sum_{i'=1}^{N} (c^k_{ii'} x_{ii'} + \bar{c}^k_{ii'} \bar{x}_{ii'}) \right] \tag{13}$$

Où $\bar{\mathbf{X}}$ et $\bar{\mathbf{C}}^k$ sont les relations complémentaires à **X** et $\mathbf{C}^k$ respectivement. D'autre part **X** doit vérifier les contraintes d'une relation d'équivalence, vues à la formule (12).

Mais l'équation (13) peut aussi s'écrire sous une forme équivalente (dite du "support de voix" dans la terminologie de Condorcet). En effet, en utilisant successivement la propriété d'additivité relationnelle associée à la linéarité du critère qui permet d'obtenir le tableau de Condorcet proprement dit, à savoir : $c_{ii'} = \sum_{k=1}^{M} c^k_{ii'} \forall i, i'$, on obtient :

$$F_C(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left[\left(\sum_{k=1}^{M} c^k_{ii'} \right) x_{ii'} + \left(\sum_{k=1}^{M} \bar{c}^k_{ii'} \right) \bar{x}_{ii'}) \right] = \sum_{i=1}^{N} \sum_{i'=1}^{N} [(c_{ii'} x_{ii'} + \bar{c}_{ii'} \bar{x}_{ii'})] \tag{14}$$

Dans le membre droit de l'expression (14) la sommation en k sur le nombre de variables disparaît. Le "vrai" critère de Condorcet est celui correspondant à la forme de droite de la formule (14), néanmoins l'avantage de l'écriture (13) c'est qu'elle permet d'ajuster le critère de Condorcet à la valeur de M car k varie de 1 à M. Comme nous allons le voir le cas $M = 1$

11. Les objets peuvent être, par exemple, un ensemble de voitures caractérisées par certains attributs : comme la couleur {rouge,blanc,bleu} ; la marque {Renault, Toyota, Mitsubishi}.

est également très intéressant et donne de la généralité au critère de Condorcet écrit sous cette forme.

Le critère de Condorcet s'écrit également comme une fonction à minimiser. Cette fonction, nommée forme "duale" du critère de Condorcet, a pour expression :

$$F_C(\bar{X}) = \sum_{k=1}^{M} \left[\sum_{i=1}^{N} \sum_{i'=1}^{N} (\bar{c}_{ii'}^{k} x_{ii'} + c_{ii'}^{k} \bar{x}_{ii'}) \right] \tag{15}$$

Revenons maintenant au problème de modularisation des graphes, en nous appuyant sur l'approche proposé par [Zahn (1964)]. Ce dernier est le premier auteur à avoir posé le problème de trouver une relation d'équivalence **X** qui approxime au mieux une Relation Binaire symétrique donnée, R, définie sur les paires des éléments d'un ensemble fini V. Pour ce faire, il définit une fonction de *distance* (en l'occurrence la distance de la "différence symétrique" entre relations) notée $F_Z(R, X)$, à minimiser, entre ces deux relations :

$$F_Z(R, X) = |X - R| + |R - X| \tag{16}$$

Où l'on a utilisé la notation $A - B \equiv A \cap \bar{B}$.

Comme la seule propriété qu'il impose à la relation R est la symétrie, celle-ci peut être facilement modelisée à partir d'un graphe $G(V, E)$ non-orienté et sans boucles où les sommets seraient les éléments de l'ensemble V et les relations entre eux seraient les arêtes E. Dans l'expression (16) la relation R est un sous-ensemble du produit cartésien $V \times V$, c'est l'ensemble de paires liées par la relation R. Ainsi, $-R$ représente, quant à elle, le complémentaire de R dans $V \times V$, c'est-à-dire l'ensemble de paires $(v_i, v_i') \in V \times V$ qui ne sont pas en relation dans R. En notations relationnelles la fonction (16) s'écrit alors :

$$F_Z(X) = \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} (\bar{a}_{ii'} x_{ii'} + a_{ii'} \bar{x}_{ii'}) \tag{17}$$

Où $a_{ii'}$ est le terme général de la matrice d'adjacence **A** du graphe induit par la relation R. La variable $x_{ii'}$ est la variable relationnelle de la relation cherchée **X**, qui doit vérifier les contraintes d'une relation d'équivalence, équation (12).

L'expression (17) n'est rien d'autre que la fonction de Condorcet dans sa formulation à minimiser, équation (15), mais pour une valeur de $k = M = 1$. Comme la constante $\frac{1}{2}$ ne modifie pas la valeur optimale du critère, nous l'omettrons dans l'expression (17) et c'est cette expression que nous nommerons le "critère de Zahn-Condorcet".

Le critère de Zahn-Condorcet vérifie les propriétés de linéarité et séparabilité. D'autre part, pour tout graphe connecté, non pondéré et non orienté la solution optimale possède la caractéristique suivante : pour toute classe d'équivalence obtenue le nombre d'arêtes intra-classe est supérieur ou égal à 50% 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.

Cette dernière caractéristique est une conséquence de la condition dite de "majorité par paires" de Condorcet. En effet, toute classe possédant moins de 50% d'arêtes intra-classe possibles aura un effet négatif sur la valeur finale du critère.

Dans certains cas cette condition de majorité absolue intra-classe 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 :

$$F_{Z_{OZ}}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}((1-\alpha)a_{ii'}x_{ii'} + \alpha\bar{a}_{ii'}\bar{x}_{ii'}) \qquad 0 < \alpha < 1 \tag{18}$$

Bien évidemment sous les contraintes d'une relation d'équivalence, équation (12).
Cette dernière écriture constitue une généralisation du critère de Condorcet, déjà proposée par [Owsiński et Zadrożny (1986)], ainsi ce critère est linéaire et séparable, . 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. Plus α est proche de 1 plus dense sont les connexions intra-classe. Si $\alpha = 0.5$ on obtient la fonction de Zahn-Condorcet.

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

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

Les auteurs ont proposé ce critère dans [Newman et Girvan (2004)] comme une mesure de la force de la structure communautaire d'un graphe. Ce critère cherche à maximiser l'écart entre le graphe original et sa version aléatoire correspondante, car un graphe aléatoire ne possède pas de structure communautaire. Plus spécifiquement, cette fonction cherche à maximiser la différence entre le nombre d'arêtes intra-classe et l'espérance de cette valeur dans un graphe où les arêtes sont placées de façon aléatoire entre les sommets.

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 et Girvan (2004)], plus connue aujourd'hui comme "fonction de modularité", est donnée par l'expression suivante :

$$F_{NG}(\mathbf{e}) = \sum_{j}^{\kappa}(e_{jj} - (\sum_{j'}^{\kappa} e_{jj'})^2) = Tr(\mathbf{e}) - (\sum_{j}^{\kappa}\sum_{j'}^{\kappa}[\mathbf{e}^2]_{jj'}) \tag{19}$$

Où $[\mathbf{e}^2]_{jj'}$ représente l'élément jj' de la matrice $\mathbf{e}^2$, donc le deuxième terme de (19) n'est autre que la somme des éléments de la matrice $\mathbf{e}^2$. Dans l'expression (19) 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 de 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 d'arêtes réelle 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 :

$$F_{NG}(\kappa) = \sum_{j=1}^{\kappa} \left(\frac{|E(C_j)|}{M} - \left(\frac{\sum_{v \in C_j} d_v}{2M} \right)^2 \right) \tag{20}$$

où M est le nombre total d'arêtes (ou la somme des poids dans le cas d'un graphe pondéré), $E(Cj)$ est l'ensemble d'arêtes intra-classe de la classe C_j et d_v est le degré du sommet v.

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

$$F_{NG}(X) = \frac{1}{2M} \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(a_{ii'} - \frac{a_{i.} a_{.i'}}{2M} \right) x_{ii'} \tag{21}$$

avec **X** vérifiant également les contraintes d'une relation d'équivalence, équation (12).

L'écriture relationnelle de l'expression (21) permet de mettre en évidence les propriétés du critère de Newman-Girvan :

- Il est linéaire et séparable.
- L'écriture relationnelle fait disparaitre le nombre de classes de la partition cherchée, ainsi on n'est pas obligé de le fixer à l'avance.
- Il s'agit d'un modèle nul, il vaut 0 pour la *partition grossière* où tous les objets sont dans la même classe. Cela signifie que pour un graphe complet, la partition optimale possède une valeur de modularité nulle.

D'autres propriétés intéressantes de ce critère sont discutées dans [Brandes et al. (2008)]. Notamment le propriété de *non-locality* (non-localité), qui met en évidence que la modularité, 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 sous cette forme relationnelle que le terme entre parenthèses de l'équation (21) 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)].

5.1.3 Le critère de "Correlation clustering" de Demaine et Immorlica (2002)

Le problème dit de "correlation clustering" fut posé initialement par [Bansal et al. (2002)] : "Étant donné un graphe $G = (V, E)$ 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.

Il s'agit donc, de modulariser un graphe pondéré avec des poids réels, i.e. $w_{ii'} \in \mathbb{R}$ où les grands poids positifs représentent une forte corrélation entre les points extrêmes alors que les grands poids négatifs représentent une forte répulsion entre les points extrêmes, et les poids à valeur absolue proche de zéro représentent peu d'information. Plus tard, [Demaine et Immorlica (2003)] ont formulé la fonction économique visant à trouver une solution optimale au problème. En notations relationnelles ce critère s'écrit dans sa version à maximiser :

$$F_D(X) = \frac{1}{2}\left(\sum_{i,i'} w^{+}_{ii'} x_{ii'} + \sum_{i,i'} w^{-}_{ii'} \bar{x}_{ii'}\right) \tag{22}$$

ou dans sa version à minimiser :

$$F_D(\bar{X}) = \frac{1}{2}\left(\sum_{i,i'} w^{+}_{ii'} \bar{x}_{ii'} + \sum_{i,i'} w^{-}_{ii'} x_{ii'}\right) \tag{23}$$

Avec $w^{+}_{ii'} = w_{ii'} 1_{(w_{ii'}>0)}$ et $w^{-}_{ii'} = |w_{ii'}| 1_{(w_{ii'}<0)}$ (où 1_u représente la fonction indicatrice de l'ensemble u).

Bien évidement les critères (22) et (23) doivent vérifier les contraintes (12).

Il est facile de remarquer que l'expression (22) n'est autre que le critère de Condorcet énoncé dans (13) avec des poids réels : $c_{ii'} \equiv w^{+}_{ii'}$ et $\bar{c}_{ii'} \equiv w^{-}_{ii'}$. De façon analogue, l'expression (23) est une formulation très voisine du critère Dual de Condorcet, expression (15) pour un graphe pondéré avec simplement des poids réels au lieu des $c_{ii'}$. De ce fait ce critère vérifie les mêmes propriétés que le critère de Condorcet, il est linéaire et séparable.

5.1.4 Le critère de Condorcet pondéré en A de Marcotorchino (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 :

$$F_{CPond}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(\hat{a}_{ii'} x_{ii'} + \bar{\hat{a}}_{ii'} \bar{x}_{ii'}\right) \tag{24}$$

où $\hat{a}_{ii'}$ et $\bar{\hat{a}}_{ii'}$ sont respectivement définis au travers des équations (5) et (7).

Pour garantir que l'optimisation du critère (24) permet d'obtenir une partition, **X** doit, là encore vérifier les contraintes d'une relation d'équivalence, énoncées dans (12).

Compte tenu de la définition de $\bar{\hat{\mathbf{A}}}$ (équation (7)), maximiser l'expression (24) revient à maximiser l'expression :

$$F_{CPond}(X) = \sum_{i=1}^{N} \sum_{i'=1}^{N} \left(2\hat{a}_{ii'} - \frac{\hat{a}_{ii} + \hat{a}_{i'i'}}{2} \right) x_{ii'} \tag{25}$$

L'écriture relationnelle (25) nous montre que ce critère est linéaire par rapport à l'inconnue **X** et séparable. Il 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.

5.2 Les critères séparables de fonctions non-linéaires de X

5.2.1 Le critère de Mancoridis-Gansner (1998)

La fonction à optimiser de [Mancoridis et al. (1998)] est issue du domaine dit de "cluster-programming", elle cherche à la fois à maximiser les connexions intra-classe et à minimiser les connexions inter-classes.

Le but des auteurs, à l'origine, était de trouver un modèle d'optimisation visant à récupérer automatiquement la structure modulaire d'un programme informatique à partir de son code source. Le but sous-jacent était de comprendre la structure des composantes du programme [12]. Cette décomposition permet aux programmeurs de faire face à la problématique de l'interpénétration des lignes de code au travers de l'identification de procédures connectées en "modules", plus indépendantes les unes des autres. Le programme informatique en question est modélisé par un graphe de la façon suivante : ses sommets constituent les composantes du système : les classes, les variables, les macros et les structures de données [13] ; tandis que ses arêtes constituent les relations (fonctions) courantes comme l'importation, l'exportation, l'héritage ou l'appel à une procédure [14].

Le critère de S. Mancoridis et Y. Gansner se base sur un compromis entre deux mesures : l'inter et l'intra-connectivité :

1. **Intra-connectivité** : cette mesure représente la fraction du nombre d'arcs existants dans la classe j par rapport au nombre maximal d'arcs qui peuvent exister [15], soit N_j^2, où N_j est l'effectif de la classe j. L'expression de l'intra-connectivité A_j de la classe j à N_j composantes et m_j arêtes intra-classe est :

12. Surtout pour les programmes les plus utilisés, qui contiennent plusieurs centaines de milliers de lignes de code encapsulées dans plusieurs modules, cette tâche devient alors indispensable.

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

14. Du point de vue informatique.

15. Les auteurs ont considéré un graphe orienté avec des boucles au moment de définir cette quantité, donc le nombre maximal d'arêtes est la taille de la classe au carré. Dans [Delest et al. (2006)] les auteurs ont adapté le modèle de Mancoridis-Gansner pour l'adapter à un graphe complet.

$$A_j = \frac{m_j}{N_j^2} \tag{26}$$

Ainsi l'intra-connectivité est un ratio, sa valeur est donc comprise entre 0 et 1. Une valeur d'intra-connectivité proche de l'unité indique que les composantes à l'intérieur de la classe sont fortement connectées.

2. **Inter-connectivité** : c'est une mesure de connectivité entre deux classes différentes. Il s'agit de 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 [16], soit $2N_jN_{j'}$. 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-classe est donnée par :

$$E_{jj'} = \begin{cases} 0 & \text{si } j = j' \\ \frac{\epsilon_{jj'}}{2N_jN_{j'}} & \text{si } j \neq j' \end{cases} \tag{27}$$

une valeur d'inter-connectivité faible indique que les groupes sont indépendants, dans une large mesure. L'inter-connectivité est un indice qui varie entre 0 et 1. Dans le cas idéal $E_{jj'}$ est nulle, il n'existe donc aucun lien entre les classes j et j'.

La fonction objectif de [Mancoridis et al. (1998)] 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 :

$$F_{MG} = \frac{1}{\kappa}\sum_{j=1}^{\kappa} A_j - \frac{1}{\frac{\kappa(\kappa-1)}{2}}\sum_{j,j'=1}^{\kappa} E_{jj'} \quad \text{si } \kappa > 1 \tag{28}$$

Le premier terme de l'équation (28) 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 $\frac{\kappa(\kappa-1)}{2}$. Le signe moins devant le deuxième terme permet de maximiser l'opposé de l'interconnectivité 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).

Compte tenu de la définition de la variable relationnelle **X**, équation (10), l'écriture relationnelle de l'équation (28) est la suivante :

$$F_{MG}(X) = \frac{1}{\kappa}\sum_{i}^{N}\sum_{i'}^{N}\frac{a_{ii'}x_{ii'}}{x_{i.}x_{.i'}} - \frac{1}{\kappa(\kappa-1)}\sum_{i}^{N}\sum_{i'}^{N}\frac{a_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}} \quad \text{si } \kappa > 1 \tag{29}$$

X est soumis aux contraintes d'une relation d'équivalence, formule (12).

Dans l'équation (29) 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 **X** rend possible implicitement le calcul de la somme des liens intra-classe et inter-classe. D'autre part le

16. La valeur 2 vient du fait que pour un graphe orienté il peut y avoir deux arcs reliant la même paire de sommets.

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 l'objet i.
L'équation (29) nous montre que ce critère est non-linéaire par rapport à **X**, séparable et non-équilibré.
Notons **U** la matrice carrée d'ordre N, de terme général :

$$u_{ii'} = \frac{x_{ii'}}{x_{i.}x_{.i'}} \qquad \forall i, i' \in V \times V$$

et sa complémentaire :

$$\bar{u}_{ii'} = \frac{\bar{x}_{ii'}}{x_{i.}x_{.i'}} \qquad \forall i, i' \in V \times V$$

La Matrice **U** peut s'interpréter comme une matrice de densité ou de taux d'occupation. Il est connu qu'après permutation des lignes et colonnes de la matrice **X** (revenant à mettre les objets appartenant à la même classe ensemble) on obtient une matrice diagonale par blocs. De plus, 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 des objets i et i'. 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'. La matrice **U** possède d'autres propriétés intéressantes :

1. Elle est symétrique et $u_{ii'} \in [0, 1] \qquad \forall i, i' \in V \times V$.
2. La somme des éléments de chaque bloc vaut 1.
3. Comme il y a κ classes, la somme de ses éléments vaut κ (conséquence de la propriété prédédente et de l'équation (8)), soit, $\sum_{i,i'} u_{ii'} = \kappa$. De plus, $\sum_{i,i'} \bar{u}_{ii'} = \kappa(\kappa - 1)$.
4. Elle a la propriété de "Blocs Diagonalisation Constante".
5. Chacun de ses termes est le carré du terme général de la matrice relationnelle pondérée inconnue : $u_{ii'} = \hat{x}_{ii'}^2 \qquad \forall (i, i') \in V \times V$.

En se servant de la propriétés 4, il est possible de démontrer que maximiser (29) revient à maximiser :

$$F_{MG}(U) = \frac{1}{\kappa} \sum_{i}^{N} \sum_{i'}^{N} a_{ii'} u_{ii'} + \frac{1}{\kappa(\kappa - 1)} \sum_{i}^{N} \sum_{i'}^{N} \bar{a}_{ii'} \bar{u}_{ii'} \tag{30}$$

Dans l'expression (30) les coefficients $\frac{1}{\kappa}$ et $\frac{1}{\kappa(\kappa-1)}$ sont des pondérations du terme d'accords positifs et du terme d'accords négatifs respectivement. Lorsque $\kappa = 2$ on retrouve le critère de Condorcet pondéré deux fois sur X (voir [Marcotorchino et El Ayoubi (1991)]) :

$$F_{MG}(U) = \frac{1}{2} \left(\sum_{i}^{N} \sum_{i'}^{N} a_{ii'} u_{ii'} + \sum_{i}^{N} \sum_{i'}^{N} \bar{a}_{ii'} \bar{u}_{ii'} \right) \tag{31}$$

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 [17] et 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 sous la forme relationnelle générique :

$$F(Z, r) = r \sum_{i}^{N} \sum_{i'}^{N} a_{ii'} z_{ii'} + (1 - r) \sum_{i}^{N} \sum_{i'}^{N} \bar{a}_{ii'} \bar{z}_{ii'} \tag{32}$$

On retrouve, encore une fois la généralisation proposée par [Owsiński et Zadrożny (1986)]. Il découle alors que :

- Nous obtenons formellement le critère de Zahn-Condorcet si dans (32) $r = 1/2$ et $z_{ii'} = x_{ii'}$.
- Nous obtenons formellement le critère de Mancoridis-Gansner si dans (32) : $r = \frac{\kappa-1}{\kappa}$ 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.

5.2.2 Le critère de Ratio-Cuts de Y. Wei and C.K. Cheng (1989)

Le critère "Ratio-cuts" de [Wei et Cheng (1989)] est né dans le domaine du partitionnement des circuits électriques. Un bon partitionnement 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 [18] (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 et 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 et Cheng (1989)] cherche à trouver la meilleure partition en 2 sous-ensembles disjoints U et W en minimisant la fonction suivante :

$$Rcut = \frac{e(U, W)}{|U| * |W|} \tag{33}$$

17. Puisque $\kappa \in [2, N]$, est un entier positif et puisqu'il dénote le nombre de classes. D'autre part, maximiser le critère de l'équation 30 revient à maximiser ce même critère multiplié par $(\kappa - 1)$

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

Où $e(V, W)$ est le nombre d'arêtes entre les classes U et W, 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 $\frac{1}{2}\sum_{i}^{N}\sum_{i'}^{N} a_{ii'}\bar{x}_{ii'}$, 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 notation relationnelle comme la quantité à minimiser suivante :

$$F_{Rcut}(X) = \sum_{i}^{N}\sum_{i'}^{N} \frac{a_{ii'}\bar{x}_{ii'}}{x_{i.}x_{.i'}} \tag{34}$$

Où **X**, comme dans les cas précédents, doit satisfaire les contraintes linéaires d'une relation d'équivalence données par l'équation (12).

On reconnaît immédiatement le terme de droite de (29). Il s'agit du nombre d'arêtes inter-classes pondéré par la taille des classes. En effet, la partie variable de (29) représente le terme terme général de la matrice $\bar{\mathbf{U}}$ (complémentaire de la matrice de taux de densité **U**). L'écriture relationnelle met en évidence que le critère est non-linéaire para rapport à **X** mais il est néanmoins séparable.

5.2.3 Le critère de la Différence de Profils (1976)

Ce critère a été introduit dans le cas d'une mesure contingentielle d'écart entre deux partitions sous le nom de *Distance du* Φ^2 (dans le livre de [Cailliez et Pagès (1976)]). Il a été écrit, avec des notations relationnelles pour la première fois dans [Marcotorchino (1991)], et étudié par [Bedecarrax et Marcotorchino (1992)] et [Marcotorchino et El Ayoubi (1991)]. Il cherche à minimiser l'expression suivante :

$$F_{DP}(X) = ||\hat{\mathbf{A}} - \hat{\mathbf{X}}||^2 = \sum_{i}^{N}\sum_{i'}^{N} \left(\hat{a}_{ii'} - \hat{x}_{ii'}\right)^2 \tag{35}$$

Le critère *différence de profils* constitue une distance euclidienne carrée entre les profils de similarité relationnels de *présence-rareté* associés aux variables **A** et **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".

En développant l'expression (35) et en tenant compte de la définition de $\hat{\mathbf{A}}$ et $\hat{\mathbf{X}}$ (équation (5)), le critère (35) s'écrit :

$$F_{DP}(X) = Cte - 2\sum_{i}^{N}\sum_{i'}^{N} \hat{a}_{ii'}\hat{x}_{ii'} + \kappa \tag{36}$$

Ce qui revient à maximiser l'expression suivante :

$$F_{DP}(X) = 2\sum_{i}^{N}\sum_{i'}^{N}\hat{a}_{ii'}\hat{x}_{ii'} - \kappa \tag{37}$$

ou encore :

$$F_{DP}(X)) = \sum_{i}^{N}\sum_{i'}^{N}\left(2\hat{a}_{ii'} - \frac{1}{x_{i.}}\right)\hat{x}_{ii'} \tag{38}$$

où **X** doit vérifier les contraintes d'une relation d'équivalence, voir l'équation (12).

L'écriture relationnelle de la "différence de profils", équations (37) et (38), permet de constater que ce critère est non-linéaire en **X** 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. Il est intéressant de mentionner qu'il a été montré dans [Marcotorchino (2008)] qu'en s'appuyant sur la méthode des moindres carrés et le Théorème spectral de [Hoffman et Wielandt (1952)] on peut définir un indice $\Omega(\kappa)$ de classes latentes κ de la partition, dont l'optimum permet d'estimer le nombre de classes κ^* minimisant la différence de profils.

5.3 Comparaison des critères

Le Tableau 3 montre l'expression des critères séparables et linéaires en **X** (tous les critères sont à maximiser) :

Critère	**Écriture Relationnelle**
Zahn-Condorcet (1785, 1964)	$F_{ZC}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(a_{ii'}x_{ii'} + \bar{a}_{ii'}\bar{x}_{ii'})$
Owsiński-Zadrożny(1986)	$F_{Z_{OZ}}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}((1-\alpha)a_{ii'}x_{ii'} + \alpha\bar{a}_{ii'}\bar{x}_{ii'})$ avec $0 < \alpha < 1$
Newman-Girvan (2004)	$F_{NG}(X) = \frac{1}{2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'}x_{ii'} + \frac{a_{i.}a_{.i'}}{2M}\bar{x}_{ii'}\right)$
Demaine-Immorlica (2002)	$F_{D}(X) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(w^{+}_{ii'}x_{ii'} + w^{-}_{ii'}\bar{x}_{ii'})$
Condorcet pondéré en **A** (1991)	$F_{C}(X,\hat{A}) = \sum_{i=1}^{N}\sum_{i'=1}^{N}(\hat{a}_{ii'}x_{ii'} + \bar{\hat{a}}_{ii'}\bar{x}_{ii'})$

TAB. 3 – *Critères linéaires et séparables.*

Ce Tableau 3 nous montre également que les critères de Zahn-Condorcet, celui d'Owsiński-Zadrożny, celui de Demaine-Immorlica et Condorcet pondéré en **A** sont en filiation directe avec le critère de Condorcet (1785). Le critère d'Owsiński-S. Zadrożny est une généralisation de Condorcet, il lui donne plus de flexibilité comme nous allons voir dans l'exemple pratique. Les critères de Demaine et Immorlica et Condorcet pondéré en **A** se distinguent de celui de Condorcet par le type des données d'entrée. Ainsi, pour Demaine-Immorlica, les données d'entrée sont des poids réels et Condorcet pondéré en **A** utilise la matrice d'adjacence pondérée. Maximiser l'expression du critère de Newman-Girvan du tableau 3 revient à maximiser (21), en effet : $\frac{1}{2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'}-\frac{a_{i.}a_{.i'}}{2M}\right)x_{ii'} = \frac{1}{2M}\sum_{i=1}^{N}\sum_{i'=1}^{N}\left(a_{ii'}x_{ii'}+\frac{a_{i.}a_{.i'}}{2M}\bar{x}_{ii'}\right)-1$

Le Tableau 4 montre les critères séparables de fonctions non-linéaires de **X** étudiés. Tous les critères sont à maximiser sauf celui de [Wei et Cheng (1989)] :
Le Tableau 4 montre des critères non-linéaires en **X**. La partie variable de ces critères est une

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

TAB. 4 – *Critères séparables fonctions non-linéaires de **X**.*

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.

6 Exemples d'application

Nous avons testé deux jeux d'école : l'un relatif aux données du réseau social dit "club de Karaté de Zachary" [Zachary (1977)] et le deuxième relatif aux réseaux "College Football" [Girvan et Newman (2002)] avec l'algorithme de Louvain [Blondel et al. (2008)] en utilisant les critères de Newman-Girvan, Zahn-Condorcet, Owsinski-Zadrozny, Condorcet pondéré en **A** et celui de la "différence de profils". La linéarité des 4 premiers critères ont permis de les adapter facilement à l'algorithme de Louvain.

Club de Karaté de Zachary Le graphe de Zachary est un réseau social réel composé de 34 membres (sommets) et 78 liens (arêtes) d'un club de karaté. Un désaccord entre l'administrateur (sommet 1) et l'instructeur du club (sommet 34) a séparé le réseau en deux groupes de

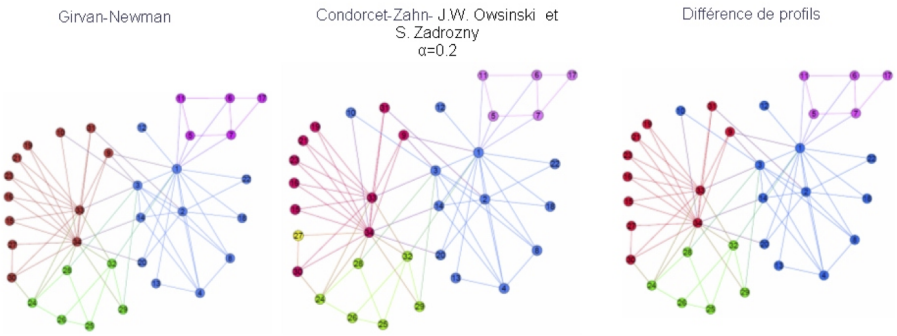

FIG. 2 – *Résultat de la Modularisation du réseaux Zachary's karate club.*

taille semblable. La FIG. 2 nous montre le résultat obtenu [19] :

La Figure 2 montre que les résultats obtenus avec les 3 critères sont similaires :
- L'optimisation directe du critère de Condorcet-Zahn partitionnait le graphe en 18 classes. L'application du critère de Owsinski-Zadrozny avec $\alpha = 0.2$ nous a permis de rendre Condorcet-Zahn plus flexible. Ainsi dans notre exemple, chaque classe doit contenir au moins 20% d'arêtes intra-classes par rapport à la quantité maximale possible. Le choix de cette valeur dépend fortement de la définition de la notion de *communauté*.
- Le critère de la différence de profils trouve un résultat proche de celui de Newman-Girvan : 4 classes.
- Dans la construction du réseau réel Zachary avait hésité avec la classe du sommet 10. Ici Newman-Girvan le classe avec l'instructeur tandis que Owsinski-Zadrozny et la différence de profils le classent avec l'administrateur.

College Football : Ce réseau, déjà étudié dans [Girvan et Newman (2002)], représente le calendrier des matchs de football américain pour la saison 2000. Chaque sommet du graphe représente une équipe (au total 115) et les arêtes (au total 1226) représentent les matchs entre équipes. Ce qui rend ce réseau intéressant, c'est qu'il intègre une structure communautaire connue. En effet, les équipes sont divisées en 12 tournois contenant entre 7 à 13 équipes chacun. Les jeux sont plus fréquents entre les membres du même tournoi qu'entre les membres des différents tournois.

Nous avons modularisé ce graphe avec 4 critères. Le tableau 5 montre les résultats obtenus avec ces critères : 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. 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%.

19. Les graphes ont été dessinés avec le logiciel Gephi

Critère	**Communautés**	**Accord avec graphe réél** (%)
Graphe Réel	12	
Newman-Girvan	10	96,9%
Zahn-Condorcet	16	97,7%
Condorcet pondéré an A	20	96,0%
Différence de profils	9	95,0%

TAB. 5 – *Résultats trouvés avec le réseau "College football"*

Les graphes des partitions obtenues ne sont pas montrés dans cet article vu la taille du graphe, cependant voici quelques remarques importantes concernant les résultats obtenus :

- Le tableau 5 montre que l'on obtient un pourcentage d'accords élevé pour les 4 critères. Cependant, le nombre de classes optimal varie d'un critère à l'autre. En effet, Newman-Girvan et la différence de profils, sous estiment le nombre de classes tandis que les partitions obtenues avec les critères de Zahn-Condorcet et Condorcet pondéré contiennent un nombre de classes supérieur au nombre de classes attendu : 12.
- Le critère de Girvan Newman n'identifie pas les tournois "IA Independents" (5 sommets) et "Sunbelt" (7 sommets), en effet la moitié de sommets de ce dernier es classée avec les sommets du tournoi "Mountain West" tandis que l'autre moitié est classé dans le tournoi "SEC".
- Les critères de Zahn-Condorcet et celui de Condorcet pondéré ont tendance à couper les classes de la partition originale en sous classes. C'est le cas pour les tournois "Mid American" (13 sommets), "Big 12" (8 sommets) et "sunbelt" (7 sommets) qui sont coupés en 2 sous graphes de tailles semblables. Le critère de Condorcet pondéré coupe aussi les tournois "SEC" (12 sommets) et "PACK10" (10 sommets) en 3 et 2 sous-groupes respectivement. D'autre part, les partitions obtenus avec ces deux critères contiennent aussi 1 et 2 classes à un seul sommet respectivement.
- Le critère de la différence de profils a plutôt tendance à fusionner les tournois, son écriture relationnelle dans sa version à maximiser (voir tableau 4) montre que la valeur du critère augmente quand les tailles des classes sont plus importantes. En effet, ce critère met dans une même classe les sommets appartenant aux 3 tournois : "PAC 10", "Mountain West" et "Sunbelt". Comme Newman-Girvan, ce critère n'identifie pas le tournoi "IA Independents".

7 Conclusions

Le tableau suivant présente une caractérisation des communautés obtenues par chaque critère, il se base sur les résultats expérimentaux obtenus :

- L'Analyse Relationnelle constitue une approche unificatrice qui permet de comparer les différents critères sous la même base avec des notations standard et simples. L'Analyse Relationnelle nous a permis d'écrire certains critères de modularisation sous forme de Modèles Linéaires en variables bivalentes.

Critère	**Résultat obtenu**
Zahn-Condorcet	Il garantit un pourcentage d'arêtes intra-classe supérieur ou égal à 50%.
Owsiński - Zadrożny	L'utilisateur peut définir le pourcentage minimal d'arêtes intra-classe requis.
Newman-Girvan	Ce critère possède une limite de résolution, les tailles de classes obtenues dépendent des caractéristiques globales du réseau, plus la taille du réseau augmente, plus il fusionne les groupes denses.
Condorcet pondéré en **A**	Il défavorise les classes très peuplées.
Différence de profils	Il favorise les classes peuplées.

TAB. 6 – *Caractérisation des communautés obtenues selon critère.*

– L'écriture Relationnelle des critères permet d'une part leur comparaison, d'autre part de caractériser leurs principales propriétés structurelles et donc de comprendre leur biais ou défauts éventuels. Cela permet de guider l'utilisateur par rapport au choix du critère en fonction du besoin spécifique. Nous avons montré avec un exemple que ce choix dépend fortement de la notion que l'utilisateur a de *communauté*.

– Nous avons montré qu'une quantité non négligeable de critères dérivait du critère de Condorcet (1785) notamment : Owsinski-Zadrozny, Zahn et Demaine. D'autres critères optimisent une version pondérée du critère de Condorcet. C'est le cas du critère de Mancoridis-Gansner et du critère *Ratio-cuts*.

Références

Bansal, N., A. Blum, et S. Chawla (2002). Correlation clustering. In *IEEE Symp. on Foundations of Computer Science*.

Barabasi, A. et J. Frangos (2002). *Linked : The New Science Of Networks Science Of Networks*. Perseus Publishing.

Barabási, L. (2012). Dossier : La théorie de la compléxité. *Journal pour la science : La Recherche*, pp. 36–49.

Bedecarrax, C. et F. Marcotorchino (1992). *"La Distance de la Différence de Profils" dans le livre "Distance"*, pp. 199–203. Publication Université de Haute Bretagne.

Belson, W. A. (1959). Matching and prediction on the principle of biological classification. *Survey Research Centre, London School of Economics and Political Science*.

Benzécri, J. et Collaborateurs (1973a). *Analyse des Données, Tome I : Classification Automatique*. Paris : Dunod Editeur.

Benzécri, J. et Collaborateurs (1973b). *Analyse des Données, Tome II : Analyse des Correspondances*. Paris : Dunod Editeur.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, et E. Lefebvre (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment P10008*.

Brandes, U., D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, et D. Wagner (2008). On modularity clustering. *IEEE Transactions on Knowledge and Data Engineering 20*(1), 172–188.

Cailliez, F. et J. Pagès (1976). *Introduction à l'analyse des données*. Société de Mathématiques Appliquées et de Sciences Humaines.

Caritat A., M. d. C. (1785). 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.

De Montgolfier, F., M. Soto, et L. Viennot (2012). Modularité asymptotique de quelques classes de graphes. In F. Mathieu et N. Hanusse (Eds.), *14èmes Rencontres Francophones sur les Aspects Algorithmiques des Télécommunications (AlgoTel)*, La grande motte, France, pp. 1–4.

Delest, M., J. Fedou, et G. Melancon (2006). A quality measure for multi-level community structure. In *Proceedings of the Eighth International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, Timisoara, Romania, pp. 63–68. IEEE Computer Society.

Demaine, E. D. et N. Immorlica (2003). Correlation clustering with partial information. pp. 1–13.

Girvan, M. et M. E. J. Newman (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America 99*(12), 7821–7826.

Guilbaud, G. (1952). *Les théories de l'intérêt général et le problème logique de l'agrégation*. Revue économique.

Hoffman, A. et H. Wielandt (1952). The variation of the spectrum of a normal matrix. *Duke Math. Journal* (20), 37–39.

Labiod, L. (2008). *Contribution au Formalisme Relationnel de la Classification Croisée de deux Ensembles*. Thèse de doctorat, Université Pierre et Marie Curie.

Labiod, L., N. Grozavu, et Y. Bennani (2010). Relationship between the modularity criterion and the relational analysis. In *Advanced Information Management and Service (IMS), 2010 6th International Conference on*, pp. 229–235.

Mancoridis, S., B. Mitchell, C. Rorres, Y. Chen, et E. Gansner (1998). 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)*, Ischia, Italy, pp. 45–52. IEEE Computer Society.

Marcotorchino, F. (1989). *Liaison Analyse Factorielle-Analyse Relationnelle (I) : "Dualité Burt-Condorcet"*. Paris : Etude du Centre Scientifique IBM France, No F142.

Marcotorchino, F. (1991). *L'analyse Factorielle-Relationnelle (parties 1 et 2)*. Paris : Etude du Centre Scientifique IBM France, M06.

Marcotorchino, F. (2000). *"Dualité Burt-Condorcet :relation entre analyse factorielle des correspondances et analyse relationnelle", dans le livre "Analyse des Correspondances et*

Techniques Connexes". Berlin : Moreau J., Doudin P.A., Cazes P. Editeurs, Springer-Verlag Berlin.

Marcotorchino, F. (2008). Analyse factorielle relationnelle : Calcul des inerties interclasses et du nombre de classes latentes.

Marcotorchino, F. et N. El Ayoubi (1991). Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association. *Revue de Statistique Appliquée 39*(2), 25–46.

Marcotorchino, F. et P. Michaud (1979). *Optimisation en Analyse ordinale des données*. Paris : Masson.

Newman, M. et M. Girvan (2004). Finding and evaluating community structure in networks. *Physical Review E 69*(2).

Owsiński, J. et S. Zadrożny (1986). Clustering for ordinal data : a linear programming formulation. *Control and Cybernetics 15*(2), 183–193.

Viennet, E. (2009). Recherche de communautés dans les grands réseaux sociaux. *Revue des Nouvelles Technologies de l'Information (RNTI-A3)*, 145–160.

Wei, Y. et C. Cheng (1989). Towards efficient hierarchical designs by ratio cut partitioning. *IEEE International Conference on Computer-Aided Design*, 298–301.

Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research 33*, 452–473.

Zahn, C. (1964). Approximating symmetric relations by equivalence relations. *SIAM Journal on Applied Mathematics 12*, 840–847.

Summary

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. In this paper, we give through a relational coding, the way of comparing different criteria of modularization such as: Newman-Girvan, Mancoridis-Gansner, or Condorcet-Zahn, etc. This representation facilitates their understanding and their usefulness in some practical contexts, where their purposes become easily interpretable and understandable.