Partition et recouvrement de communautés dans les graphes bipartis, unipartis et orientés

Michel Crampes, Michel Plantié

Laboratoire LGI2P, École des Mines d'Ales, Site de Nîmes
Parc Georges Besse, 30035 Nîmes Cedex, France
michel.crampes-@-mines-ales.fr http://www.lgi2p.ema.fr/plantie

Résumé :
Le classement d'entités est à la base de la production de nouvelles connaissances que ce soit pour identifier des concepts, ou bien pour propager des propriétés aux membres d'une classe. De très nombreuses méthodes de classification ont vu le jour. Les méthodes récentes de recherche de communautés dans les réseaux sociaux apportent un nouvel éclairage dans le domaine. Dans un premier temps les travaux ont porté principalement sur la détection de communautés partitionnées dans les graphes unipartis non orientés. A l'inverse tant la détection de communautés dans les graphes bipartis que le recouvrement de communautés dans les deux types de graphes ont été beaucoup moins explorés. Partant des graphes bipartis nous proposons dans cet article une méthode simple et originale qui unifie la détection de communautés partitionnées et leur recouvrement dans les graphes bipartis, les graphes unipartis non orientés et les graphes orientés. Nous montrons sur des exemples concrets comment notre méthode peut s'étendre à l'analyse de données plus générales et permet d'extraire de la connaissance en juxtaposant le partitionnement et le recouvrement.

1. Introduction

Avec le développement d'Internet et l'importance prise par les réseaux sociaux, la recherche de communautés à partir de graphes fait l'objet de nombreux travaux. Le principe général consiste à regrouper les individus de telle manière que les liens qu'ils entretiennent avec ceux présents à l'intérieur de leur communauté soient plus nombreux ou plus forts que les liens qu'ils entretiennent avec ceux présents dans les autres communautés. Beaucoup d'algorithmes aux approches très variées ont été proposés. Les plus importants sont rapportés dans (Papadopoulos *et al.* (2011); Yang *et al.* (2010); Porter *et al.* (2009)) et de manière plus détaillée dans (Fortunato (2009)). En grande majorité ils portent sur la détection de communautés partitionnées (ensembles disjoints de noeuds) dans les graphes unipartis. Dans ce type de graphes tous les noeuds peuvent éventuellement être reliés. Ces deux conditions limitent fortement le champ d'application de ces algorithmes. En effet dans la réalité les individus appartiennent à plusieurs

communautés. Le recouvrement est plutôt la règle. De plus de nombreuses relations sociales sont médiatisées par des propriétés communes entre individus. On parle alors de graphes multimodaux, le cas le plus fréquent étant le graphe bimodal (ou biparti).

Des travaux récents se sont intéressés à la détection de communautés partitionnées dans les graphes bipartis. D'autres travaux à l'inverse se sont intéressés à la recherche de communautés recouvrantes dans les graphes unipartis. Les recherches sur la détection de communautés recouvrantes dans les graphes bipartis sont rares. Dans un article précédent nous nous étions intéressés à ce problème en utilisant les treillis de Galois (Crampes & Plantié (2012)). Dans le présent article nous proposons une méthode différente. Elle utilise tout algorithme de partitionnement de graphes unipartis pour traiter à la fois les graphes unipartis, les graphes bipartis et les graphes orientés et produire des communautés partitionnées et recouvrantes. Concrètement nous utilisons la méthode de Louvain (Blondel *et al.* (2008)). Au delà de son caractère unificateur notre méthode tranche aussi par la présentation des résultats qui sont compréhensibles et accessibles aux non spécialistes. Cette dernière caractéristique est essentielle pour l'analyse et l'extraction de connaissance. Appliquée à des benchmarks bien connus nous obtenons des résultats au moins aussi bons et souvent meilleurs que ceux obtenus par d'autres auteurs. De plus la méthode peut être appliquée à des domaines d'analyse de données autres que les réseaux sociaux. En particulier nous l'avons appliquée au clustering de 'régions d'intérêt' dans le cortex à partir de données de tractographie cérébrale. Ainsi il apparait que la méthode proposée peut être utilisée au même titre que des méthodes d'analyse de données telles que K-means ou le clustering spectral, avec l'avantage qu'elle ne nécessite pas l'intervention de l'expérimentateur pour choisir un nombre de clusters ou un seuil de clustering. En juste retour des choses, après avoir historiquement emprunté ses méthodes à l'analyse de données, la recherche en matière de détection de communautés peut proposer de nouvelles méthodes applicables à d'autres domaines avec des éclairages différents.

2. État de l'art

Les nombreux travaux de recherche se sont focalisés dans un premier temps sur la détection de communautés partitionnées dans des graphes unipartis. Ils ont exploré de nombreuses pistes empruntées à l'analyse de données (méthodes de clustering), à des métaphores physiques (recuit simulé) ou biologiques (algorithmes génétiques), à l'analyse de graphes (cliques, centralité, marche aléatoire), etc. Le lecteur intéressé en trouvera une liste conséquente dans (Fortunato (2009)) et une évaluation sociologique de certains algorithmes dans (Cazabet *et al.* (2012)). Parmi elles certaines retiennent plus particulièrement notre attention. La première, de type hiérarchique, peut prendre une forme ascendante (agrégative) ou descendante (Newman & Girvan (2004)). Le résultat peut être représenté par un dendogramme. L'arrêt et donc le nombre de communau-

tés relève d'un choix extérieur à l'algorithme. Une autre approche combine l'analyse spectrale et une méthode de projection sur les axes des vecteurs propres (Pothen *et al.* (1990)). Elle reprend en cela une méthode éprouvée dans l'analyse de données. Récemment ces méthodes se sont enrichies de la définition de la modularité de l'organisation d'un graphe en communautés (Newman (2006)). La plupart des algorithmes proposent des stratégies de partitionnement qui cherchent à maximiser la modularité. Dans cette veine l'algorithme de Louvain (Blondel *et al.* (2008)) consiste à agréger itérativement les noeuds du graphe. Il retient notre attention du fait qu'il ne nécessite pas de donner à priori le nombre de clusters ou un seuil de clustering.

En relation avec le présent article nous nous focalisons sur les travaux récents qui cherchent à produire pour l'essentiel soit des communautés partitionnées à partir de graphes bipartis soit à l'inverse des communautés recouvrantes à partir de graphes unipartis. Dans le premier cas une tendance est d'abord de construire un modèle de modularité inspiré de Newman mais spécifique aux graphes bipartis avec l'argument que les deux ensembles, dans le cas général, ne devraient pas générer le même nombre de communautés (Murata (2010); Suzuki & Wakita (2009)). Cependant à partir de (Barber (2007)) une expression de la modularité pour les graphes bipartis est directement dérivée de Newman et reprise par divers auteurs pour appliquer des méthodes classiques telles que le recuit simulé (Guimerà *et al.* (2007)), le clustering spectral (Barber (2007)), les algorithmes génétiques (Nicosia *et al.* (2009)), la transmission de labels (Liu Xin & Murata Tsuyoshi (2010)), ou encore l'analyse spectrale dichotomique (Leicht & Newman (2007)).

De leur coté les stratégies de détection de communautés recouvrantes à partir de graphes unipartis éventuellement pondérés étendent bien souvent les méthodes de partitionnement. (Palla *et al.* (2005)) utilisent une métaphore de percolation de k-cliques. (Davis & Carley (2008)) font appel à la marche aléatoire dans un graphe. (Gregory (2009)) utilise les algorithmes de propagation de labels. Certains auteurs proposent des méthodes spécifiques. (Wu *et al.* (2012)) cherchent les recouvrements entre communautés partitionnées. (Reichardt & Bornholdt (2006)) combinent le modèle d'interaction de spins de Pott et le recuit simulé. (Lancichinetti *et al.* (2009)) optimisent un fonction statistique locale. (Evans & Lambiotte (2009)) traitent le problème dual de partitionner les liaisons pondérées.

Dans la littérature les travaux qui portent à la fois sur les graphes bipartis et le recouvrement des communautés se font plus rares. On retrouve l'extension de méthodes telle la recherche de bicliques recouvrantes (Lehmann *et al.* (2008)). D'autres méthodes originales font appel aux résultats connus dans les treillis de Galois (Crampes & Plantié (2012)) et (Roth *et al.* (2008)), mais la représentation et les algorithmes sont complexes. Nos travaux exposés ici se distinguent par le fait que nous unifions les graphes unipartis, bipartis et orientés pour produire des communautés à la fois partitionnées et recouvrantes.

3. Unification des graphes bipartis, unipartis et orientés

Afin d'unifier ces trois types de graphes, nous les ramenons à un modèle unique de graphe biparti susceptible d'être interprété comme un graphe uniparti pour ensuite en détecter les communautés. La construction de ce modèle de graphe biparti passe par la construction de la matrice d'adjacence.

3.1. Graphes bipartis

Formellement un graphe biparti $G = (U, V, E)$ est un graphe $G' = (N, E)$ dont l'ensemble des noeuds N est l'union de deux ensembles de noeuds indépendants U et V et dont les arêtes connectent uniquement les paires de noeuds (u, v) tels que u appartient à U et v appartient à V.

$N = U \cup V$,
$U \cap V = \emptyset$,
$E \subseteq U \times V$

De manière évidente si $r = |U|$ et $s = |V|$, alors $|N| = n = r + s$

La matrice binaire de biadjacence d'un graphe biparti est par définition la matrice B à r lignes et s colonnes dans laquelle :

$B_{i,j} = 1$ ssi $(u_i, v_j) \in E$ et $\mathrm{B}_{i,j} = 0$ ssi $(u_i, v_j) \notin E$.

Il est important de souligner que la marge des lignes de B représente les degrés des noeuds u_i et la marge des colonnes représente les degrés des noeuds v_j.

Par définition la matrice d'adjacence A' de G' est la matrice bloc :

$A' = \begin{pmatrix} 0_r & B \\ B^t & 0_s \end{pmatrix}$ dans laquelle 0_r est une matrice carrée nulle d'ordre r et 0_s est une matrice carrée nulle d'ordre s.

L'utilisation de ces définitions mathématiques connues nous permet de substituer au graphe biparti $G = (U, V, E)$ le graphe uniparti $G' = (N, E)$. La recherche de communautés dans G peut se faire en recherchant les communautés dans G' par l'application à la matrice A' d'un algorithme pour les graphes unipartis. Le procédé d'utilisation du graphe uniparti en lieu et place du graphe biparti pour rechercher des communautés a déjà été envisagé par (Barber (2007)). Mais cet auteur n'en tire pas toutes les conclusions unificatrices et n'en exploite pas les possibilités en termes de détection de communautés à la fois partitionnées et recouvrantes.

3.2. Graphes orientés

Un graphe orienté $G^d = (N^d, E^d)$ est un graphe où N^d est l'ensemble des noeuds et E^d est un ensemble d'arcs entre les noeuds N^d : $E^d \subseteq N^d \times N^d$. Il a pour matrice d'adjacence la matrice asymétrique A^d. Pour transformer un graphe orienté en un graphe biparti nous dupliquons les noeuds pour leur attribuer un rôle. Nous obtenons ainsi un

ensemble N^{out} identique à l'ensemble des noeuds N^d vus comme noeuds de départ des arcs, et un ensemble N^{in} identique à l'ensemble des noeuds N^d vus comme noeuds d'arrivée des arcs. Le graphe orienté G^d est en conséquence transformé en un graphe biparti $G = (N^{out}, N^{in}, E)$ dans lequel N^{out} et N^{in} sont des bijections de N^d et, E en bijection avec E^d, est l'ensemble des arêtes reliant un noeud de départ à un noeud d'arrivée. La matrice d'adjacence de G^d joue le rôle de la matrice de biadjacence de G. Cette transformation a été aussi suggérée par (Guimerà *et al.* (2007)). Pour rechercher des communautés dans un graphe orienté il est donc possible comme pour tout autre graphe biparti de considérer la transformation du graphe G^d en un graphe uniparti $G' = (N, E)$ avec sa matrice d'adjacence bloc A'. Dans A' les noeuds N sont construits par dédoublement des noeuds N^d en un ensemble de noeuds départs N^{out}, et en un ensemble de noeuds arrivés N^{in}.

3.3. Graphes unipartis

Les deux procédés ci-dessus permettent de transformer un graphe biparti ou un graphe orienté en un graphe uniparti pour lequel il nous est ensuite possible de faire une recherche de communautés avec un algorithme dédié aux graphes unipartis. Nous montrons ici de manière originale comment il est aussi possible de transformer un graphe uniparti en un graphe biparti pour ensuite revenir à un graphe uniparti que nous appellerons le graphe uniparti étendu. Ce procédé peut paraitre contre productif. Quel intérêt de transformer un graphe uniparti en un graphe biparti pour ensuite le transformer de nouveau en un graphe uniparti étendu ? Deux raisons plaident pour cette transformation.

D'une part en terme de validation, comme il est possible d'appliquer un même algorithme de graphe uniparti sur un graphe uniparti d'origine et sur son graphe transformé en un graphe uniparti étendu il sera possible de vérifier que cet algorithme donne les mêmes communautés dans les deux cas. C'est effectivement l'un des procédés de validation que nous avons utilisé, sachant qu'il est difficile de dire en général qu'un algorithme donne les 'bonnes' communautés sauf à se comparer à d'autres auteurs.

D'autre part il existe peu d'algorithmes de recherche de recouvrements de communautés pour les graphes unipartis. Appliquant un même procédé pour les trois types de graphes (biparti, orienté et uniparti) le deuxième intérêt est que nous disposons d'une technique unique pour l'identification et la visualisation de recouvrements qui fait appel au caractère biparti comme nous le verrons ci-dessous.

Concrètement la transformation originale que nous proposons est la suivante. Étant donné un graphe uniparti $G^m = (N^m, E^m)$ ayant pour matrice d'adjacence la matrice symétrique A^m d'ordre N^m, nous construisons le graphe biparti $G = (U, V, E)$ où $U = V = N^m$ et E est l'ensemble des arêtes tel que d'une part $\forall x, y \in E^m$ si $(x, y) \in E^m$, $(x, y) \in E$ et $(y, x) \in E$, et d'autre part $(x, x) \in E$. Une autre manière de considérer cette transformation est de dire que la matrice de biadjacence B de G est la matrice d'adjacence A^m de G^m à laquelle a été rajoutée la matrice unité I :

$B = A^m + I$.

Nous justifierons par la suite le fait d'avoir rajouté à A^m la matrice unité I. Il est essentiel de noter que bien que carrée et symétrique, la matrice B représente un graphe biparti dans un cas particulier où ce sont les mêmes noeuds clonés qui composent les deux ensembles U et V et que quand il y a une liaison entre deux noeuds, il existe une liaison entre les deux clones. Il est alors possible de transformer ce nouveau graphe biparti en un graphe uniparti selon le procédé d'extension ci-dessus en construisant A' avec la particularité que $B = B^t$ ce qui n'est évidemment pas le cas dans les graphes bipartis en général.

4. Détection de communautés partitionnées, recouvrantes et outils d'analyse

4.1. Définition de la modularité pour les graphes bipartis, unipartis et orientés

Ayant ramené les trois types de graphes à des graphes bipartis, puis ayant transformé ceux-ci en des graphes unipartis, il est possible d'appliquer des algorithmes connus et efficaces pour la recherche de communautés dans ces graphes. Nous nous inscrivons dans la suite des méthodes qui font appel à la modularité selon (Newman (2006)). Le principe consiste à rassembler les noeuds de telle manière qu'il y ait le maximum de liaisons à l'intérieur d'une communauté et le minimum de liens entre les communautés. Formellement, étant donné un graphe uniparti $G^m = (N^m, E^m)$, la modularité Q d'une partition de graphe est définie :

$$Q = \sum_c [\frac{|e_c|}{m} - (\frac{d_c}{2m})^2] \tag{1}$$

où $|e_c|$ est le nombre de liaisons dans la communauté c , d_c est la somme des degrés des noeuds appartenant à c et m est le nombre total de liaisons dans le graphe : $m = \frac{\sum_c d_c}{2}$. Il est possible de reformuler la modularité en prenant en compte la matrice d'adjacence A. Dans le cas plus général d'un graphe pondéré :

$$Q = \frac{1}{2m} \sum_{i,j} \left[A^m_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{2}$$

où A^m_{ij} représente le poids de la liaison entre i et j, $k_i = \sum_j A^m_{ij}$ est la somme des poids des arcs attaché au noeud i, c_i est la communauté à laquelle appartient le noeud i, la fonction de Kronecker $\delta(u, v)$ est égale à 1 si $u = v$ et 0 sinon et $m = 1/2 \sum_{ij} A^m{}_{ij}$. Pour l'instant nous ne considérons que des graphes binaires et dans ce cas les poids A^m_{ij} prennent les valeurs 1 ou 0 selon que la liaison existe ou n'existe pas.

L'interprétation de cette formule est la suivante : la modularité est la somme pondérée pour toutes les communautés de la différence entre les liaisons observées à l'intérieur

de la communauté (terme A_{ij}^m) et la probabilité de ces liaisons (terme $\frac{k_i k_j}{2m}$ dont le numérateur est le produit des marges correspondant à la cellule i, j).

Dans le cas qui nous intéresse des graphes bipartis, après avoir proposé différents modèles de modularité spécifique, les publications les plus récentes partent de la matrice de biadjcence B pour proposer un modèle probabiliste intuitif directement inspiré de celui de Newman pour les graphes unipartis :

$$Q = \frac{1}{m} \sum_{i,j} \left[B_{ij} - \frac{k_i k_j}{m} \right] \delta(c_i, c_j) \tag{3}$$

Notre formulation sera proche. En effet comme nous appliquerons un algorithme de graphes unipartis, nous conservons la formule 2 pour ce type de graphe et l'appliquons à la matrice d'adjacence A'.

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij}^{'} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{4}$$

Dans cette matrice A', les noeuds de chaque ensemble U et V apparaissent par construction à la fois en colonnes et en lignes. En conséquence en recherchant les communautés d'un graphe biparti comme s'il était uniparti les deux types de noeuds U et V vont produire un même nombre de communautés dans lesquelles ils seront associés (nous en ferons une démonstration formelle dans un autre article).

Ce type de résultat a été fortement contesté par des auteurs qui considèrent que le nombre de communautés n'a aucune raison d'être identique pour les deux ensembles (Murata (2009); Suzuki & Wakita (2009)). Mais d'autres auteurs comme nous rendent solidaires les deux types d'éléments lors de l'affectation dans les mêmes communautés (Liu Xin & Murata Tsuyoshi (2010); Lehmann *et al.* (2008); Roth *et al.* (2008)).

Dans le débat ainsi ouvert entre les tenants d'une recherche de communautés différentes pour les deux types de noeuds et les tenants d'une recherche de communautés associant les deux types de noeuds il nous semble que les arguments plaident en faveur de la deuxième position. D'une part certains auteurs comme Murata d'abord en faveur de la première position se sont ralliés ensuite à la seconde en prenant acte de la pertinence de la formule 3. En effet dans cette formulation les deux types de noeuds sont solidaires au travers de la fonction de Kronecker $\delta(c_i, c_j)$. Par ailleurs, au plan expérimental, dans le tableau comparatif présenté dans (Murata (2010)) il apparait que les performances en terme d'optimisation de modularité sont aussi bonnes quand les deux types de noeuds sont associés et quand il ne le sont pas.

Prenant acte de ces constats, nous rajoutons l'intérêt majeur que présente l'association des deux types de noeuds dans les mêmes communautés. Quand on se focalise sur un ensemble, elle nous permet de donner de la sémantique à une communauté à l'aide des éléments de l'autre ensemble présents dans la même communauté. De plus nous

nous servons de cette association pour identifier les recouvrements et leur associer des mesures comme nous le voyons ci-après.

4.2. Détection de communautés partitionnées

Nous avons montré dans la section précédente que la recherche de communautés dans un graphe biparti pouvait se ramener à la recherche de communautés dans un graphe uniparti en utilisant la modularité de Newman. Il suffit de considérer la matrice d'adjacence A' et d'appliquer l'algorithme pour graphes unipartis qui parait le plus efficace et sémantiquement pertinent. Après expérimentation nous avons choisi l'algorithme de Louvain (Blondel *et al.* (2008)) étant donné son efficacité sur de grands graphes, le fait qu'il prenne en compte des graphes pondérés, et la facilité à interpréter son mode de calcul pour des non experts potentiellement utilisateurs contrairement à d'autres algorithmes tels que l'analyse spectrale, la recherche de cliques, ou bien la transmission de labels. Le lecteur intéressé pourra trouver le détail de cet algorithme dans la référence citée. Après avoir utilisé une version disponible sur Matlab, nous avons développé la nôtre en Java pour mieux ensuite la moduler selon nos objectifs.

Concrètement on fournit à l'algorithme la matrice étendue A' quel que soit le type de graphe (biparti, uniparti ou orienté) et on récupère en sortie les communautés partitionnées dans lesquelles sont répartis les deux types de noeuds (clonés pour les graphes orientés et unipartis). La possibilité d'appliquer la méthode sur des graphes unipartis transformés en graphe bipartis nous permet de vérifier sur des benchmarks classiques de graphes unipartis, comme par exemple le karaté club ou les dauphins (non présentés dans les limites de cet article) qu'elle donne bien les mêmes résultats que si l'algorithme avait directement été appliqué sur les graphes unipartis initiaux.

Le premier résultat original atteint est le suivant : les trois types de graphes peuvent être traités avec une méthode unique là où les autres auteurs, à quelques rares exceptions, appliquent des algorithmes spécifiques à chaque type de graphes. De plus l'algorithme mis en oeuvre est un de ceux utilisés pour traiter les graphes unipartis. Comparés aux autres auteurs, les résultats sont tout à fait pertinents comme nous le montrerons dans les expérimentations ci-dessous.

4.3. Sémantique des communautés partitionnées

L'approche qui permet d'optimiser la modularité en associant les deux types de noeuds dans les mêmes communautés permet de donner une signature sémantique duale à chaque communauté. En effet le regroupement des noeuds d'un type dans une communauté est le résultat de leur attirance par les noeuds de l'autre type dans cette même communauté. Ainsi un ensemble de noeuds d'un même type dans une communauté est justifié par l'ensemble des noeuds de l'autre type associés à cette communauté. On trouve ici un processus similaire à la projection des facteurs dans l'Analyse Factorielle

des Correspondances. On verra dans les expérimentations que ce caractère dual des communautés offre des possibilités d'analyse qui n'ont pas été suffisamment relevées dans la littérature.

4.4. Recouvrements et fonctions de recouvrement des communautés

L'affectation des deux types de noeuds dans les communautés nous permet d'observer les recouvrements entre communautés et de donner différentes mesures à ces recouvrements. En particulier nous avons défini une mesure de probabilité d'appartenance, une mesure floue de légitimité et une mesure de réaffectation. Dans les limites du présent article nous ne détaillons que la mesure de légitimité, les autres mesures faisant l'objet d'un autre article.

Nous définissons la fonction de légitimité $L(u_i \in c)$ qui mesure l'implication d'un noeud u_i dans une communauté c :

$$L(u_i \in c) = \frac{\sum_j B_{ij}\delta(c_j)}{|\{v \in c\}|} \tag{5}$$

où $\delta(c_j)$ vaut 1 si $c_j = c$ et vaut 0 sinon ;

Cette fonction peut s'interpréter de la manière suivante : un noeud est attiré par une communauté d'autant plus qu'il a un nombre relatif élevé de relations avec cette communauté indépendamment de la taille de cette communauté.

Pour exemple, dans le benchmark SW qui sera détaillé ci-dessous, des dames ont participé à des évènements. On peut observer que c_1 contient 7 évènements, c_2 5 évènements et c_3 2 évènements. En appliquant la formule ci-dessus on obtient pour w_1 une légitimité de $\frac{2}{7}$ en direction de c_1, $\frac{1}{5}$ en direction de c_2 et $\frac{1}{2}$ en direction de c_3. On ne démontrera pas dans les limites de cet article qu'il s'agit là d'une fonction d'appartenance floue. Cependant on verra plus loin qu'il sera possible d'appliquer une $\alpha - coupe$ pour mieux observer les recouvrements et surtout les noeuds déviants (ceux qui sont mal classés).

5. Expérimentation

Nous considérons ici quelques unes des expérimentations que nous avons menées avec notre méthode. Faute de place nous ne présentons pas deux benchmarks standards de graphes unipartis (le karaté club de Zachary et les dauphins) qui valident la démarche en produisant un partitionnement et un recouvrement parfaitement pertinents par rapport aux autres auteurs et à la structure du graphe.

5.1. Southern Women (SW)

Ce benchmark est un grand classique pour tous les auteurs qui veulent comparer leurs algorithmes d'extraction de communautés partitionnées dans les graphes bipartis. Dans les années trente une équipe d'ethnologues américains recueille de nombreuses données comportementales dans les populations du Sud des États Unis. Parmi les résultats de cette étude un tableau montre la participation de 18 dames à 14 types d'évènements sociaux différents (tea party, aide à la maison, etc.) Dans son article de référence de méta analyse (Freeman (2003)) compare les résultats de 21 méthodes de calculs de regroupements des dames en fonction de leur participation aux mêmes évènements. La plupart repèrent 2 communautés partitionnées. Depuis d'autres résultats différents sont rapportés. Nous avons extrait de l'article de Freeman le tableau Dames x Évènements, avons formalisé la matrice de biadjacence, et suivant en cela notre méthode avons construit la matrice d'adjacence A'. Nous avons ensuite appliqué l'algorithme de Louvain pour obtenir un partitionnement dans lequel les évènements sont associés aux dames. Enfin nous avons construit la matrice pondérée des recouvrements pour la mesure de légitimité et la mesure de réaffectation. Cette dernière n'est pas présentée ici.

Résultats. La Figure 1 synthétise une partie de nos résultats. Afin de disposer d'un référent visuel, le graphe biparti d'origine est représenté en deux couches, les dames sur la couche supérieure et les évènements au dessous. Une liaison entre une dame et un évènement témoigne de la participation de la dame à cet évènement. L'application de notre méthode permet de distinguer 3 communautés (1-rouge, 2-verte et 3-jaune). Ce résultat est plus précis que ceux présentés dans (Freeman (2003)) où un seul auteur voit 3 communautés (qui sont d'ailleurs très proches des nôtres) et les autres n'en voient que deux.

Au delà du partitionnement la Figure 1 montre les recouvrements avec la mesure de légitimité pour les dames et les évènements. Les meilleures valeurs sont soulignées. Seulement une dame (w8) et un évènement (e8) - en italique - présentent des valeurs de légitimité supérieures pour des communautés autres que celles auxquelles ils ont été affectés. Notre méthode met ainsi en relief des litiges d'affectation qu'aucun autre auteur n'avait fait apparaitre. Ils correspondent à un optimum local trouvé par Louvain, cette heuristique faisant partie de la classe des algorithmes gloutons pour contourner le caractère NP-complet du problème de regroupement. Nos travaux actuels montrent qu'il est possible de sortir de l'optimum local de la modularité en réaffectant les éléments litigieux.

5.2. Comptes Facebook

Avec cette expérimentation nous rentrons dans des données plus conséquentes et originales. Trois comptes Facebook ont été en parti chargés avec l'autorisation de leurs propriétaires, aucun de ceux-ci n'étant membre de notre équipe de recherche. Nous en

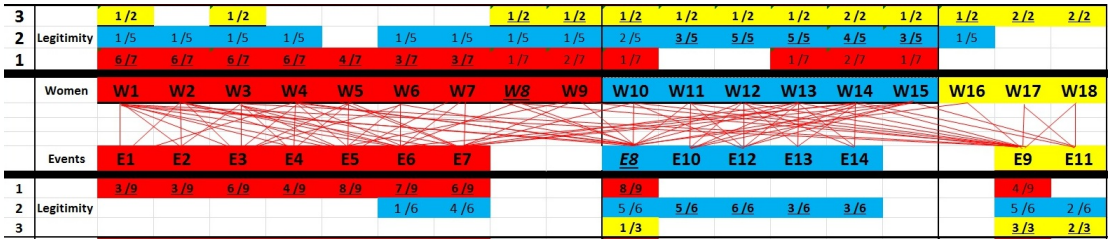

FIGURE 1: Communautés et mesures de modularités pour : Women Events

avons extrait les photos et les tags associés aux photos. Ces deux types de données reliées constituent un graphe biparti binaire (non pondéré) : une photo est reliée à un tag (en général une personne) si la photo est marquée du tag. Les photos ne sont pas directement reliées, ni les tags. Nous appliquons à ces données notre méthode de détection de communautés pour obtenir partitionnement et recouvrements. Nous présentons ci-dessous le résultat pour l'un des trois jeux de données (nous appellerons le propriétaire D), les deux autres ayant une structure similaire.

Résultats. Le jeu de données comporte environ 1000 photos pour 350 tags différents. Nous avons supprimé les tags autres que les personnes (par exemple 'paysage'). On observe tout d'abord plus de 300 communautés comportant une photo sans tag, ou une photo avec un tag unique. Ces communautés singletons s'expliquent par le fait qu'en l'absence d'autres photos partageant le même tag la connexité est brisée et une communauté spécifique est créée. Nous les ignorons parce qu'elles ne comportent qu'une seule photo. Les communautés autres que les singletons sont présentées dans la Figure 2.

En final on observe 16 communautés peu recouvrantes. Cela peut être expliqué par le fait que pour que plusieurs individus soient présents sur une même photo, il faut qu'ils aient été rassemblés au même instant dans un même lieu. En conséquence il apparait deux types d'individus : ceux liés à un moment précis de la vie de D (par exemple une promotion d'étudiants) sont présents dans une seule communauté et ceux (quelques proches ou membres de sa famille) présents à différents moments de sa vie apparaissent dans plusieurs communautés.

La première communauté se distingue du fait qu'elle contient un individu (le premier à gauche sur la figure) très présent dans presque toutes les autres communautés. Il s'agit en fait de D lui même. L'essentiel de cette communauté est composé de photos sur lesquelles D figure seul (plus de 200 photos) ou en présence d'une autre personne qui n'est pas présente dans les autres communautés. Il s'agit donc de la communauté caractéristique de D. On peut regretter que D soit associé à une communauté où il apparait bien isolé de ses groupes d'amis présents dans les autres communautés. Mais grâce au recouvrement on note aussi que D est bien présent dans presque toutes les autres communautés.

Cette expérience montre que sur ce jeu de données si on se limite au partitionnement

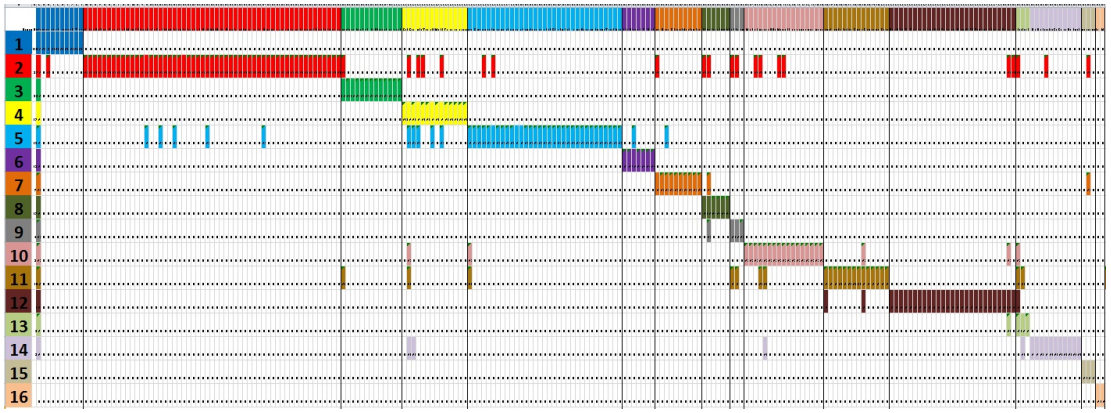

FIGURE 2: Communautés et indicateurs des recouvrements : Compte Facebook

les résultats ne peuvent être analysés et il apparait des incohérences comme par exemple l'isolement de D (il y en a d'autres que nous ne discutons pas faute de place). A l'inverse grâce aux recouvrements et à l'association des deux ensembles d'entité une analyse fine peut être opérée. La suite de l'expérimentation portera sur l'extraction d'autres connaissances telles que le rôle du temps, du lieu ou de certains individus communs à différentes communautés. De même on pourra étudier l'exploitation de ces résultats pour la constitution d'albums et la diffusion des photos.

5.3. Tractographie du cerveau

Alors que notre méthode d'extraction de communautés avait été conçue pour les réseaux sociaux, il nous a été proposé de l'utiliser pour analyser des données de tractographie du cerveau. Il s'agit, à partir de matrices probabilistes de relations d'extraire des regroupements entre régions du cortex. Les enjeux scientifiques et médicaux sont considérables.

Les différents jeux de données très récents étaient produits et fournis par des chercheurs de l'équipe 'Cognition, neuroimaging and brain diseases' du laboratoire CRICM. Ceux-ci utilisent des nouvelles méthodes de tractographie de l'imagerie par résonance magnétique (IRM) pondérée en diffusion (DWI) pour fournir un aperçu des connexions du cerveau humain et pour mettre en évidence les voies neuronales qui relient différentes régions du cortex (Catani & de Schotten (2012)). La tractographie probabiliste produit des matrices de connectivité entre des 'Régions d'Intérêt' (ROI). Les chercheurs utilisent ensuite avec succès des méthodes de clustering spectral pour regrouper ces régions en fonction de leur similarité de connexion. Mais il est difficile de valider les résultats. Si elle est applicable notre méthode permettrait d'obtenir un autre éclairage. Parmi les différentes expérimentations que nous avons menées nous présentons ici celle qui utilise la matrice de liaison entre 374 ROI du lobe occipital (LO) et 1914 ROI de

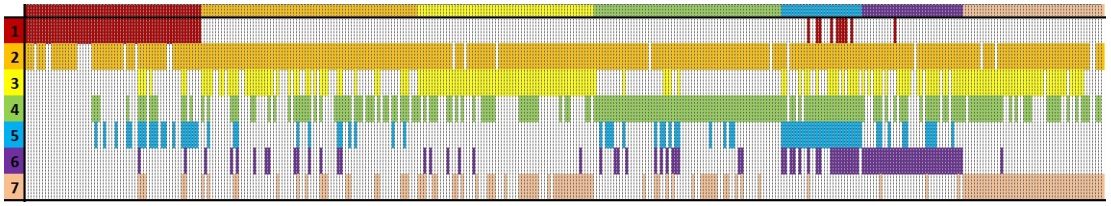

FIGURE 3: Communautés et indicateurs des recouvrements : Régions du Cerveau

l'hémisphère gauche entier (HGE). Chaque cellule représente la probabilité de liaison entre une ROI-LO et une ROI- HGE. Par rapport aux autres expérimentations ci-dessus on notera la difficulté supplémentaire d'avoir des liaisons pondérées représentant des probabilités. Cependant l'algorithme de Louvain permet de traiter des graphes unipartis pondérés. La matrice est donc considérée comme une matrice de biadjacence pondérée à laquelle nous appliquons notre méthode pour extraire des communautés partitionnées et recouvrantes.

Résultats. La Figure 3 montre l'obtention de 7 communautés alors que les chercheurs en avaient trouvé 8 par clustering spectral. Il est important de rappeler que notre méthode fournit les communautés à postériori alors que le clustering spectral suppose l'intervention de l'expérimentateur soit pour choisir à priori le nombre de clusters, soit pour le décider en fonction de la répartition de l'énergie du nuage de points sur les vecteurs propres. En l'occurrence les chercheurs s'étaient arrêtés à 8 clusters après observation de la courbe de scree test (Catani & de Schotten (2012)). Pour mieux faire apparaitre la répartition des ROI dans les clusters nous avons adapté la mesure de légitimité aux graphes pondérés ; cette adaptation n'est pas présentée dans cet article. Il est alors possible de disposer d'une fonction d'appartenance floue comme pour les graphes binaires et d'appliquer un seuil $(\alpha - cut)$. La politique de seuillage que nous avons appliquée est celle qui consiste à monter le seuil d'appartenance jusqu'au point où, dans le partitionnement, une ROI d'un cluster serait retirée de son cluster. Alors qu'avant l'application de l'alpha-coupe les clusters étaient très superposés, ils sont après seuillage bien tranchés sauf le cluster 2 qui semble être très interconnecté.

En conclusion de cette expérimentation il est intéressant de noter que la méthode de clustering spectral a fait apparaitre 8 communautés là où nous en avons fait apparaitre 7. Il n'y pas contradiction puisque la première méthode fait intervenir une décision de l'expérimentateur alors que la nôtre donne un optimum sans intervention. L'interprétation des résultats et l'exploration de nouvelles données sont en cours.

6. Conclusion

Nous avons introduit une méthode de détection de communautés pour les graphes bipartis. Son originalité tient au fait que d'une part elle permet de traiter indifféremment

plusieurs types de graphes et que d'autre part elle produit à la fois un partitionnement et les recouvrements. De plus nous avons proposé différentes mesures de recouvrement (une seule a été présentée dans l'article) et une méthode de visualisation pour les graphes de taille moyenne afin d'analyser les résultats. Enfin nous pouvons utiliser n'importe quel algorithme de recherche de communautés partitionnées dans les graphes unipartis. Cela nous permet de choisir celui qui nous parait le plus efficace et le plus simple à implémenter et à comprendre pour les non experts. Notre choix s'est porté sur Louvain. Outre le fait qu'il est facile à comprendre et qu'il donne de bons résultats, cet algorithme permet de traiter efficacement de grands graphes. Ce point permet de corriger un point faible de notre méthode que nous n'avons pas encore cité.

En effet la matrice d'adjacence A' est quatre fois plus importante que la matrice d'adjacence initiale A^m des graphes unipartis ou que la matrice de biadjacence B des graphes bipartis. En pratique, en utilisant la méthode Louvain, le volume des données à traiter est simplement le double puisque cet algorithme ne s'intéresse qu'aux liaisons existantes et la moitié de la matrice A' est composée de zéros. De plus l'efficacité de Louvain est prouvé sur des graphes comportant des millions de noeuds. Multiplier par deux le volume est donc un inconvénient limité. Le vrai problème est la visualisation des résultats comme on a pu le voir dans les expérimentations pour des matrices de taille moyenne supérieure à 500 x 500. Ici se situe le vrai enjeu parce que l'on souhaite visualiser pour analyser et ensuite interpréter.

Nous avons montré sur des benchmarks classiques des résultats expérimentaux particulièrement pertinents comparés à ceux des autres auteurs et nous avons montré sur un cas réel que notre méthode pouvait aussi s'appliquer de manière plus générale à l'analyse de données. Sur ce dernier point il est intéressant de noter que l'expérimentateur n'intervient pas dans le processus d'optimisation contrairement aux autres méthodes connues d'analyse des données. Parmi les perspectives nos travaux actuels portent sur 1) la comparaison de notre méthode avec le clustering spectral étant donné son utilisation en tractographie du cerveau en comparant les valeurs des modularités finales, 2) l'exploitation des résultats et leur visualisation pour l'analyse et l'extraction de connaissances à l'aide de diverses mesures de recouvrement, et 3) la définition de méthodes de réaffectation, sujet sur lequel aucun auteur ne s'est penché faute de disposer de mesures de recouvrements. Parmi ces chantiers, la visualisation des résultats pour les graphes de grande taille est un verrou majeur. Nous cherchons quelles sont les méthodes de visualisation les plus appropriées, sachant que nous avons pour l'instant privilégié la présentation sous forme de matrices comme dans les figures de cet article.

Au delà des qualités unificatrices et sémantiques que nous avons présentées pour notre méthode, nous apprécions de pouvoir généraliser son utilisation aux méthodes de clustering pour l'analyse de données. Elle apporte en effet un éclairage nouveau : l'expérimentateur n'intervient pas sur le nombre de clusters. La détection de communautés a longtemps emprunté ses méthodes à l'analyse de données. Nous montrons dans cet article qu'en retour l'analyse de données peut dorénavant bénéficier de méthodes nou-

velles issues de la détection de communautés.

Remerciements : Nous remercions l'équipe du projet ANR-09-RPDOC-004- 01 ainsi que le groupe de recherche CRICM UPMC U975/UMRS 975/UMR 7225 de l'Hôpital de la Salpêtrière pour nous avoir fournit les données de tractographie du cerveau.

Un projet phare ou prototype de cette communauté est le projet Human Brain Connectome : http ://www.humanconnectomeproject.org/

Références

BARBER M. (2007). Modularity and community detection in bipartite networks. *Physical Review E*, **76**(6), 1–9.

BLONDEL V. D., GUILLAUME J.-L., LAMBIOTTE R. & LEFEBVRE E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics : Theory and Experiment*, **2008**(10), P10008.

CATANI M. & DE SCHOTTEN M. T. (2012). *Atlas of human brain connections*. Oxford University Press, 2012.

CAZABET R., LEGUISTIN M. & AMBLARD F. (2012). Automated community detection on social networks : useful ? efficient ? asking the users. *Proceedings of the 4th International Workshop on Web Intelligence and Communities*, p. 1–6.

CRAMPES M. & PLANTIÉ M. (2012). Détection de communautés dans les graphes bipartis. In *IC 2012 Ingénierie des connaissance*.

DAVIS G. B. & CARLEY K. M. (2008). Clearing the FOG : Fuzzy, overlapping groups for social networks. *Social Networks*, **30**(3), 201–212.

EVANS T. S. & LAMBIOTTE R. (2009). Line Graphs, Link Partitions and Overlapping Communities. *Physical Review E*, **80**(1), 9.

FORTUNATO S. (2009). Community detection in graphs. *Physics Reports*, **486**(3-5), 103.

FREEMAN L. C. (2003). Finding social groups : A meta-analysis of the southern women data. In *Dynamic Social Network Modeling and Analysis. The National Academies*, p. 39—-97. Press.

GREGORY S. (2009). Finding overlapping communities in networks by label propagation. *New Journal of Physics*, **12**(10), 103018.

GUIMERÀ R., SALES-PARDO M. & AMARAL L. (2007). Module identification in bipartite and directed networks. *Physical Review E*, **76**(3).

LANCICHINETTI A., FORTUNATO S. & KERTÉSZ J. (2009). Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, **11**(3), 033015.

LEHMANN S., SCHWARTZ M. & HANSENLARS K. (2008). Biclique communities. *Physical review. E, Statistical, nonlinear, and soft matter physics*, **78**(1 Pt 2).

LEICHT E. A. & NEWMAN M. E. J. (2007). Community structure in directed networks. *Physical Review Letters*, **100**(11), 118703.

LIU XIN & MURATA TSUYOSHI (2010). An Efficient Algorithm for Optimizing Bipartite Modularity in Bipartite Networks. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, **14**(4), 408–415.

MURATA T. (2009). Modularities for bipartite networks. *Proceedings of the 20th ACM conference on Hypertext and hypermedia HT 09*, **90**(6), 245–250.

MURATA T. (2010). Detecting communities from tripartite networks. *WWW*, p. 0–1.

NEWMAN M. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, **74**(3 Pt 2), 036104.

NEWMAN M. & GIRVAN M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, **69**(2).

NICOSIA V., MANGIONI G., CARCHIOLO V. & MALGERI M. (2009). Extending the definition of modularity to directed graphs with overlapping communities. *Journal of Statistical Mechanics : Theory and Experiment*, **2009**(03), P03024.

PALLA G., DERENYI I., FARKAS I. & VICSEK T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, **435**(7043), 1–10.

PAPADOPOULOS S., KOMPATSIARIS Y., VAKALI A. & SPYRIDONOS P. (2011). Community detection in Social Media. *Data Mining and Knowledge Discovery*, **June**, 1–40.

PORTER M. A., ONNELA J.-P. & MUCHA P. J. (2009). Communities in Networks.

POTHEN A., SIMON H. D. & LIOU K.-P. (1990). Partitioning Sparse Matrices with Eigenvectors of Graphs. *SIAM Journal on Matrix Analysis and Applications*, **11**(3), 430.

REICHARDT J. & BORNHOLDT S. (2006). Partitioning and modularity of graphs with arbitrary degree distribution. *Physical Review E - Statistical, Nonlinear and Soft Matter Physics*, **76**(1 Pt 2), 015102.

ROTH C., OBIEDKOY S. & KOURIE D. G. (2008). On succinct representation of knowledge community taxonomies with formal concept analysis. *International Journal of Foundations of Computer Science*, **19**(2), 383.

SUZUKI K. & WAKITA K. (2009). Extracting Multi-facet Community Structure from Bipartite Networks. *2009 International Conference on Computational Science and Engineering*, **4**, 312–319.

WU Z., LIN Y., WAN H., TIAN S. & HU K. (2012). Efficient overlapping community detection in huge real-world networks. *Physica A : Statistical Mechanics and its Applications*, **391**(7), 2475 – 2490.

YANG B., LIU D., LIU J. & FURHT B. (2010). *Discovering communities from Social Networks : Methodologies and Applications* . Boston, MA : Springer US.