Détection de communautés chevauchantes dans les graphes bipartis

Michel Crampes1**, Michel Plantié** [1]

Laboratoire LGI2P, Ecole des Mines d'Ales
{michel.crampes, michel.plantie}@mines-ales.fr

RÉSUMÉ : *La détection de communautés dans les réseaux sociaux est devenue un champ de recherche majeur. La plupart des méthodes, à l'exemple de l'algorithme de Louvain, s'intéressent aux graphes de personnes (graphes monopartis) pour extraire des communautés disjointes. Elles s'appuient de manière globale sur l'optimisation d'un critère appelé modularité. En s'inspirant des treillis de Galois nous présentons une méthode originale de détection de communautés chevauchantes pour les graphes bipartis. Contrairement aux quelques méthodes qui exploitent aussi des treillis de Galois notre approche est de complexité polynomiale et prend en compte en final tous les individus. Dans le même esprit que la modularité nous proposons un critère de regroupement, l'autonomie, qui combine la cohésion interne d'une communauté et son indépendance des autres communautés. Plusieurs exemples illustrent notre approche et en montrent la validité sémantique et pragmatique.*

MOTS-CLÉS: *détection de communautés, réseaux sociaux, acquisition de connaissances, treillis de Galois, hypergraphe.*

1. Introduction

L'usage de la diffusion d'information dans les réseaux sociaux (textes, photos, vidéos, etc.) est très sensible et en même temps très stratégique. Les exemples abondent comme dans la diffusion d'articles scientifiques pour des groupes d'intérêts de chercheurs, la diffusion personnalisée de photos sur Facebook, la diffusion d'informations (tweets, personnes, favoris, etc.) sur Twitter, les applications personnalisées de recommandation, la détection de tendances, etc. Les enjeux scientifiques et industriels importants associés aux réseaux sociaux ont conduit un grand nombre de chercheurs à s'intéresser au problème déjà ancien de l'extraction automatique de communautés. La grande majorité des méthodes proposées cherchent à effectuer une partition des entités à partir du graphe initial des relations observées. La sémantique de ces relations est rarement considérée. Le manque d'information sur les éléments qui rapprochent ou éloignent les individus en est bien sûr la cause principale. Dans cet article, nous portons notre intérêt sur deux problèmes peu traités, d'une part la détection de communautés dans des graphes

[1] Les deux auteurs ont contribué à part égale à cet article. Par ailleurs, Dorian Bosatelli a participé à la programmation des algorithmes.

bipartis, et d'autre part la détection de communautés chevauchantes. Les graphes bipartis ont l'ensemble de leurs sommets partitionnés en deux sous-ensembles. Les arêtes ne relient que des sommets n'appartenant pas au même sous-ensemble. Un exemple de ce type de graphe est constitué d'un sous-ensemble de personnes et d'un sous-ensemble de propriétés associées à ces personnes. Ces propriétés confèrent au graphe une dimension sémantique. Pour bien appréhender les possibilités sémantiques nouvelles de ces structures, nous utilisons les propriétés des treillis de Galois, en évitant cependant la complexité importante liée à ces treillis. Les communautés chevauchantes sont, du point de vue sémantique, très intéressantes et se représentent aisément par le concept d'hypergraphe. Tout individu dans la réalité, appartient à plusieurs communautés reflétant leurs préférences sémantiques, et les hypergraphes ou communautés chevauchantes représentent bien cet état de fait.

2. Etat de l'art

La détection de communautés a fait l'objet de nombreux travaux de recherche, et quatre documents constituent des états de l'art complets de ce domaine : (Papadopoulos et al., 2011) (Porter et al., 2009) (Yang et al., 2010) (Fortunato 2009). Ils sont majoritairement focalisés sur la partition de graphes monopartis. Chaque personne appartient au final à une seule communauté. Les calculs se basent sur la maximisation d'un critère mathématique comme la modularité de (Newman and Girvan 2004). Elle représente le fait qu'il existe dans la communauté à l'étude un nombre maximum de liaisons à l'intérieur et un nombre minimum de liaisons des nœuds avec les communautés extérieures.

Différentes méthodes ont été établies pour trouver l'optimum comme plusieurs algorithmes gloutons proposés par (Newman and Girvan 2004), (Noack and Rotta 2008), ou bien une méthode d'analyse spectrale (Newman 2006), ou encore une recherche des arcs les plus centraux (Newman and Girvan 2004). L'algorithme le plus connu du fait de son efficacité dans le traitement de grands graphes pour la partition en communautés est celui de Louvain (Blondel et al., 2008). Une variante traite les graphes possédant des poids sur ses arcs. Dans son état de l'art très exhaustif (Fortunato 2009) décrit des méthodes plus récentes de détection de communautés fondées sur le partitionnement. Le partitionnement des communautés, bien que séduisant mathématiquement, n'est pas satisfaisant pour décrire correctement la réalité. Chaque individu a « plusieurs vies » et appartient généralement à plusieurs communautés en fonction de ses activités familiales, professionnelles, et autres. Par exemple un chercheur peut être affilié à plusieurs communautés partiellement recouvrantes en fonction de ses domaines scientifiques, de ses centres d'intérêts et de ses activités de veille scientifique. Certains travaux considèrent cet aspect en proposant une réponse sous forme d'hypergraphe de communautés comme (Estrada and Rodriguez-Velazquez 2005).

Prise en compte sémantique : la grande majorité des techniques de détection de communautés dans les graphes ne s'intéresse pas à la raison des relations entre deux nœuds. Pour cela il faut disposer de graphes bipartis ou multipartis, c'est-à-dire des

graphes dont l'ensemble des nœuds est réparti en plusieurs sous-ensembles disjoints et dont les arcs ne relient que des nœuds de sous-ensembles différents. Un exemple de ce type de graphe est l'ensemble des photos d'un compte facebook avec leurs « tags » respectifs (Plantié and Crampes 2010) ou encore celui des réseaux épistémiques tripartis de (Roth and Bourgine 2005) reliant des chercheurs avec leurs publications et les mots clés de ces publications. L'extraction de communautés est souvent effectuée par la conversion d'un graphe multipartis vers un graphe monoparti, en attribuant un lien entre deux nœuds s'ils partagent une propriété commune. (Guimerà, Sales-Pardo, and Amaral 2007) attribue une mesure de modularité aux graphes bipartis utilisant même une pondération en fonction du nombre de propriétés communes et peut ainsi se ramener à un problème de partitionnement classique de graphe. Cependant plusieurs chercheurs conservent les propriétés des graphes multipartis en étendant la notion de modularité à ces types de graphes (Newman and Girvan 2004) ou en adaptant des algorithmes initialement prévus pour les graphes monopartis (Suzuki and Wakita 2009) (Neubauer & Obermayer, 2009) (Barber 2007) (Murata 2009). Les travaux sur les méthodes de clustering recouvrant apportent des éclairages les études en cours actuellement voir (Ganti, Gehrke, and Ramakrishnan 1999; Palla et al. 2005; Cleuziou 2008). Une autre approche utilise les treillis de Galois pour obtenir des résultats sémantiquement plus riches. Les premiers travaux effectués par (Freeman and White 1993) sont les précurseurs de plusieurs contributions en la matière comme nous allons le voir dans la section suivante.

3. Treillis de Galois et graphes bipartis

3.1. Définitions

L'analyse formelle de concepts (AFC) définit un treillis de Galois ou treillis de concepts comme l'organisation d'un ensemble d'objets associés à des propriétés aussi appelées attributs (Ganter B. and Wille R. 1999). C'est un graphe hiérarchisé dont les nœuds, appelés concepts, sont les regroupements d'objets qui partagent les mêmes propriétés, et dont les arêtes représentent une relation d'ordre entre les concepts en fonction de cette fonction d'appartenance des propriétés. L'ensemble des objets d'un concept est l'extension, et l'ensemble des propriétés partagées est l'intension du concept. Pour une expression formelle le lecteur pourra se référer à (Crampes et al., 2011b). Le caractère dual d'un treillis de concept permet de faire basculer les rôles. Les propriétés deviennent alors les objets et les objets deviennent les propriétés. Le treillis de Galois correspond au maximum d'information sur un possible regroupement des individus en communautés. En effet chaque extension peut être considérée comme une communauté puisqu'elle contient un groupe d'individus qui possèdent exactement le même sous-ensemble de propriétés (Freeman and White 1993). Cela correspond à l'agencement le plus fin de communautés. Chaque groupe correspondant à un concept ainsi constitué est parfaitement défini au niveau sémantique par l'intension du concept considéré. Nous obtenons ainsi un hypergraphe constitué de toutes les extensions non vides dont les hyper-arêtes sont sémantiquement définies par les intensions associées. Toutefois

dans la pratique le nombre de concept devient très rapidement élevé en rapport exponentiel au nombre d'attributs ou d'objets. Nous devrons donc réduire le nombre de concepts en déterminant des critères de sélection de ces concepts.

3.2 Exemple d'un cas réel et méthodologie expérimentale

L'exemple que nous utilisons comme fil conducteur présenté dans (Crampes and Plantié 2011) et reproduit dans la figure 1 est structuré autour de 145 photos prises

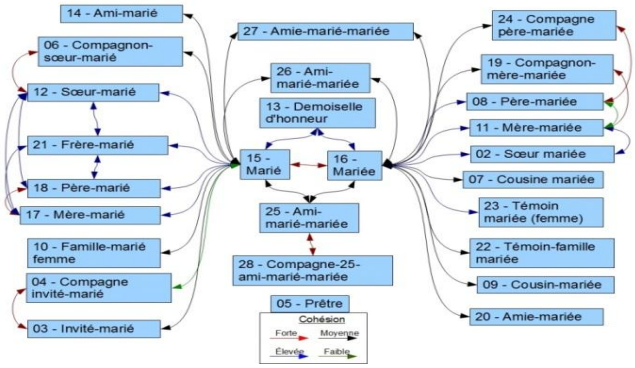

Figure 1. *Le graphe référent du mariage*

lors d'un mariage constitué de 27 personnes. Nous n'avons retenu qu'une photo pour un même groupe de personne en prenant la meilleure. Chaque photo a donc un contenu différent. Le tableau photos/personnes constitue donc le contexte au sens de l'AFC. La figure 1 représente le graphe construit manuellement que nous considérons comme référent sémantique. Il a été tracé en utilisant les informations disponibles sur les rôles et relations civiles des personnes relativement aux mariés. Il détient l'information sociale des personnes. Le déroulement de la cérémonie est exprimé par le témoignage des photos qui nous permettent de construire un contexte dont les objets sont les individus et les photos les propriétés. Nous nommons ce graphe « référent » car il constitue pour le groupe de personnes étudié la référence élaborée par des experts humains et toutes relations établies par calcul entre les personnes du groupe pourront être jugée à leur proximité par rapport au groupe référent. A partir de ce graphe biparti nous avons produit un treillis de Galois de 184 concepts à l'aide de l'outil Galicia (Crampes and Plantié 2011). Les extensions du treillis obtenu ne fournissent pas un ensemble de communautés satisfaisant contrairement à (Freeman and White 1993). En effet les communautés sont toutes de petite taille et en très grand nombre, ce qui ne correspond pas à une réalité sémantique évidente pour un expert humain. Nous devons réduire le nombre de communautés en utilisant des méthodes qui diminuent la précision sémantique mais renforce par exemple la solidarité à l'intérieur des communautés obtenues. Peu de chercheurs ont abordé ce problème. (Roth and Bourgine 2005) recherchent des communautés épistémiques ou groupes disjoints de chercheurs partageant les mêmes connaissances. Ils proposent 5 critères pour réduire le nombre de concepts du treillis de Galois. Ils obtiennent ainsi des communautés de taille significative (extension) portant une sémantique intéressante (intension). La taille significative, se définit en

rapport au jugement d'un expert humain pour en observant le groupe obtenu pourrait en déduire qu'un tel groupe peut correspondre à une réalité observée dans la nature. Sémantique intéressante est une notion qui se rapporte au fait qu'un groupe de personne constitué par l'algorithme partage des éléments sémantiques communs comme des photos communes, des documents communs etc. Cependant ils confessent eux-mêmes que les heuristiques mises en œuvre nécessitent d'être validées par une analyse extérieure. Dans une version ultérieure pour pallier aux défauts des heuristiques précédentes (Roth et al., 2008) utilisent des méthodes connues dans la communauté AFC afin de réduire le nombre de concepts, comme (Stumme G. et al. 2002) qui recherchent des concepts dits « iceberg » ayant une intension fréquente, supérieure à un seuil. Cependant cette méthode peut laisser de côté des concepts intéressants mais à faible support comme l'observent (Jay et al. 2008). (Kuznetsov 2007) propose une méthode moins empirique fondée sur la notion de stabilité des concepts. Mais elle présente quelques limitations. D'une part des seuils doivent être fixés, d'où une part arbitraire ; d'autre part les calculs du treillis initial et de la stabilité pour chaque concept qui nécessitent le dénombrement de l'ensemble des parties de chaque extension sont de type NP-complet. Une heuristique a été présentée dans (Roth, Obiedkoy, and Kourie 2008) pour lever le second niveau de complexité. Il faut cependant toujours calculer le treillis qui est potentiellement de taille exponentielle pour ensuite identifier les concepts de plus forte stabilité. Une autre objection importante porte sur le fait que certains objets vont être ignorés en ne retenant que les concepts au-dessus des seuils. Ce manquement n'est pas acceptable dans de nombreux cas où tous les individus doivent faire partie en final d'au moins une communauté. Dans la section suivante nous introduisons une méthode qui s'inspire des treillis de Galois mais qui ne calcule qu'une partie de celui-ci de manière polynomiale et qui classe tous les objets dans des communautés qui se chevauchent. La méthode proposée répond ainsi à deux problèmes ci-dessus : elle est de complexité polynomiale, et tous les individus sont classés dans au moins une communauté. Cependant la détermination d'un seuil est toujours nécessaire.

4. Modularité d'un hypergraphe

4.1. Hypergraphes de communautés initiales

La première phase consiste à identifier dans un treillis de Galois un hypergraphe de communautés initiales déjà représentatives. Le concept « top » est écarté parce que son intension est vide. Dans le cas du mariage on peut artificiellement le rendre vide en supprimant une photo de tout le mariage qui n'aide pas à la constitution de communautés. La première couche supérieure du treillis est composée des concepts qui ont pour unique parent ce top. Dans notre cas, comme chaque photo est distincte par filtrage initial, ces concepts ont une intension ne comptant qu'une seule photo. L'union des extensions de ces concepts redonne nécessairement l'ensemble des individus. Ces concepts dotés d'une seule propriété sont donc un premier niveau de découpage en communautés qui est le plus fédérateur puisque l'extension d'un concept inférieur sera toujours un sous-ensemble de l'extension d'au moins un

d'entre eux. Sur l'exemple du mariage, nous diminuons à 23 le nombre de communautés obtenues comparativement aux 184 concepts du treillis complet. Ce nombre est majoré par le nombre de photos (cas dégénéré où il y une photo par individu). Ces concepts de première couche sont facilement accessibles sans calculer tout le treillis. Nous évitons ainsi la complexité exponentielle du nombre total de concepts du treillis. Ce premier hypergraphe des communautés peut paraître trivial et sémantiquement limité. Mais il est déjà significatif puisqu'il permet de couvrir à minima l'ensemble des individus et à les séparer en groupes en prenant en compte au moins un attribut. Le choix de ne retenir que le premier niveau est avantageux dans la mesure où il est simple. Mais il présente de sérieux inconvénients. Il peut y avoir de nombreux concepts au premier niveau, au pire autant que de sous-ensembles des individus. Par ailleurs on ne tient pas compte des attributs qui n'ont pas été utilisés pour réaliser ce premier niveau. Enfin il peut induire des communautés qui éventuellement se recouvrent fortement comme par exemple dans le cas où il n'y a que deux individus différents entre deux photos retenues au premier niveau.

4.2. Affinage de la détection de communautés partitionnées

Nous avons proposé dans (Crampes and Plantié 2012) une heuristique simple pour répondre partiellement au premier point, à savoir réduire le nombre de communautés détectées au premier niveau. L'idée directrice repose sur l'hypothèse *pragmatique* que deux communautés qui ont beaucoup d'individus en commun et peu d'individus spécifiques sont susceptibles de fusionner. Cette hypothèse renvoie à beaucoup de stratégies de fusion d'entreprise ou de collectifs d'individus. A cette fin, nous calculons l'indice de Jaccard pour tous les couples constitués par appariement des extensions du premier niveau. C'est le rapport entre le nombre d'individus communs rapporté au nombre d'individus total pour chaque couple [a,b] : $J(Ga, Gb) = |Ga \cap Gb\ |/|Ga \cup Gb\ |$. S'il existe au départ n communautés initiales (concepts de premier niveau) le nombre de paires de communautés susceptibles de fusionner que nous obtenons est n (n-1)/2. Nous nous limitons à celles qui donnent le plus fort indice de Jaccard avec pour objectif le recouvrement de la population d'individus. Pour le mariage, nous obtenons ainsi un hypergraphe de 11 communautés soit environ deux fois moins qu'initialement, chacune décrite par deux photos. Les calculs utilisés sont polynomiaux et simples car ils peuvent se passer de la construction du treillis de Galois en se limitant à l'extraction de la première couche. Cependant cette heuristique simple qui s'appuie sur une logique pragmatique (unir les communautés qui se recouvrent le plus) présente les limites suivantes. Les autres attributs ne sont pas pris en compte. Des individus sont unis alors que rien ne justifie leur union sauf leur présence commune avec d'autres individus communs à deux concepts. Le premier point a été traité en partie dans (Crampes and Plantié 2012) en faisant intervenir tous les attributs pour donner du poids aux relations entre individus. Un algorithme de partitionnement peut ensuite être appliqué, différent de celui de Louvain, mais avec la même philosophie : partitionner l'ensemble de tous les individus. L'idée de partitionnement a été

critiquée dans la section 4 où nous avons mis en avant l'intérêt de produire des hypergraphes communautaires. Dans la suite de la contribution nous nous proposons de produire en sortie un hypergraphe en cherchant à maximiser une fonction collective de constitution de communautés. Cette fonction doit chercher à conserver ensemble les individus ayant le plus de traits communs et à séparer au maximum ceux qui en ont le moins. A cette fin nous introduisons une fonction de cohésion d'une communauté et une fonction de séparation entre les communautés. La cohésion représente la difficulté à rompre la proximité sémantique qui unit une communauté. La séparation représente l'aptitude à se différencier des autres communautés. La combinaison de ces deux critères trouve un écho dans la modularité du partitionnement d'un graphe entre plusieurs communautés mais appliquée ici à des hypergraphes. La section suivante définit la h-modularité c'est-à-dire la modularité de communautés recouvrantes.

5. Recherche de communautés recouvrantes

Pour repérer des communautés, nous sélectionnons d'abord les concepts de premier niveau, c'est-à-dire ceux déterminés par une seule propriété. Ils sont calculables de manière polynomiale. L'étape suivante consiste à retenir parmi ces concepts candidats ceux qui sont les plus à même de former des communautés. A l'image de la modularité pour générer des partitions de graphes nous calculons la h-modularité en considérant dans quelle mesure la cohésion et la séparation de chaque concept candidat contribue à la h-modularité.

5.1. La cohésion d'une communauté

La cohésion d'une communauté est l'aptitude de ses membres à se retrouver ensemble sur des propriétés partagées. Empiriquement, si les propriétés sont les photos, on cherche pour une communauté donnée les photos qui renforcent la cooccurrence des membres de cette communauté. Dans la suite nous décrivons les modalités de calcul de la cohésion en faisant appel à cet exemple. Soit une communauté candidate a où g_a est l'ensemble des individus qui la composent. Sémantiquement nous cherchons toutes les photos qui montrent un sous-groupe de g_a. En prenant l'indice de Jaccard d'un couple de photos de ce type, limité aux sous-groupes de g_a qu'elles montrent, on mesure le chevauchement de ces sous-groupes de g_a. Pour tous les couples de photos concernées, on a ainsi une mesure de la solidarité des sous-groupes de g_a et donc de g_a dans son ensemble. Soit $\{p_b\}$ l'ensemble des photos qui contiennent au moins un membre de g_a et g_b les personnes présentes sur une photo p_b. Nous posons $g_{a,b} = g_a \cap g_b$. Pour deux photos b et c qui portent chacune un sous ensemble de g_a , l'indice de Jaccard permet de mesurer le recouvrement des deux sous-ensembles : $J_{a,b,c} = |g_{a,b} \cap g_{a,c}| / |g_{a,b} \cup g_{a,c}|$. Nous définissons la cohésion de a :

$$\textbf{Cohesion(a)} = \Sigma_{b,c|b<c} |g_{a,b} \cap g_{a,c}| / |g_{a,b} \cup g_{a,c}| / (n\,(n-1)/2),$$

où n est le nombre de photos telles que $g_{a,b} \neq \varnothing$.

5.2. La séparation d'une communauté

Nous mesurons ici dans quelle mesure une communauté est séparée des autres. C'est la mesure opposée du recouvrement de cette communauté par les autres communautés. Pour l'illustrer, nous prenons ici encore l'exemple du mariage avec les personnes et les photos. Si on considère deux communautés leur recouvrement peut être observé par les photos qui contiennent une partie des personnes communes aux deux communautés. Soit a et b les deux communautés dont on veut mesurer le recouvrement : $g_{a,b} = g_a \cap g_b$.

En extrapolant l'indice de Jaccard, une photo c qui représente une partie de ce recouvrement donne une mesure : $K_{a,b,c} = |g_{a,b} \cap g_c| / |g_a \cup g_b|$. Sémantiquement cet indice peut s'interpréter comme la contribution de la photo c à l'observation de l'intersection entre a et b. L'indice de séparation des communautés a et b est l'opposé de la somme des contributions de toutes les photos en nombre n pour lesquelles $g_{a,b} \cap g_c \neq \varnothing$: Séparation(a,b) = $1 - \Sigma_{c|\ ga,b \cap gc \neq \varnothing} |g_{a,b} \cap g_c| / |g_a \cup g_b| / n$. Il est alors possible de donner un indice de séparation de la communauté *a* par rapport à toutes les autres communautés :

Séparation(a) = $\Sigma_{b \neq a}$ Séparation(a,b) /m,

où m est le nombre de communautés candidates.

5.3. L'autonomie d'une communauté

La cohésion fournit un indice de la capacité d'une communauté à ne pas s'émietter indépendamment des autres communautés. La séparation donne sa capacité à conserver ses membres vis-à-vis des autres communautés. Nous retrouvons l'équivalent des deux termes de la modularité présentée dans l'équation 1 de l'article de (Blondel et al., 2008). Mais alors que dans cette équation on mesurait la modularité d'ensemble du graphe monoparti, ici nous sommes intéressés par la capacité de chaque communauté à rester unie. Nous définissons l'autonomie d'une communauté a par rapport aux autres communautés candidates comme suit :

Autonomie(a) = Cohesion(a) x Séparation(a)

Ainsi toutes les communautés candidates sont dotées d'un indice d'autonomie par rapport aux autres. L'étape suivante va consister à sélectionner les meilleurs candidats et à réaffecter les communautés qui auront échoué.

5.4. Algorithme de choix des communautés

Deux méthodes sont possibles : asynchrone et synchrone. Nous présentons ici la méthode la plus simple mais la moins précise, la méthode asynchrone :

1) Calculer le premier niveau du treillis de Galois : pour chaque attribut regrouper les objets qui le possèdent et vérifier que cet ensemble n'est pas un sous-ensemble d'un attribut sélectionné. Si cette extension est originale, conserver le concept comme étant de premier niveau.

2) Calculer le second niveau du treillis de Galois : pour chaque couple de concepts de premier niveau, calculer l'intersection des extensions et créer autant de concepts enfants que d'intersections non vides.
3) Pour tous les concepts de premier niveau, calculer la cohésion, la séparation puis l'autonomie.
4) Créer une liste de communautés sélectionnées, et y placer séquentiellement les concepts qui ajoutent de nouveaux membres par ordre décroissant d'autonomie jusqu'à la prise en compte de tous les objets.
5) Sélectionner dans la liste, les communautés à autonomie plus élevée qu'un seuil S défini. Ces communautés élues forment le noyau des communautés finales (variante : sélectionner les N premières communautés).
6) Pour chaque communauté non élue, la fusionner à celle des communautés élues qui en est le moins séparée.

Lors de l'étape 5, plusieurs stratégies de choix et de réaffectation sont possibles. Par exemple soit on retient les communautés dont l'autonomie est au-dessus d'un certain seuil, soit on retient un nombre prédéfini de communautés. Les deux solutions peuvent se justifier par des critères pragmatiques. Un seuil intéressant est celui de 50% sur l'autonomie pour retenir la liste des communautés formant le noyau des communautés finales. Ce seuil signifie que les membres des communautés en question sont à la fois peu séparés et peu dispersés dans d'autres communautés. Pour l'exemple du mariage, nous retenons les deux possibilités. Pour la deuxième solution, nous retenons un nombre N de communautés (dans l'expérimentation ci-dessous, ce nombre sera fixé à 5). Les N communautés les plus autonomes sont déclarées 'élues'. Les communautés restantes sont affectées chacune en entier à une communauté au plus. Cette politique est discutable et fera l'objet d'une analyse pragmatique dans un prochain article. Pour réaffecter une communauté non élue, on calcule sa séparation avec toutes les communautés élues et on l'affecte à celle avec laquelle elle est le moins séparée. Si une communauté non-élue est totalement séparée de toutes les autres communautés élues, nous ne disposons pas de critère de réaffectation. En conséquence elle devient élue et soit une communauté déjà élue doit être réaffectée, soit le nombre de communautés devient supérieur à N.

6. Expérimentation

Nous avons expérimenté nos méthodes sur trois exemples différents : le mariage déjà évoqué ci-dessus, et deux comptes « Facebook » de personnes possédant un nombre conséquent de photos étiquetées avec des personnes. Les comptes facebook donnent une expérimentation plus proche de la réalité. Le premier compte comporte 305 photos portant sur 88 personnes, le deuxième compte possède 644 photos portant sur 274 personnes.

Communauté	Membres	Cohésion	Séparation	Autonomie finale
1	3 - 4	1	1	1
2	8 - 24	0,78	0,98	0,77
4	16 - 20 - 22	0,63	0,95	0,6
6	25 - 28	0,57	0,98	0,55
3	6 - 10 - 12 - 14 - 16 - 21 - 22 - 23	0,35	0,94	0,33
5	2 - 5 - 12 - 14 - 15 - 16 - 21 - 24 - 28	0,31	0,94	0,29
7	2 - 5 - 7 - 8 - 9 - 10 - 11 - 12 - 13 - 14 - 15 - 16 - 17 - 18 - 19 - 21 - 22 - 23 - 26 - 27	0,14	0,96	0,13

Tableau 1. *Les communautés détectées du mariage*

6.1. Analyse des résultats du mariage

En fixant le seuil S à 0,5 nous obtenons 7 communautés pour le mariage. Le tableau 1 ci-dessus montre les communautés finales dont la valeur d'autonomie résultante diffère de celle utilisée pour le seuil et est quelquefois inférieure à 0,5 parce que certaines communautés en fusionnant voient leur autonomie baisser.

6.2. Analyse des résultats des comptes Facebook

Dans le tableau 2-a est listé le résultat obtenu sur le premier compte facebook (FB 1), lorsque nous choisissons un seuil d'autonomie de 0,7. Les autonomies des communautés de ce compte facebook sont plutôt élevées, provenant du fait que les communautés sont assez séparées : les personnes appartiennent à des cercles différents qui ne se connaissent que par l'intermédiaire du propriétaire du compte.

Concernant le deuxième compte facebook (FB 2) dans le tableau 2-b ci-dessus,

Communauté compte FB 1	Nombre de membres	Cohésion	Séparation	Autonomie
1	6	0,94	0,94	0,89
2	9	0,75	0,94	0,71
3	9	0,76	0,94	0,71
4	42	0,7	0,98	0,7
5	11	0,69	0,94	0,65
6	13	0,66	0,94	0,63
7	12	0,66	0,94	0,62
8	33	0,61	0,96	0,59

Seuil	0,50	0,60	0,70	0,80	0,90	0,95
Nombre de communautés compte FB 2	83	74	59	46	27	13

Tableau 2 a. b *Communautés de comptes Facebook*

les tailles des communautés sont importantes. Nous avons mentionné le nombre des communautés obtenues en fonction du seuil d'autonomie choisi. Pour obtenir un faible nombre de communautés, le seuil devient très élevé.

6.3. Comparaison des résultats

Les valeurs d'autonomies sur les communautés du mariage sont bien plus faibles que sur les communautés des comptes facebook. On peut l'interpréter par le fait que les personnes du mariage se connaissent pratiquement toutes entre elles, alors que sur les comptes facebook des groupes provenant de différents horizons et intérêts sont présents. Les autonomies associées sont donc plus importantes montrant un cloisonnement plus élevé entre les différents groupes. Ce simple résultat est déjà intéressant : arriver à montrer le cloisonnement des communautés. Il est difficile d'évaluer pratiquement les résultats sur les communautés. Il n'existe pas de benchmark permettant de valider les résultats obtenus. Les seules références possibles doivent être données par les experts, si tant est que cela est possible. En effet il est très envisageable que l'on ne trouve pas de consensus viable pour définir les bonnes communautés dans un groupe de personnes. Concernant l'exemple du mariage, nous avons un graphe référent qui peut servir de base. L'algorithme de Louvain qui effectue une partition du graphe donne un résultat singulier sur le graphe du mariage, en affectant à des communautés différentes le marié et la mariée, alors que le prêtre se trouve dans une communauté ou il est le seul membre. Nos algorithmes donnent un résultat bien différent : le prêtre est inclus dans une grande communauté autour des deux mariés. Des petites communautés isolées reflètent les situations exprimées par les photos, où des amis des mariés sont souvent pris

ensemble et peu avec le reste de la noce. Les communautés élues sont souvent les plus petites : cela s'explique par la facilité d'obtenir un haut degré d'autonomie avec un plus faible nombre de personne. Elles sont plus difficiles à fractionner, et servent de noyau auxquelles les communautés moins autonomes se rattachent pour se renforcer. Un autre point à considérer est l'efficacité de la mesure de cohésion à exprimer la ressemblance des individus deux à deux. Dans de futurs travaux, il pourrait être intéressant de considérer la notion d'« air de famille » de Wittgenstein pour comparer les résultats avec notre calcul de cohésion. Un point important à considérer également est la difficulté de visualisation des communautés. Lorsqu'il s'agit d'une partition la visualisation est du même niveau de difficulté que la visualisation de graphes, les auteurs s'employant à colorier les nœuds en fonction de l'appartenance à une communauté unique. Dans les cas de communautés chevauchantes la visualisation devient plus délicate ; aucune méthode ne permet de visualiser un hypergraphe dans un cas général de manière pertinente. En termes de performances on peut observer que sur le cas le plus lourd (FB-2) le temps de calcul reste acceptable (de l'ordre de la minute), reflétant le caractère non exponentiel des algorithmes. Ces temps de calcul raisonnables proviennent du choix de limiter le calcul du treillis au premier niveau. Il pourrait être intéressant de considérer d'autres niveaux du treillis dans les calculs, mais la complexité devient importante dès le deuxième niveau. Pour cette raison nous limitons le calcul au 1er niveau du treillis.

7. Conclusion

Nous avons présenté une méthode de détection de communautés chevauchantes à partir de graphes bipartis. C'est une méthode originale partant d'un hypergraphe pour obtenir un nouvel hypergraphe de communautés. Comparativement à d'autres auteurs, elle a l'avantage de prendre en compte la sémantique (c'est-à-dire le partage de propriétés communes) et permet une flexibilité pragmatique en autorisant le choix de différentes stratégies d'agrégation en fonction de critères de ressemblance entre communautés plus ou moins importants. Elle ouvre de nouvelles perspectives en proposant d'aller au-delà de la simple partition du graphe initial et d'offrir une variation dans les stratégies d'affectation des personnes en fonction des propriétés partagées. Elle s'inspire des treillis de Galois sans toutefois tomber dans une complexité exponentielle observée chez certains auteurs. De plus tous les individus trouvent au moins une communauté. Les expériences menées donnent des résultats intéressants qui demandent à être prolongés en particulier en étudiant des variantes de l'algorithme plus complexes mais plus précises.

Références

Barber, Michael. 2007. "Modularity and Community Detection in Bipartite Networks." *Physical Review E* 76 (6): 1–9.

Blondel, Vincent D, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. "Fast Unfolding of Communities in Large Networks." *Journal of Statistical Mechanics: Theory and Experiment* 2008 (10) (October 9).

Cleuziou, G. 2008. *An Extended Version of the K-means Method for Overlapping Clustering*. *2008 19th International Conference on Pattern Recognition*. Ieee.

Crampes, Michel, Jeremy de Oliveira-Kumar, Sylvie Ranwez, and Jean Villerd. 2009. "Indexation De Photos Sociales Par Propagation Sur Une Hi{é}rarchie De Concepts." *Actes De La Conférence IC 2009, Hammamet Tunisie*: 13–24.

Crampes, Michel, and Michel Plantié. 2012. “Détection De Communautés Dans Les Graphes Bipartis.” In *IC 2012 Ingénierie Des Connaissance*.

Estrada, Ernesto, and Juan A Rodriguez-Velazquez. 2005. “Complex Networks as Hypergraphs.” *Systems Research*: 16. http://arxiv.org/abs/physics/0505137.

Fortunato, Santo. 2009. “Community Detection in Graphs.” *Physics Reports* 486 (3-5) (June 3): 103.

Freeman, Linton C, and D R White. 1993. “Using Galois Lattices to Represent Network Data.” *Sociological Methodology*,*23*: 127–146.

Ganter B., and Wille R. 1999. *Formal Concept Analysis: Foundations and Applications*. Ed. Springer.

Ganti, Venkatesh, Johannes Gehrke, and Raghu Ramakrishnan. 1999. “CACTUS---clustering Categorical Data Using Summaries.” *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 99*: 73–83.

Guimerà, Roger, Marta Sales-Pardo, and Luís Amaral. 2007. “Module Identification in Bipartite and Directed Networks.” *Physical Review E* 76 (3) (September).

Jay, Nicolas, François Kohler, and Amedeo Napoli. 2008. “Analysis of Social Communities with Iceberg and Stability-based Concept Lattices.” In *ICFCA'08 Proceedings of the 6th International Conference on Formal Concept*, 258–272.

Kuznetsov, Sergei O. 2007. “On Stability of a Formal Concept.” *Annals of Mathematics and Artificial Intelligence* 49 (1-4) (June 6): 101–115. doi:10.1007/s10472-007-9053-6. http://dblp.uni-trier.de/db/journals/amai/amai49.html#Kuznetsov07.

Michel Crampes, and Michel Plantié. 2011. “Méthodes D’extraction De Réseaux Sociaux Et De Diffusion De Photos Sociales.” In *IC 2011 Ingénierie Des Connaissances*.

Murata, Tsuyoshi. 2009. “Modularities for Bipartite Networks.” Ed. Ciro Cattuto, Giancarlo Ruffo, and Filippo Menczer. *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia HT 09* 90 (6): 245–250.

Neubauer Nicolas, and Obermayer Klaus. “Towards Community Detection in k-Partite k-Uniform Hypergraphs.” In *Proceedings NIPS 2009*

Newman, Mark. 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, Mark, and M. Girvan. 2004. “Finding and Evaluating Community Structure in Networks.” *Physical Review E* 69 (2) (February).

Noack, Andreas, and Randolf Rotta. 2008. “Multi-level Algorithms for Modularity Clustering”. Data Structures and Algorithms; Statistical Mechanics; Discrete Mathematics; Physics and Society (December 22): 12.

Palla, Gergely, Imre Derényi, Illés Farkas, and Tamás Vicsek. 2005. “Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society.” *Nature* 435 (7043) (June 9): 814–8.

Papadopoulos, S, Y Kompatsiaris, A Vakali, and P Spyridonos. 2011. “Community Detection in Social Media.” *Data Mining and Knowledge Discovery* (June): 1–40.

Plantié, Michel, and Michel Crampes. 2010. “From Photo Networks to Social Networks, Creation and Use of a Social Network Derived with Photos.” *Proceedings of the ACM International Conference on Multimedia , Firenze, Italy, October*.

Porter, Mason A., Jukka-Pekka Onnela, and Peter J. Mucha. 2009. “Communities in Networks.” *Notices of the American Mathematical Society, Vol. 56, No. 9, 2009*. SSRN.

Roth, Camille, and Paul Bourgine. 2005. “Epistemic Communities: Description and Hierarchic Categorization.” *Mathematical Population Studies: An International Journal of Mathematical Demography* 12 (2): 107–130.

Roth, Camille, Sergei Obiedkoy, and Derrick G Kourie. 2008. “On Succinct Representation of Knowledge Community Taxonomies with Formal Concept Analysis.” *International Journal of Foundations of Computer Science* 19 (2): 383.

STUMME G., TAOUIL R., BASTIDE Y., PASQUIER N., and LAKHAL L. 2002. “Computing Iceberg Concept Lattices with TITANIC .” *Data & Knowledge Engineering* 2 (42): 189–222.

Suzuki, Kenta, and Ken Wakita. 2009. “Extracting Multi-facet Community Structure from Bipartite Networks.” *2009 International Conference on Computational Science and Engineering* 4: 312–319.

Yang, Bo, Dayou Liu, Jiming Liu, and Borko Furht. 2010. *Discovering Communities from Social Networks: Methodologies and Applications* . Ed. Borko Furht. Boston, MA: Springer US. doi:10.1007/978-1-4419-7142-5.