Détection de communautés dans les grands graphes d'interactions (multiplexes) : état de l'art

Rushed Kanawati

Détection de communautés dans les grands graphes d'interactions (multiplexes) : état de l'art

Application aux systèmes de recommandation

Rushed Kanawati
LIPN CNRS UMR 3070, USPC
e-mail : rushed.kanawati@lipn.univ-paris13.fr

Résumé

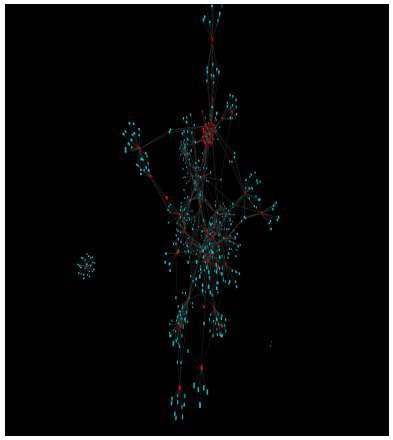
Nous présentons dans ce rapport une brève état de l'art des techniques de détection de communautés dans les grands graphes d'interactions. Nous motivons d'abords l'intérêt de l'étude de cette problématique dans le contexte de systèmes de recommandation. Puis nous passons en revue les principales approches proposées pour traiter ce problème dans le cas des graphes simples, puis dans des graphes mutiplexes qui correspondent mieux au cas des systèmes réels. L'accent est mis aussi dans cette étude sur les différentes approches d'évaluation des communautés détectées par les différents algorithmes.

1 Introduction

Les traces, aujourd'hui souvent informatisées, de certaines activités humaines, peuvent être modélisées sous forme de réseaux complexes. Des exemples sont : les traces d'utilisation de moteurs de recherche sue le web [Baeza-Yates, 2007, Kanawati, 2013b], les traces d'échanges de ressources sur des réseaux pair à pair [Shahabi and Kashani, 2007, Aidouni et al., 2009], les logs de communications électroniques et/ou téléphoniques [Bouveyron and Chipman, 2007, Gutierrez et al., 2013, Guigourès et al., 2013], les interactions dans les réseaux sociaux en-ligne, comme facebook, twitter ou encore LinkedIn [Archambault and Grudin, 2012], les jeux en réseaux et l'immersion dans des mondes virtuels [Szell and Thurner, 2012, Kappe et al., 2009, Shah and Sukthankar, 2011], l'implication dans le blogosphère et autres sites de partage social de ressources (i.e. les folksonomies)

[Papadopoulos et al., 2010, Pujari and Kanawati, 2012] et les collaborations scientifiques [Newman, 2001, Newman, 2004a, Benchettara et al., 2010].

La plupart des graphes obtenus lors de la modélisation des activités cités ci-avant exhibent des propriétés topologiques non-triviales mais similaires à d'autres graphes d'interactions observés dans d'autres contextes et d'autres domaines comme la biologie (ex. réseaux d'interactions entre protéines), les réseaux technologiques (ex. Internet, le web) [Tarissan et al., 2013], et les réseaux de transport (réseaux de routes, réseaux ferrés, réseau d'interconnexion entre aéroports) [Roth et al., 2011, Barthelemy, 2010]. Nous désignons ces graphes, modélisant des systèmes réels, par le nom générique de *graphes de terrain*. La figure 1 illustre deux exemples de ces graphes.



(a) Réseau (bi-partite) d'achats sur un site de e-commerce

(b) Réseau de collaborations scientifiques extrait de DBLP

FIGURE 1: Deux exemples de graphes de terrain

Par souci de simplification, nous considérons, sauf mentionne explicite, des graphes simples, non dirigés et non pondérés. Le tableau 1 donne les principales notations employées dans le reste du document.

Certaines des caractéristiques topologiques des graphes de terrain sont aussi partagées par les graphes aléatoires [Erdös and Rényi, 1959, Watts and Strogats, 1998]. C'est notamment le cas de deux caractéristiques emblématiques suivantes :

La faible densité : La densité d'un graphe est définie par la proportion de liens observés par rapport au nombre de liens potentiels. Dans un graphe simple, composé de n nœuds reliés par m liens la densité est donné par $\frac{m}{n\times(n-1)}$. Les graphes de terrain ont souvent une densité très

TABLE 1: Notations utilisées

Notation	Description
$G =< V, E >$	Graphe non orienté, V : ensemble de nœuds, E : ensemble de liens
$n =\parallel V \parallel$	Nombre de nœuds
$m =\parallel E \parallel$	Nombre de liens
A_G	La matrice d'adjacence du graphe G
$\Gamma(x)$	Ensemble de voisins directes d'un nœud
$d_x =\parallel \Gamma(x) \parallel$	Degré de x
$dist(x, y)$	Distance géodésique entre les nœuds x et y

faible. Nous observons souvent que le nombre de liens est proportionnel au nombre de nœuds. A titre d'exemple les densités des deux graphes illustrés à la figure 1. sont respectivement 2×10^{-2} et 10^{-4}.

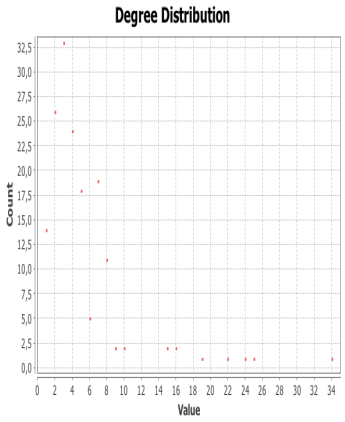
L'effet petit-monde Cette propriété, mise en évidence historiquement par la fameuse expérience de Milgram [Milgram and Travers, 1969], exprime le fait que les graphes de terrain ont souvent des diamètres très faibles. Le diamètre d'un graphe est donné par la longueur de plus long des courts chemins entre chaque couple de nœuds. Les diamètres des graphes pris en exemple ici sont respectivement 8 et 24.

D'autres caractéristiques des graphes de terrain sont bien propres à ce type de graphes et en font un objet d'étude à part entière. Les principales de ces caractéristiques sont :

Distribution hétérogène de degrés : Dans un graphe simple, le degré d'un nœud est donné par le nombre de ses voisins directs. Dans les graphes de terrain on observe souvent qu'il y a certains nœuds qui ont des degrés très élevés et beaucoup d'autres ont des degrés très faibles. La distribution de degrés est souvent décrite par une distribution en loi de puissance de le forme $P(k) = \beta k^{-\gamma}$. où $P(k)$ désigne la probabilité d'un nœuds d'avoir k voisins. Les distributions de degrés de deux graphes pris en exemple sont illustrées à la figure 2.

Un coefficient de clustering local élevé ; Cette propriété exprime le fait que la probabilité que deux nœuds, ayant au moins un voisin commun, soient liés est bien plus grande que la probabilité de lien entre deux nœuds aléatoirement choisis. Par exemple, dans les réseaux sociaux on observe généralement que les amis d'une personne ont tendances à être liés entre eux aussi. Le coefficient de clustering est donné

(a) Distribution de degrés du réseau de co-achat de la figure 1 a)



(b) Distribution de degrés du réseau de collaboration scientifique de la figure 1 b)]

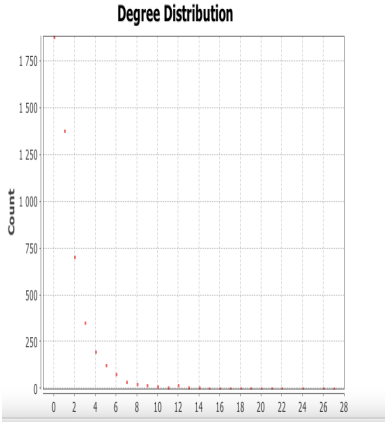


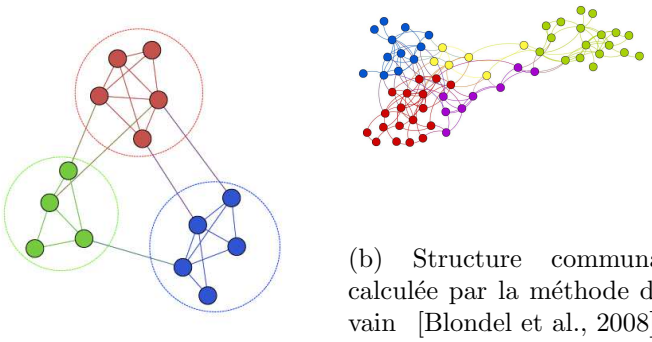
FIGURE 2: Distribution de degrés des graphes illustrés à la figure 1

par la formule suivant :

$$\sum \frac{3 \times \#\triangle}{\#\wedge} \tag{1}$$

où $\#\triangle$ est le nombre de triangles dans le graphe et $\#\wedge$ est le nombre de triades. Noter que dans un graphe aléatoire le coefficient de clustering sera de l'ordre de la probabilité de l'existence d'un lien. Pour les graphes exemples nous avons les coefficients de clustering suivants : 0.84 et 0.67.

Structure communautaire : Une autre caractéristique phare des graphes de terrain est la possibilité de les diviser en modules ou en sous-graphes cohésives et denses et qui sont faiblement connectés entre eux. Une illustration d'un graphe exemple est donné à la figure 3 où on montre à gauche un exemple d'une structure communautaire sur un graphe d'exemple. A droite nous montrons les communautés détectées automatiquement dans un graphe réel modélisant des interactions observées au sein d'un groupe de dauphins dont le comportement a été étudié dans [Lusseau et al., 2003].



(a) Structure communautaire dans un graphe d'exemple

(b) Structure communautaire calculée par la méthode de Louvain [Blondel et al., 2008] dans un graphe d'interaction entre dauphins [Lusseau et al., 2003]

FIGURE 3: Exemple de structures communautaires

Détection de communautés et systèmes de recommandation : Les applications de la détection de communautés sont nombreuses. C'est une étape souvent nécessaire pour nombre d'opérations de traitement de grands graphes notamment pour la visualisation [Bastian et al., 2009], la compression [Hernández and Navarro, 2012] et la parallélisation des calculs. Un autre champ d'applications important qui nous intéresse ici, est le calcul de recommandations. Dans le contexte de réseaux sociaux en ligne, la détection de communautés peut servir pour recommander d'établir de nouveaux liens *d'amitié*, un service fréquemment proposé dans les sites des réseaux sociaux en-ligne. Dans le contexte de réseaux bibliographiques on peut penser à la recommandation de nouvelles collaborations scientifiques [Kanawati, 2013a, Benchettara et al., 2010]. Dans le cadre de réseau d'achats le concept de communauté peut être vu comme une généralisation de l'approche classique de filtrage collaboratif [Resnick et al., 1994] où on peut recommander à une personne les produits bien évalués par les membres de sa communauté. Les produits peuvent être aussi regroupés en communautés selon les motifs de leurs achats ce qui permet de recommander à un client des produits similaires à ce qu'il a aimé au passé.

Dans beaucoup d'applications réelles, les réseaux d'interactions sont des réseaux hétérogènes composés de plusieurs types de nœuds et de liens. Prenons par exemple les réseaux issus des sites de partage social ou les folksonomies (ex. Flicker, citeUlike, bibsonomy) qui font intervenir souvent trois types de nœuds : les utilisateurs, les ressources annotées et les tags utilisés

pour l'annotation [Pujari and Kanawati, 2012]. Les techniques de détection de communautés peuvent servir pour traiter le problème difficile de recommandation de tags [Papadopoulos et al., 2011, Schifanella et al., 2010]. Des réseaux multiplexes sont aussi utilisés pour modéliser des interactions typées comme c'est le cas dans des sites de notation où les consommateurs peuvent évaluer sur une échelle de valeurs des produits proposés (hôtels, restaurants, films, etc.). Chaque niveau d'évaluation peut correspondre à une couche dans un réseau multiplex.

Le reste de ce rapport est organisé comme suit. Dans la section 2 nous passons en revue les principales approches pour la détection de communautés dans les graphes simples. Les différentes approches d'évaluation des communautés détectées sont présentées dans la section 3. Dans la section 4 nous présentons les principales approches d'extension des approches étudiées dans la section 2 pour le cas des graphes multiplexes. Enfin nous concluons dans la section 5.

2 Détection de communautés dans des graphes simples

La ressemblance du problème d'identification de communautés avec beaucoup d'autres problèmes traités dans d'autres domaines, comme le *clustering* de données, le problème de calcul de cut dans des graphes ou encore les problèmes d'optimisation font qu'il existe une grande variété d'approches pour l'identification de communautés. Trois études de synthèse intéressantes mais non-exhaustives sont présentées dans [Fortunato, 2010, Tang and Liu, 2010, Papadopoulos et al., 2012]. Ici, nous proposons de classifier les approches existantes dans quatre classes non exclusives entre elles :

- *Approches centrées groupes* où des nœuds sont regroupés en communautés en fonction de propriétés topologiques partagées.
- *Approches centrées réseau* où la structure globale du réseau est examinée pour la décomposition du graphe en communautés.
- *Approches centrées propagation* qui appliquent souvent une procédure d'émergence de la structure communautaire par échange de messages entre nœuds voisins.
- *Approches centrées graines* où la structure communautaire est construite autour d'un ensemble de nœuds choisis d'une manière informée.

2.1 Approches centrées groupes

Le principe consiste à confondre la définition d'une communauté avec un groupe de nœuds ayant certaines caractéristiques topologiques communes. L'exemple le plus trivial est d'assimiler une communauté à une clique maximale dans le graphe ou à une γ-dense quasi clique. Une clique est un sous-graphe complet. Une clique est maximale si on ne peut l'étendre en ajoutant de nouveaux nœuds. Une γ-dense quasi clique est un sous-graphe dont la densité est supérieur à un certain seuil $\gamma \in [0, 1]$. Or, le problème de calcul de cliques maximales est un problème NP- difficile, ce qui rends difficile d'envisager son utilisation dans le contexte de très grands graphes. Une autre concept utile, souvent employé dans le domaine de l'analyse des réseaux sociaux, est le concept de *K-core*. Un K-core est un sous-graphe connexe maximal dans lequel le degré de chaque nœud est supérieur ou égale à k. les graphes de terrain sont principalement des graphes très parcimonieux, de telle structures sont souvent minoritaire dans les graphes. Par contre de groupements denses de nœuds peuvent servir comme des graines pour la détection des communautés (voir section 2.4).

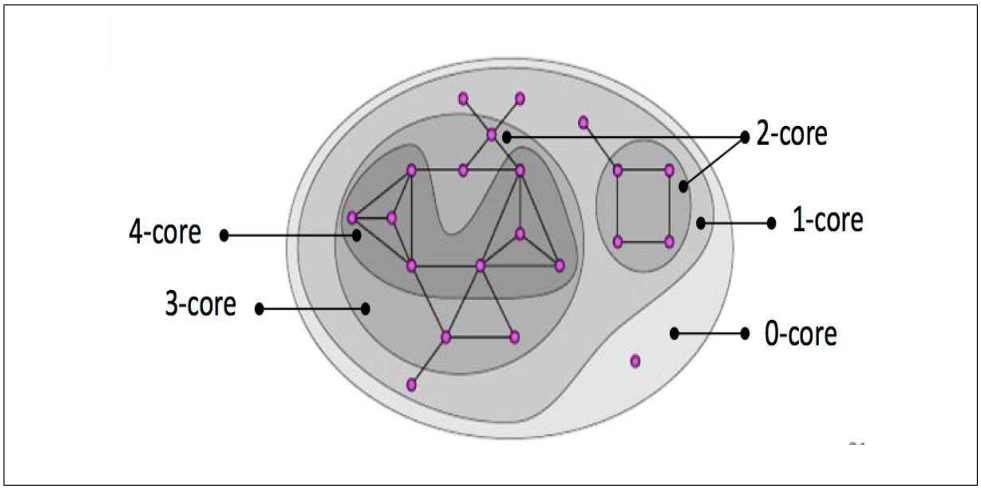


FIGURE 4: Exemple de K-core dans un graphe - Exemple tiré de [Papadopoulos et al., 2012]

2.2 Approches centrées réseau

La majeur partie des approches proposées dans la littérature s'appuient sur un schéma de calcul prenant en compte la connexion globale du graphe

cible. Différentes approches ont été proposées. Nous reprenons dans la suite la classification proposée dans [Tang and Liu, 2010] des approches centrées réseau où on distingue trois familles d'approches :

2.2.1 Approches de clustering

Une approche simple pour la détection de communautés consiste à transformer ce problème en problème classique de clustering de données [Aggarwal and Reddy, 2014]. Etant donné n individus à regrouper en clusters, beaucoup d'algorithmes classiques calculent d'abord une matrice de similarité $\mathcal{S}$ de dimension $n \times n$ où un élément S_{ij} exprime la similarité entre deux individus i et j selon une mesure de similarité donnée. Dans le cas d'un graphe G de n nœuds il est aussi possible de construire une matrice de similarité entre les nœuds du graphe en utilisant une mesure de similarité topologique entre les nœuds du graphe. Différentes mesures de similarité topologiques dyadiques peuvent être définies. Nous les classifions en trois grandes catégories :

- Les mesures basées sur le voisinage des nœuds, dites aussi *mesures locales*.
- Les mesures basées sur les chemins entre les nœuds, dites aussi *mesures globales*.
- Les mesures semi-locales.

Le tableau 2 résume les principales mesures locales les plus utilisées.

Mesures basées sur les chemins Concernant les mesures basées sur les chemins les principales sont :

La proximité : $sim^{proxi}(x,y) = \frac{1}{dist(x,y)}$: Plus la distance géodésique entre deux nœuds est petite plus la proximité des deux nœuds est grande. Or, rappelons qu'une caractéristique phare des graphes de terrain est la faible degré de séparation. Autrement dit, la distance moyenne entre chaque couple de nœuds est faible. Ce qui rends une telle mesure peu discriminant dans beaucoup de situations.

La mesure de Katz : soit $\sigma^l(x,y)$ l'ensemble de chemins de longueur l reliant deux nœuds x et y. La mesure de Katz proposée initialement dans [Katz., 1953] est définie par :

$$sim^{katz}(x,y) = \sum_{l=1}^{\infty} \beta^l \times \parallel \sigma^l(x,y) \parallel \tag{2}$$

TABLE 2: Mesures de similarité dyadiques centrées voisinage

Mesure	*Formule*	Référence
Voisins communs (VC)	$sim^{VC}(x,y) = \parallel \Gamma(x) \cap \Gamma(y) \parallel$	[Lü and Zhou, 2011]
Cosine (ou indice de Salton)	$sim^{cos}(x,y) = \frac{\|\Gamma(x)\cap\Gamma(y)\|}{\sqrt{\|\Gamma(x)\|\times\|\Gamma(y)\|}}$	[Salton and McGill, 1983]
Jaccard	$sim^{Jaccard}(x,y) = \frac{\|\Gamma(x)\cap\Gamma(y)\|}{\|\Gamma(x)\cup\Gamma(y)\|}$	[Jaccard, 1901]
Adamic-Adar (AA)	$sim^{AA}(x,y) = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{log(\|\Gamma(z)\|)}$	[Adamic and Adar, 2003]
Allocation de ressource (RA)	$sim^{RA}(x,y) = \sum_{z\in\Gamma(x)\cap\Gamma(y)} \frac{1}{\|\Gamma(z)\|}$	[Zhou et al., 2009]
Attachement Préférentiel (AP)	$sim^{AP}(x,y) = d_x \times d_y$	[Barabási and Albert, 1999]
Sørensen Index	$sim^{Sørensen}(x,y) = \frac{2\times\|\Gamma(x)\cap\Gamma(y)\|}{\|\Gamma(x)\|+\|\Gamma(y)\|}$	[Sørensen, 1948]
HPI [1]	$sim^{HPI}(x,y) = \frac{\|\Gamma(x)\cap\Gamma(y)\|}{min(\|\Gamma(x)\|,\|\Gamma(y)\|)}$	[Ravasz et al., 2002]
HDI [2]	$sim^{HPI}(x,y) = \frac{\|\Gamma(x)\cap\Gamma(y)\|}{max(\|\Gamma(x)\|,\|\Gamma(y)\|)}$	[Ravasz et al., 2002]

où $\beta << 1$ est un facteur qui va favoriser la prise en compte des chemins de longueurs courtes. Dans [Fouss et al., 2007] on montre que si β est inférieur à la plus grande valeur propre de A_G alors le calcul de cette mesure pour chaque couple de nœuds converge pour les valeurs calculées par la formule matricielle suivante :

$$sim^{Katz} = (I - \beta \times A_G)^{-1} - I \tag{3}$$

où I est la matrice identité. Le calcul de cette mesure est très couteuse pour les grands graphes. En pratique nous nous contentons d'une formule simplifiée comme nous le montrons lors de la discussion des mesures semi-locales.

Indice de Leicht-Holme-Newman (LHN) Cette mesure initialement proposée dans [Leicht et al., 2006] est une variante de la mesure de Katz où on introduit l'idée que des nœuds sont similaires si ils sont connectés à des nœuds similaires. La formulation matricielle récursive de cette mesure est donnée par :

$$sim^{LHN} = \phi A sim^{LHN} + \psi I \tag{4}$$

où ϕ, et $\pm psi$ deux facteurs de pondération des poids des similarités entre voisins par rapport à la similarité directe entre les deux nœuds cibles. Dans [Lü and Zhou, 2011], on montre que le calcul de cette indice peut se résumer au calcul suivant :

$$sim^{LHN} = 2m\lambda_1 D^{-1}(I - \frac{\phi A}{\lambda_1})^{-1} D^{-1} \tag{5}$$

où D est la matrice des degrés du graphe cible ($D_{xy} = \delta_{xy} d_x$). δ_{xy} est la fonction de Kronecker : $\delta_{xx} = 1$ et $\delta_{xy} = 0$ si $x \neq y$.

L'intermédiarité de chemin (PBC) [3] Cette mesure, proposée dans [Pujari, 2013], est une généralisation de la mesure d'intermédiarité de groupe [Kolaczyk et al., 2009]. Soit $\sigma^{dist(x,y)}(x, y)$ l'ensemble de plus courts chemins reliant x et y. D'une manière analogue à l'intermédiarité d'un lien, l'intermédiarité d'un chemin $p \in \sigma^{dist(x,y)}(x, y)$ est donnée par :

$$BC(p) = \sum_{i,j \in V} \frac{\| \sigma^{dist(i,j)}(i, j|p) \|}{\| \sigma^{dist(i,j)}(i, j) \|} \tag{6}$$

où $\sigma^{dist(i,j)}(i, j|p)$ est l'ensemble de plus courts chemins reliant i à j et ayant p comme un sous-chemin. L'intermédiarité de chemin entre

deux nœuds x, y est donnée par :

$$sim^{PBC}(x,y) = \max_{p \in \sigma^{dist(x,y)}(x,y)} BC(p) \tag{7}$$

Le temps de commutation moyen (CT) Le temps de commutation moyen est donné par le nombre d'étapes nécessaires à un marcheur aléatoir qui part de x d'atteindre y puis retourner à x. Dans [Fouss et al., 2007] on montre que le calcul de cette indice est donné par la formule suivante :

$$sim^{CT}(x,y) = \frac{1}{L^+_{xx} + L^+_{yy} + 2L^+_{xy}} \tag{8}$$

où $L^+ = (D - A)^{-1}$ est le pseudo-inverse de la matrice laplacienne du graphe cible.

Indice de forêts de matrices (MFI) Selon cette mesure proposée dans [Chebotarev and Shamis, 1997], la similarité entrée deux nœuds x et y est exprimée par le ratio du nombre de forêts recouvrant tel que x et y sont dans le même arbre recouvrant enraciné dans x sur le nombre total de forêts recouvrant dans le graphe. Dans [Fouss et al., 2007] on montre que le calcul de cette mesure peut être donné par la formule matricielle suivante :

$$sim^{MFI} = (I + L)^{-1} \tag{9}$$

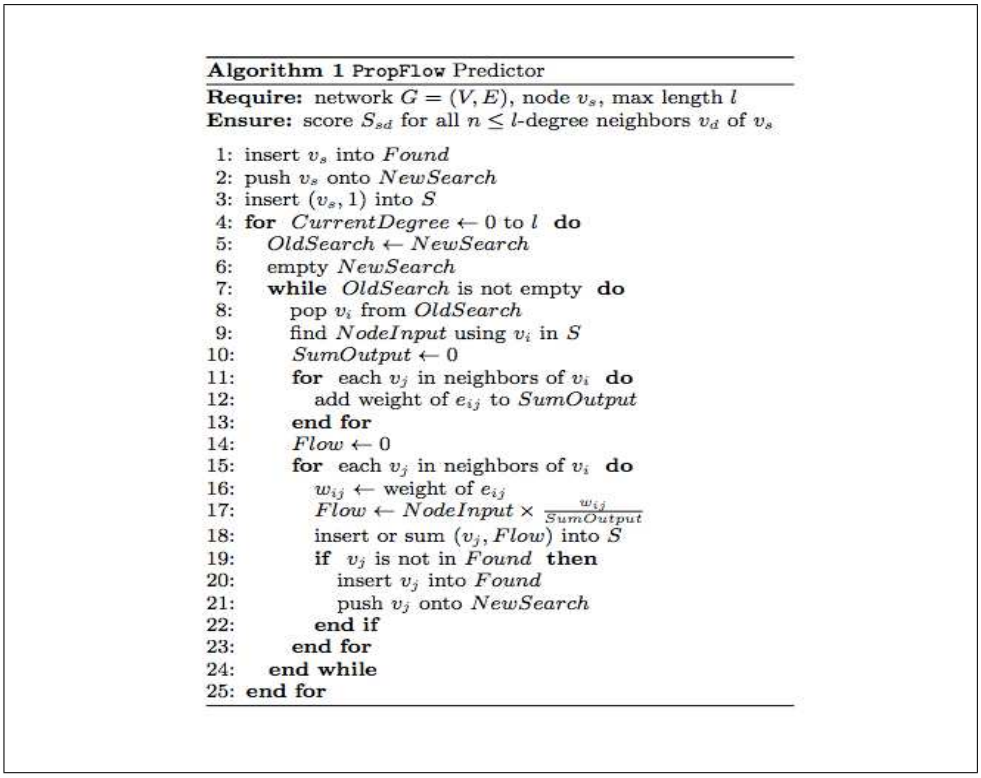
où L est la matrice laplacienne du graphe cible.

Mesures semi-locales Une mesure semi-local vise à réaliser un compromis entre une exploration de la structure du graphe qui dépasse le simple voisinage d'une part, et l'efficacité computationelle d'autre part. Souvent Une mesure semi-locale est dérivée d'une mesure centrée chemin comme c'est le cas par exemple de la mesure tronquée de Katz définie par :

$$sim^{t-katz} = \sum_{l=1}^{l_{max}} \beta^l A^l \tag{10}$$

Une autre mesure similaire est l'indice de chemin local proposée dans [Zhou et al., 2009] et définie par :

$$sim^{LPI} = A^2 + \epsilon A^3 \tag{11}$$



Algorithm 1 `PropFlow` Predictor

Require: network $G = (V, E)$, node v_s, max length l
Ensure: score S_{sd} for all $n \leq l$-degree neighbors v_d of v_s

```
1: insert v_s into Found
2: push v_s onto NewSearch
3: insert (v_s, 1) into S
4: for CurrentDegree ← 0 to l do
5:    OldSearch ← NewSearch
6:    empty NewSearch
7:    while OldSearch is not empty do
8:       pop v_i from OldSearch
9:       find NodeInput using v_i in S
10:      SumOutput ← 0
11:      for each v_j in neighbors of v_i do
12:         add weight of e_ij to SumOutput
13:      end for
14:      Flow ← 0
15:      for each v_j in neighbors of v_i do
16:         w_ij ← weight of e_ij
17:         Flow ← NodeInput × w_ij / SumOutput
18:         insert or sum (v_j, Flow) into S
19:         if v_j is not in Found then
20:            insert v_j into Found
21:            push v_j onto NewSearch
22:         end if
23:      end for
24:   end while
25: end for
```

FIGURE 5: L'algorithme de calcul de la mesure PropFlow - exemple tiré de [Lichtenwalter et al., 2010]

où ϵ est un paramètre à fixer pour pondérer l'apport de nombre de chemins de longueur 3 à la valeur de cette mesure. Noter que si $\epsilon = 0$ cette mesure reviens à calculer le nombre de voisins communs (cf. 2).

Une autre mesure semi-local, nommée *PropFlow*, proposée dans [Lichtenwalter et al., 2010] repose sur l'idée de la propagation bornée entre les deux nœuds cibles. La figure 5 donne l'algorithme de calcul de cette mesure. La similarité entre deux nœuds x et y est assimilée à la probabilité d'un marcheur aléatoire de partir de x et d'arriver à y en l étapes. Les poids des liens du graphe cible sont pris pour être les probabilités de transition d'un nœud vers un autre. Le processus d'exploration applique une stratégie de recherche en largeur, mais l'exploration s'arrête à l'arrivé au nœud cible ou à l'arrivé à un nœud déjà visité (y compris si c'est le nœud x).

Dans [Benchettara et al., 2010], les auteurs font remarquer que beaucoup de réseaux de collaboration réels sont issus de la projection d'un réseau

bipartite sur l'un des deux ensembles le constituant. Par exemple, un réseau de co-achat est la projection sur l'ensemble de clients du graphe bipartite d'achats, reliant les clients aux produits achetés. De nouvelles mesures de similarités entre les clients peuvent être définies d'une manière indirecte en considérant les relations induites sur l'espace de produits. D'une manière plus générale, soit un graphe bipartite G_{bip} défini sur deux ensembles $\top$ et $\perp$. Soit $G_\top$ (reps. $G_\perp$) le graphe projeté sur l'ensemble $\top$ (reps. $\perp$). Une mesure indirecte définie sur les éléments de $\perp$ est donnée par :

$$A_\perp(x,y) = \Phi_{u \in \Gamma_{G_{bip}}(x), v \in \Gamma_{G_{bip}}(y)} A_\top(u,v) \tag{12}$$

où Φ est une fonction d'agrégation. $A_\top$ (reps $A_\perp$) est une mesure dyadique quelconque définie sur le graphe $G_\top$ (reps. $G_\perp$).

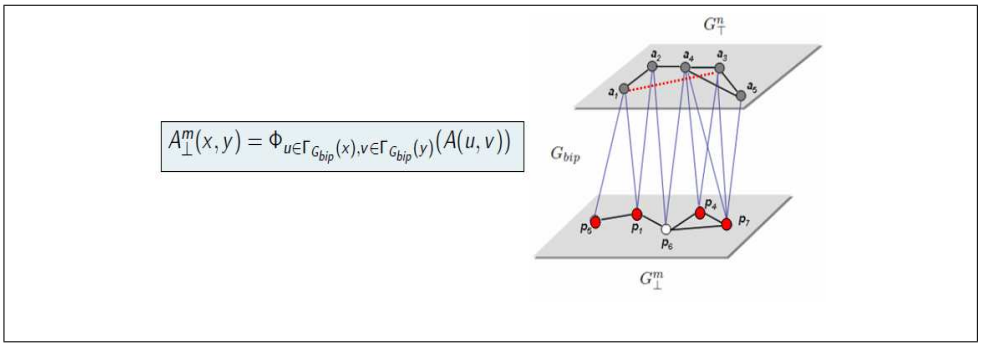


FIGURE 6: Définition de mesures indirectes dans des graphes projetés d'un graphe bipartite

Clustering spectral Un problème similaire au problème de partitionement de graphe est celui de calcul de coupe minimale. Une coupe correspond à la partition de l'ensemble de sommets V d'un graphe en deux ensembles disjoints S et T de sorte que l'ensemble de liens du graphe aient extrémités dans chaque sous-ensemble de la partition. La coupe est minimale si l'ensemble de liens entre les deux sous-ensembles est minimale. Pour éviter le calcul de coupes triviales et peu intéressantes où l'un des sous-ensembles se résume à un singleton, on change souvent la fonction objective de sorte que les tailles des sous-ensemble soient prises en compte. Soit $\pi = (C_1, C_2, \ldots, C_k)$ une partition de l'ensemble des nœuds. Les deux variantes de la fonction objective les plus employées sont [White and Smyth, 2005] :

– Le ratio de la coupe donné par :

$$Ratio\_Cut(\pi) = \sum_{i=1}^{k} \frac{cut(C_i, \overline{C_i})}{\parallel C_i \parallel} \tag{13}$$

où $\overline{C_i}$ est le complément de C_i

– La coupe normalisée donnée par :

$$Ncut = \sum_{i=1}^{k} \frac{cut(C_i, \overline{C_i})}{vol(C_i)} \tag{14}$$

où $vol(C_i) = \sum_{v \in C_i} d_v$

Dans les deux cas, l'optimisation de la fonction objective peut être ramenée à un problème de minimisation de trace formulée par :

$$\min_{S \in \{0,1\}^{n \times k}} Tr(S^T \tilde{L} S) \tag{15}$$

où :

$$\tilde{L} = \begin{cases} D - A & \text{cas de Ratio Cut} \\ I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}} & \text{cas de Ncut} \end{cases} \tag{16}$$

Dans les deux cas, la valeur optimale de S corresponds aux premières vecteurs propres de la matrice $\tilde{L}$ qui sont associés aux plus petites valeurs propres [Tang and Liu, 2010].

2.2.2 Approche fondée sur les modèles de blocks

Le principe de cette approche décrite dans [Tang and Liu, 2010] est d'approximer la structure du graphe, représentée par la matrice d'adjacence A par une structure de blocks. La figure 7 présente une illustration visuelle de l'approche où la matrice visualisée à droite est obtenue à partir de celle affichée à gauche après réorganisation des lignes et des colonnes. Chaque block obtenu corresponds à une communauté. On peut approximer la matrice d'adjacence A par le produit suivant :

$$A \approx S \Sigma S^T \tag{17}$$

où $S \in \{0,1\}^{n \times k}$ est la matrice d'appartenance des nœuds aux blocks, k est le nombre de blocks, et Σ est la matrice de densité d'interactions dans les blocks. Une fonction objective naturelle à minimiser est la suivante :

$$min \parallel A - S \Sigma S^T \parallel_F^2 \tag{18}$$

où $\| \cdot \|_F^2$ est la norme de Frobenius : $\| A \|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} |A_{ij}|^2$.

Ce problème de minimisation est connu pour être NP-difficile quand S est à valeurs discrètes. Une approximation consiste à assouplir la contrainte et considérer la matrice S avec valeurs continues mais en imposant l'orthogonalité des vecteurs de S. Autrement dit on impose que $SS^T = I_k$. Dans ce cas la valeur optimale de S sera les k premiers vecteurs propres de la matrice A associés aux k plus grandes valeurs propres de cette matrice.

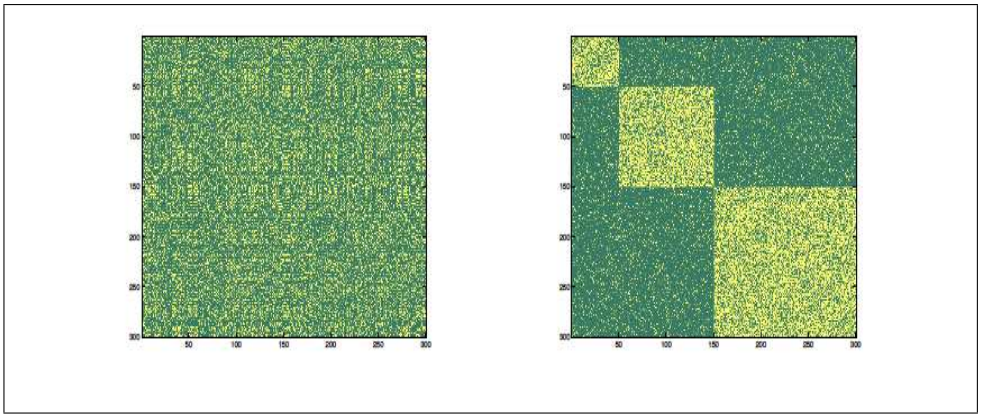


FIGURE 7: Illustration de l'approche de modèle de blocks - exemple tiré de [Tang and Liu, 2010]

2.2.3 Approches d'optimisation

Soit $\Pi = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_{2^n}\}$ l'ensemble de partitions possibles d'un graphe G où n est le nombre de nœuds du graphe. Le problème de détection de communautés peut être ramené à un problème classique d'optimisation avec la définition d'une fonction objective de qualité d'une partition. Le critère de la *modularité* proposé initialement dans [Newman, 2004a] est la fonction objective la plus utilisée depuis. De façon informelle, la modularité d'une partition mesure la différence entre la proportion de liens internes aux communautés et la même quantité dans un modèle nul ou aucun structure communautaire n'est attendue. Le modèle nul est donné par un graphe aléatoire ayant le même nombre de nœuds et de liens et la même distribution de degrés. Plus formellement étant donné une partition $\mathcal{P} = \{c_1, \ldots, c_k\}$ en k communautés. pour une communauté C_i la qualité est donnée par $\sum_{i,j \in c_i} (A_{ij} - \frac{d_i d_j}{2m})$. Pour une partition, la qualité est égale à la somme des qualités de chacune de

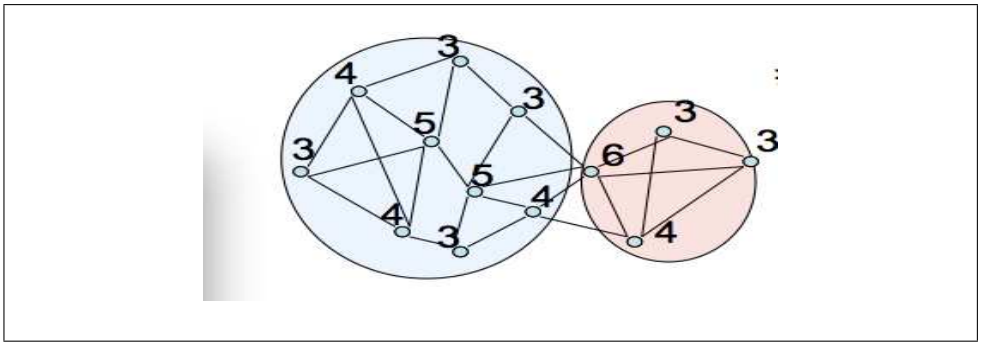


FIGURE 8: Exemple de calcul de la modularité : $Q = \frac{(15+6)-(11.25+2.56)}{25} = 0.275$

ses composantes : $\sum_{c_i \in \pi} \sum_{i,j \in C_i} (A_{ij} - \frac{d_i d_j}{2m})$. La modularité d'une partition $\mathcal{P}$ est alors donnée par la formule suivante :

$$Q(\mathcal{P}) = \frac{1}{2m} \sum_{c \in \mathcal{P}} \sum_{i,j \in c} (A_{ij} - \frac{d_i d_j}{2m}) \tag{19}$$

Le terme $\frac{1}{2m}$ est ajouté afin de normaliser les valeurs possibles de Q dans l'intervalle $[-1, 1]$.

La maximisation de la modularité est un problème NP-difficile [Brandes et al., 2008]. Des méthodes d'optimisation approchées sont proposées pour calculer, en temps et espace polynomiaux, des partitions que l'on espère proches de l'optimum. Des méthodes d'optimisation directe utilisant les techniques d'algorithmique génétique [Li and Song, 2013, Pizzuti, 2012, Cai et al., 2011], de recuit-simulé [Reichardt and Bornholdt, 2006, Guimera et al., 2004] ou de l'optimisation extrêmale [Duch and Arenas, 2005] ont été proposées. Cependant les heuristiques les plus appliquées sont fondées sur le principe de la classification hiérarchique. Deux approches contraires sont largement expérimentées :

- Les approches agglomératives (ou ascendantes) selon lesquelles on part de la partition atomique (ensemble des singletons), et on fusionne deux communautés à chaque itération. Les communautés à fusionner sont celles qui promettent une modularité maximum. Des exemples de ces approches sont données dans [Blondel et al., 2008, Newman, 2004b, Donetti and Munöz, 2004, Pons and Latapy, 2006].
- Les approches divisives (de descendantes) dans lesquelles on part du graphe entier. A chaque itération, on cherche à scinder une

communauté en deux de sorte à maximiser la modularité. Des exemples de cette approche sont données dans [Newman, 2004a, Radicchi et al., 2004].

Les deux types d'approches produisent des hiérarchies de communautés. La figure 14 illustre un exemple d'hiérarchie retournée par un algorithme d'optimisation de la modularité. Généralement, on retient une partition à nombre de communautés voulu, ou celle qui maximise la modularité.

Dans [Tang and Liu, 2010] une formulation matricielle du problème de l'optimisation de la modularité est proposée. On définit la matrice

$$B = A - \frac{dd^T}{2m} \tag{20}$$

L'expression de la modularité d'une partition (voir 19) peut alors être formulée comme suit :

$$Q = \frac{1}{2m} \sum_C S_C^T B S_C = \frac{1}{2m} Tr(S^T B S) = Tr(S^t \tilde{B} S) \tag{21}$$

où $S_C \in \{0,1\}^n$ est le vecteur d'appartenance communautaire des nœuds dans C, S est la matrice d'indication de l'appartenance d'un couple de nœuds à une même communauté, et

$$\tilde{B} = \frac{1}{2m} B = \frac{A}{2m} - \frac{dd^T}{(2m)^2} \tag{22}$$

La maximisation de Q peut se ramener alors au calcul des k premiers vecteurs propres associés aux k valeurs propres les plus grandes de la matrice $\tilde{B}$ sous condition de relaxation spectrale de S (i.e. $SS^T = I$) [Newman, 2006].

Limites de l'optimisation de la modularité Les approches fondées sur l'optimisation de la modularité font implicitement les hypothèses de travail suivantes :

(i) La meilleur décomposition en communautés d'un graphe est celle correspondant à la modularité maximale.

(ii) Si un réseau a une structure communautaire alors on peut trouver une partition précise pour laquelle la modularité est maximale.

(iii) Pour un réseau à structure communautaire, les partitions correspondant à des grandes valeurs de modularité sont structurellement similaires.

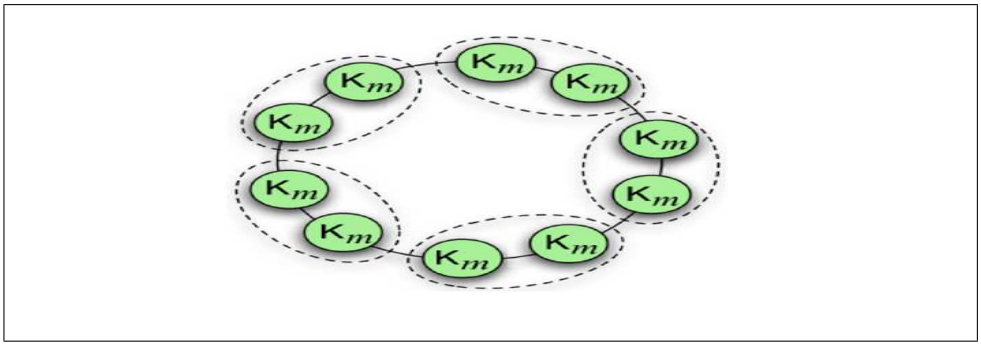


FIGURE 9: Exemple du problème de limite de la résolution de la modularité : La maximisation de la modularité conduit à grouper les cliques deux à deux - Exemple tiré de [Seifi, 2012]

Or, des récentes études ont montré que les trois hypothèses énumérées ci-avant sont toutes fausses. Dans [Fortunato and Barthélemy, 2007] les auteurs montrent que les algorithmes fondés sur l'optimisation de la modularité souffrent d'un problème de limite de résolution dans le sens qu'ils ne peuvent pas distinguer des communautés plus petites d'un certaine taille limite . Pour des graphes non pondérés la maximisation de la modularité ne permet pas de distinguer de communautés ayant un nombre de liens inférieur à $\sqrt{\frac{m}{2}}$. La figure 9 illustre un graphe type qui montre le problème de limite de résolution. Dans ce graphe composé d'un ensemble de cliques de m nœuds connectées en anneau, la maximisation de la modularité va conduire à regrouper les cliques deux à deux. Afin de fixer les idées, considérons le graphe illustré à la figure 9 dans le cas où $m = 3$. La modularité de la partition naturelle sera alors $Q = 0.650$ tandis que la modularité de la partition où les cliques sont regroupées deux à deux sera de $Q = 0.675$.

Dans une tentative de traitement de ce problème de limite de la résolution, une correction de la fonction de la modularité est proposée dans [Reichardt and Bornholdt, 2006] en ajoutant un paramètre de résolution λ comme suit :

$$Q(\mathcal{P}) = \frac{1}{2m} \sum_{c \in \mathcal{P}} \sum_{i,j \in c} (A_{ij} - \boldsymbol{\lambda} \frac{d_i d_j}{2m}) \tag{23}$$

Plus la valeur de λ est grande plus les communautés de petites tailles seront favorisées par Q puisque la maximisation de Q nécessite la minimisation du terme $\boldsymbol{\lambda} \frac{d_i d_j}{2m}$. Inversement, les communautés de grandes tailles seront favorisées en diminuant λ. Noter que pour $\lambda = 1$, on obtient la

même fonction de modularité initiale. Si cette nouvelle fonction de modularité, appelée *modularité mutli-résolution*, peut être réglée pour explorer de communautés à différentes échelles, elle apporte néanmoins une réponse partielle au problème de la limite de résolution puisque les tailles de communautés dans les réseaux réels sont très hétérogènes et suivent aussi une distribution selon une loi de puissance. D'autre part, on montre dans [Lancichinetti and Fortunato, 2011] que la maximisation de la modularité n'a pas seulement tendance à fusionner les petits groupes, mais aussi à éclater des grandes communautés, et il semble impossible d'éviter simultanément les deux problèmes.

Plus sérieux encore sont les leçons tirées de l'étude reportée dans [Good et al., 2010] où les auteurs démontrent l'existence d'un grand nombre de partitions très différentes entre elles mais qu'ont une valeur de modularité optimales. Ce plateau étendu de partitions différentes entre-elles mais qu'ont des modularités maximales explique les différences dans les résultats des différentes approches d'optimisation de la modularité. Dans [Aynaud and Guillaume, 2010] on montre que les algorithmes de maximisation de la modularité sont très sensibles à des perturbations minimales appliquées au graphe étudié.

Ces sérieux inconvénients remettent en cause les nombreuses approches développées pour la détection de communautés par maximisation de la modularité. Cependant, la modularité reste un critère, parmi d'autres, pour pouvoir distinguer et qualifier les qualités des partitions retrouvées mais on sais aujourd'hui que celui-ci ne peut pas être le seul critère possible.

2.2.4 Modèle unifié de détection de communautés

Dans un travail de synthèse, les auteurs de [Tang and Liu, 2010] présentent une approche unifiée pour les méthodes de détection de communautés basées sur l'analyse de la connexion globale du réseau. L'approche unifiée est illustrée à la figure 11. Cette approche est structurée en trois étapes :

Calcul de matrice d'utilité M A partir de la matrice d'adjacence du graphe cible A_G et une fonction objective, on construit une matrice d'utilité M. Dans les cas que nous avons traité ci-avant, M peut être

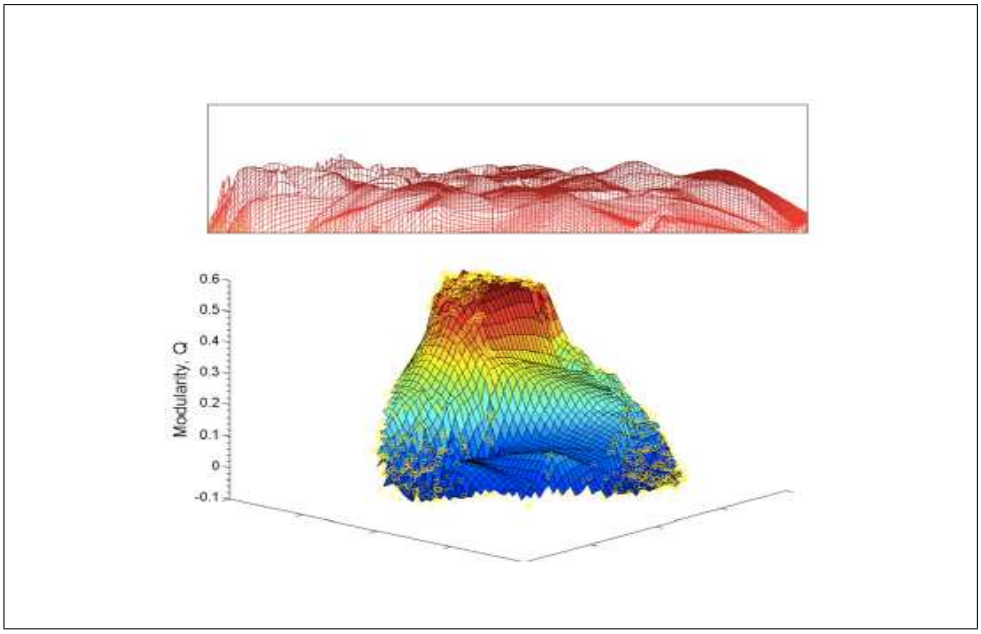


FIGURE 10: Distribution de partitions à modularités maximales - exemple tiré de [Good et al., 2010]

définie comme suit :

$$M = \begin{cases} A \text{ (voir formule 17)} & \text{Modèle de blocks} \\ \tilde{L} \text{ (voir formule 16)} & \text{Clustering spectral} \\ \tilde{B} \text{ (voir formule 22)} & \text{Maximisation de modularité} \end{cases} \tag{24}$$

Calcul de la matrice d'indication S A partir de la matrice d'utilité M on calcule la matrice formée par les K-top vecteurs propres de M. Ces vecteurs représentent l'essentiel des interactions dans le réseau en fonction de la fonction objective employée.

Calcul de la partition du graphe les nœuds du réseau seront recodés en fonction de la nouvelle base définie par la matrice S. L'algorithme k-means peut être appliqué afin de trouver une partition du graphe d'origine dans l'espace défini par S.

2.3 Approches centrées propagation

Les approches centrées propagation exploitent la propriété de la densité des liens intra-communauté. En effet, en raison de la densité relative des communautés et des faibles liens intercommunautaire, on peut

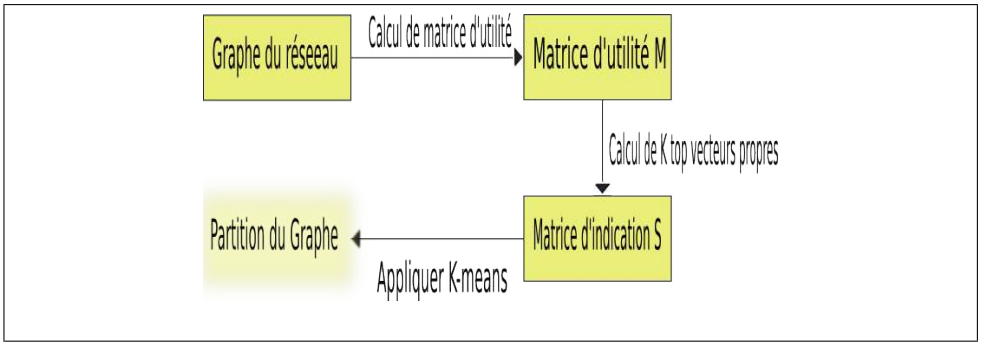


FIGURE 11: Approche unifiée pour la détection de communautés [Tang and Liu, 2010]

raisonnablement admettre qu'un *signal* émis par un nœud et retransmis par ses voisins a plus de chance de rester dans la communauté du nœud source, que de se propager aux autres communautés. Différents algorithmes exploitent cette propriété différemment. Par exemple, l'algorithme *WalkTrap* [Pons and Latapy, 2006] calcule pour chaque nœud dans le graphe un vecteur qui donne la probabilité qu'un marcheur aléatoire arrive aux autres nœuds du réseau en k pas de temps. Les vecteurs de probabilité ainsi calculés pour chaque nœud sont utilisés pour calculer des similarités entre les nœuds. D'autres algorithmes centrés propagation sont les algorithmes basés sur les techniques de propagation de labels [Bajec, 2011, Cordasco and Gargano, 2012, Corlette and III, 2010, Gregor, 2010, Raghavan et al., 2007, Xie and Szymanski, 2011].

2.4 Approches centrées graines

Le schéma général d'une approche centrée graine est structuré en deux étapes :

- Déterminer un ensemble de nœuds ou groupes de nœuds dans le graphe qu'on désigne par des *graines* et qui constituent en quelque sorte les *centres* de communautés à retrouver.
- Appliquer une procédure d'expansion autour des graines afin d'identifier les communautés dans le réseau.

Différentes heuristiques de choix de graine ont été proposées. Une graine peut être composé d'un seul nœud sélectionné en utilisant les mesures classiques de centralité comme c'est fait dans [Khorasgani et al., 2010, Shah and Zaman, 2010]. Dans d'autres algorithmes la graine est composé d'un ensemble de nœuds qui ont une certaine connectivité

[Papadopoulos et al., 2011].

Différentes stratégie d'expansion des graines sont aussi proposées. Dans beaucoup d'algorithmes on utilise les heuristiques développées pour l'identification de communautés locales [Clauset, 2005, Chen et al., 2009, Ngonmang et al., 2012]. Ces approches ne peuvent pas garantir de couvrir l'ensemble de nœuds d'un graphe dans la structure communautaire ainsi calculée. Dans [Kanawati, 2011] une approche plus originale est proposée où après la détection de graines, chaque nœud dans le graphe (graine ou non) calcule un vecteur de préférence d'appartenance aux communautés de chaque graine. L'appartenance communautaire des nœuds est le résultat d'un processus de vote local impliquant le nœud et ses voisins directs. Une étude comparative des approches centrées graines est présentée dans [Kanawati, 2013c]

2.5 Exemples d'algorithmes de détection de communautés

Dans cette section nous présentons quelques exemples des algorithmes les plus connus pour le calcul de partition de graphes de terrain.

2.5.1 Percolation de cliques

L'algorithme est structuré en trois principales étapes :

1. Calculer l'ensemble de cliques de taille k dans le graphe cible G. k est un paramètre de l'algorithme. Soit $\mathcal{C} = \{c_1, \ldots, c_i\}$ l'ensemble des k-cliques
2. Construire un graphe de cliques où chaque clique $c_i \in \mathcal{C}$ est représentée par un nœud. Deux nœuds c_i, c_j sont connectés par un lien si les deux cliques associées partagent $k-1$ nœuds dans le graphe G.
3. Les communautés dans le graphe G sont alors les composantes connexes identifiées dans le graphe de cliques construit à l'étape 2

Ce type d'algorithme fonctionne dans les graphes plutôt denses mais donne des performances limités dans les graphes parcimonieux tels la plupart des graphes de terrain.

2.5.2 Approches divisives pour l'optimisation de la modularité

Une approche divisive consiste à considérer initialement le graphe entier comme une seule communauté. Puis, itérativement supprimer un lien du graphe de sorte que la modularité soit optimisée. Un premier exemple de cette approche est l'algorithme de *Girvan-Newman* [Newman, 2004a] où

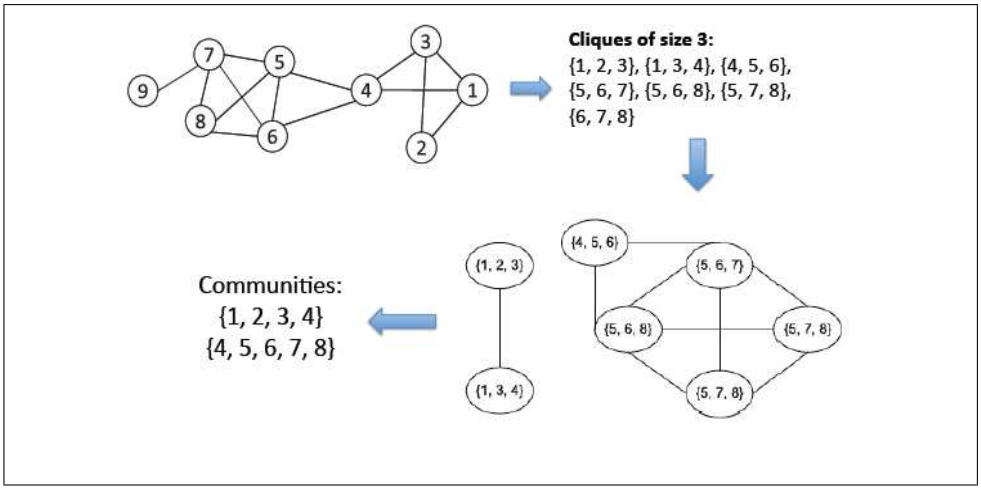


FIGURE 12: Exemple d'exécution de l'algorithme de percolation de cliques - exemple tiré de [Tang and Liu, 2010]

l'heuristique appliquée pour le choix du lien à supprimer à chaque itération consiste à choisir le lien dont la centralité d'intermédiarité est maximale. Rappelons que la centralité d'intermédiarité d'un lien e est donnée par la fraction du nombre de plus courts chemins passant par e et reliant n'import quels couples de nœuds dans le graphe sur le nombre totale de plus courts chemins dans le graphe. L'heuristique appliquée dans [Newman, 2004a] repose sur le fait que les liens inter-communautaire auront forcément une centralité d'intermédiarité élevée. L'inconvénient principale de cet algorithme est sa complexité qui est de l'ordre de $\mathcal{O}(n^3)$.

Une autre algorithme similaire est celui proposé dans [Radicchi et al., 2004]. Dans cet algorithme le lien à supprimer à chaque itération est celui dont le coefficient de clustering est maximal. Le coefficient de clustering d'un lien est défini d'une manière analogue au coefficient de clustering d'un nœud. Il est donné par le nombre de circuits qui passent par le lien sur le nombre de circuits possibles. Le calcul de ce coefficient requiert des calculs locales seulement contrairement à la centralité d'intermédiarité. La complexité de calcul de l'algorithme est de l'ordre de $\mathcal{O}(n^2)$. Le critère d'arrêt de l'itération est aussi différente de celui adopté dans l'algorithme de [Newman, 2004a]. Deux critères d'arrêt sont définis : un pour détecter des communautés fortes où le degré intra-communauté de chaque nœud est supérieur à son degré inter-communauté. Le deuxième critère concerne la détection de communautés faibles et consiste à continuer l'itération tant que la somme des degrés intra-communauté des nœuds dans une communauté

est supérieur à la somme de leurs degrés inter-communautés.

2.5.3 L'algorithme de Louvain

La méthode de Louvain [Blondel et al., 2008] implante une méthode d'optimisation gloutonne locale de la modularité. A l'état initial chaque nœud est affecté à un communauté différente des autres. L'algorithme applique ensuite une itération de succession de deux phases :

Phase d'affectation des nœuds : Pour chaque nœud x on évalue le gain de la modularité si on le déplace dans la communauté de ses voisins directs. On déplace x dans la communauté du voisin qui maximise le gain de la modularité. Si aucun gain n'est trouvé le nœud reste dans sa communauté.

Phase de compression : On compresse le graphe obtenu en remplaçant chaque communauté par un seul nœud. Deux nœuds c_x, c_y dans le nouveau graphe sont liés par un lien s'il existe un lien entre un nœud de la communauté représentée par c_x et un nœud de la communauté représentée par c_y. Le poids de lien entre deux communautés est égale à la somme des poids des liens reliant des nœuds de deux communautés.

La figure 13 illustre l'exécution de ces deux phases dans un double itération. L'algorithme s'arrête s'il a y plus de possibilité de reaffectation de nœuds ou si un maximum de modularité soit atteint. La complexité théorique de l'algorithme n'est pas étudiée, mais d'une manière expérimental, cette complexité est évaluée à $\mathcal{O}(nlogn)$ ce qui fait de Louvain la méthode la plus rapide pour l'identification de communautés.

2.5.4 Propagation de labels

Le premier algorithme implanta l'idée de propagation de labels est proposé dans [Raghavan et al., 2007]. C'est un algorithme itératif où à chaque itération un nœud envoi son label à ses voisins directs, et reçoit ceux de ses voisins. Chaque nœud détermine le label majoritaire qu'il adopte pour l'itération suivante. Ce processus itératif mène à un consensus sur un label précis pour chaque groupe de nœuds. La propagation de labels peut se faire en mode synchrone ou asynchrone. L'avantage du premier mode est qu'il peut facilement passer à l'échelle vu le parallélisme du calcul. Par contre, il peut y avoir un problème de convergence lié à un échange infini de label entre deux nœuds. Ce problème a été évité dans le mode asynchrone. Une version semi-synchrone qui tente d'avoir les avantages des ces deux versions a été proposée dans [Bajec, 2011].

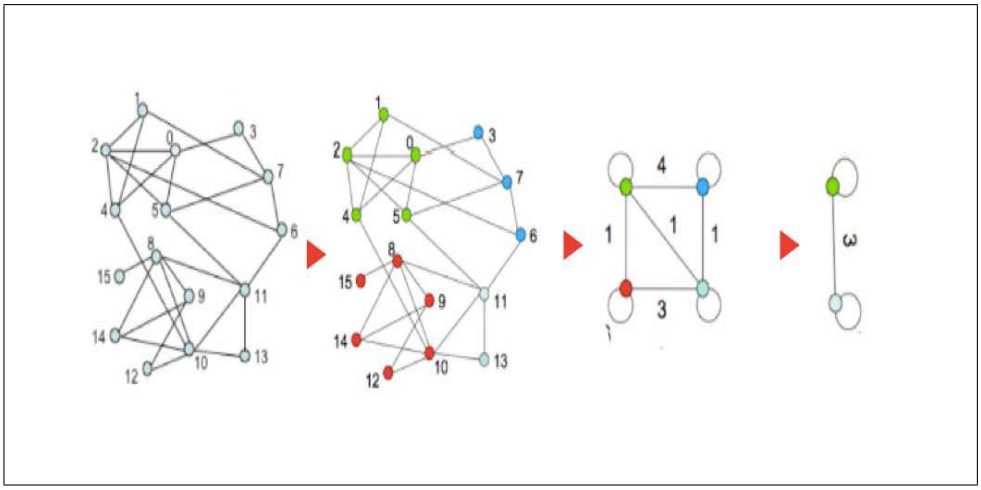


FIGURE 13: Illustration de l'exécution de la méthode de louvain

2.5.5 Walktrap

Cet algorithme proposé dans [Pons and Latapy, 2006] est basé sur l'idée qu'une marche aléatoire partant d'un nœud a plus de probabilité à rester piégée pendant un certain temps dans la communauté du nœud de départ. Supposons que nous effectuions une marche aléatoire courte sur le graphe partant d'un sommet v. Alors, la probabilité d'accéder à chacun des voisins de v en une étape est de $\frac{1}{\|\Gamma(v)\|}$. On peut donc calculer de même manière, la probabilité de se trouver au sommet j en partant de i après avoir effectué aléatoirement k pas. Cette probabilité permet de définir une distance entre les paires de sommets du graphe dans laquelle deux sommets u et v sont proches si leur vecteurs de probabilité d'atteindre les autres sommets sont similaires. Une fois ces probabilités calculées pour tous les paires de sommets, l'algorithme les utilise pour partitionner le graphe par l'intermédiaire d'une méthode de clustering hiérarchique. Commençant par n communautés ne contenant chacune qu'un seul sommet, l'algorithme cherche les deux communautés les plus proches, les fusionne, recalcule les distances, puis effectue une nouvelle fusion et ainsi de suite, jusqu'à n'obtenir qu'une seule communauté recouvrant tout le graphe.

2.5.6 L'algorithme LICOD

Le principe de l'algorithme `Licod`, proposé dans [Kanawati, 2011] est que les communautés se forment autour de nœuds spécifiques appelés *leaders*.

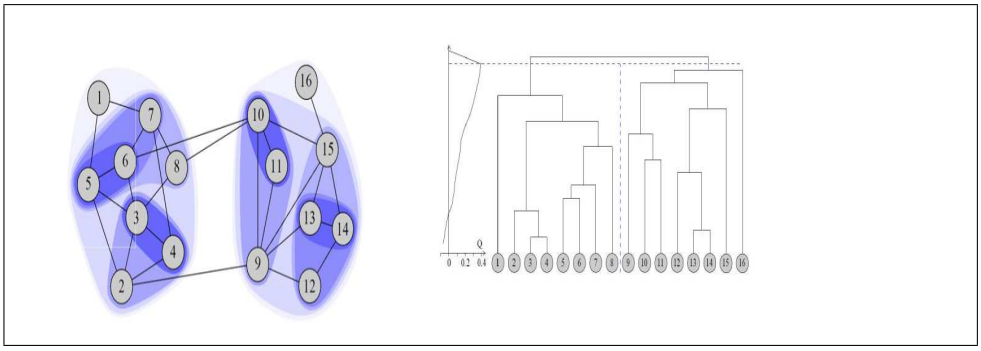


FIGURE 14: Exemple de l'exécution de Walktrap et la structure hiérarchique communautaire détectée

`Licod` se résume en trois étapes :

1. Identification de l'ensemble des Leaders $\mathcal{L}$. Un nœud est déclaré leader si sa centralité est supérieure à la centralité de la plupart de ses voisins. Différentes mesures de centralités peuvent être employées notamment la centralité de degré, la centralité d'intermédiarité et la centralité de proximité . Ici nous utilisons la centralité de *proximité*. Après, $\mathcal{L}$ est réduit en un ensemble $\mathcal{C}$ de communautés de Leaders. Deux leaders sont regroupés s'ils ont un nombre de voisins communs élevé.
2. Chaque $x \in V$ forme un vecteur de préférence P_x^0 où les communautés identifiées dans $\mathcal{C}$ sont triées en ordre décroissant. Dans la version actuelle, le degré d'appartenance d'un nœud x à une communauté $c \in \mathcal{C}$ est simplement donné par $min_{c_i} dist(x, c_i) c_i \in c$. (c peut être représentée par un ensemble de nœdus leaders c_i)
3. Une fois chaque nœud a son vecteur de préférence, on commence une phase d'intégration ou le vecteur de préférence d'un nœud est fusionné avec ceux de ses voisins directes. Ceci permet de favoriser la classe dominante dans l'ensemble des nœuds voisins. Des algorithmes issue de la théorie de choix social sont utilisés dans ce phase de vote [Chevaleyre et al., 2007]. A la stabilisation, chaque nœud x est affecté aux communautés placées en tête de vecteur de préférence.

3 Evaluation des communautés

Les différents algorithmes de détection de communautés calculent souvent des structures communautaires différentes pour un même graphe. Si les

différents algorithmes peuvent être comparés en termes de leurs complexités de calcul et d'espace mémoire requis, la qualité des communautés retrouvées reste un indicateur important de la performance de ces algorithmes. Or, l'évaluation de la qualité des communautés est aujourd'hui encore une question ouverte malgré le nombre important de travaux dans ce domaine. Trois grandes familles d'approches sont proposées dans la littérature :

- Les indices d'évaluation de la similarité d'une partition retrouvée avec une partition de référence.
- Les indices d'évaluation des qualités topologiques des communautés.
- Evaluation guidée par une tâche.

3.1 Indices d'évaluation de communautés par rapport à une partition de référence

La disponibilité d'une partition de référence pour un graphe G peut être le résultat d'un des trois processus suivants :

Annotation par un expert : Les graphes pour lesquels des experts ont défini des partitions de références sont souvent des graphes de très petites tailles. Le tableau 3 décrit les caractéristiques de principaux réseaux réels annotés par des experts et qui sont souvent utilisés comme un benchmark pour les algorithmes de détection de communautés [4]. La figure 15 montre le réseau le plus connu peut être : le réseau de club de Karaté de Zachary, étudié initialement dans [Zachary, 1977] et composé de deux communautés disjointes comme illustré par le code de couleur.

TABLE 3: Quelques réseaux réels souvent utilisés comme un benchmark pour les algorithmes de détection de communautés

Réseau	n	m	**# communautés**
Club de Karaté de Zachary	34	78	2
Football	115	616	11
Strike	24	38	3
Livres politiques	100	441	3
Dauphins	62	159	2

4. La plupart de ces réseaux de benchmark sont disponibles sur la page de jeux de donnée de Pakek : http ://vlado.fmf.uni-lj.si/pub/networks/data/esna/

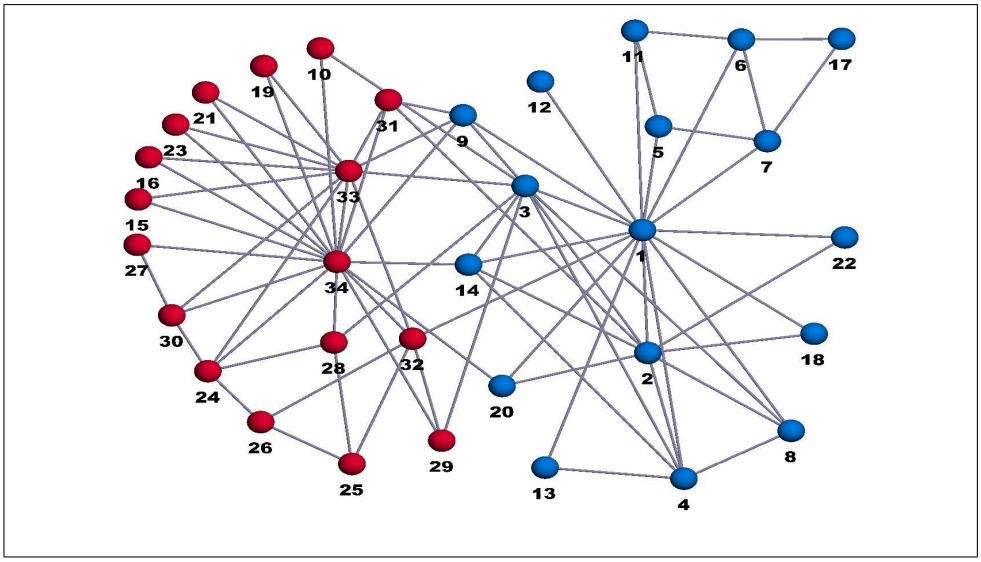


FIGURE 15: Le réseau de club de Karaté de Zachary : il décrit les relations d'amitiés entre 34 membres d'un club de Karaté observées lors de la gestion d'un conflit durable entre l'entraîneur et l'administrateur du club [Zachary, 1977]

Définition implicite : Les approches basées sur *l'inférence* de communautés en fonction de certaines informations sémantiques décrivant les nœuds et/ou les liens d'un réseau. C'est l'approche adoptée dans le travail récent de *Jure Leskovec et al.* [Yang and Leskovec, 2012]. Or, certaines règles appliquées pour l'inférence de communautés sont plus que discutables : Par exemple, dans le cas du réseau de publications reportées dans la fameuse base DBLP, les auteurs proposent que deux auteurs qui publient dans une même conférence appartient à une même communauté !. Dans le cas du réseau social *Live Journal*, ils proposent d'assimiler les groupes de fans d'artistes à des communautés. Il est bien sûr difficile de définir de règles plus précise sans analyse approfondie, mais les communautés ainsi définies soit souvent très nombreuses, de petits tailles par rapport à la taille du réseau et sont mono-thématique.

Génération par un modèle artificielle Le principe est de générer des graphes artificiels avec des structures communautaires paramétrable. Le modèle générateur le plus récent est le modèle *LFR* [Lancichinetti and Radicchi, 2008]. Les générateurs sont basés sur la définition de communautés denses qu'on relie entre elles avec une densité pa-

ramétrable afin de contrôler la difficulté de la reconnaissance de la structure communautaire. Cette approche a le mérite de la simplicité et de la possibilité de générer de graphes de différents tailles et de différents degrés de difficulté à décomposer en communautés. Mais rien ne prouve que les graphes générés par ces modèles sont bien similaires dans leurs mécanismes de formation aux graphes de terrain. L'historique de l'évolution de nos connaissances sur la caractérisation des graphes de terrain ne permet pas de dire qu'on connais déjà la liste ultime de leurs caractéristiques topologiques. Quant à la dynamique de l'évolution de communautés rien n'est encore bien établie de sorte à permettre d'utiliser les actuels modèles génératifs pour la détection de communautés dynamiques.

La disponibilité d'une partition de référence permet d'utiliser les différentes mesures de *distance* entre clusters développées pour l'évaluation des approches de classification non-supervisé (ou clustering) [Aggarwal and Reddy, 2014]. Soit V un ensemble de nœuds d'un graphe G. On désigne par $R = \{r_1, \ldots, r_n\}$ une partition de référence de V. Soit $U = \{u_1, \ldots, u_m\}$ une partition calculée par un algorithme de détection de communautés. Les mesures suivantes sont fréquemment employées pour mesurer la similarité entre deux partitions R et U.

La pureté On définit la pureté d'une communauté $u_i \in U$ par rapport une partition R par :

$$purity(u_i, R) = \max_{j=1 \to n} \frac{\parallel u_i \cap r_j \parallel}{\parallel u_i \parallel} \tag{25}$$

Cette fonctions calcule le taux de recouvrement maximale entre la communauté u_i et les communautés définies dans R. La pureté de la partition U par rapport à R est simplement définie comme la somme pondérée de la pureté de chaque communauté de U par rapport à R :

$$purity(U, R) = \sum_{i=1}^{m} w_{u_i} \times purity(u_i, R) \tag{26}$$

où $w_{u_i} = \frac{\|u_i\|}{\sum_{l=1}^{m} \|u_l\|}$ est la prévalence de la communauté u_i dans U.

L'indice de Rand Cette mesure, initialement proposée dans [Rand, 1971], est basée sur le comptage de nombre d'accords entre deux partitions sur l'appartenance communautaire de chaque pairs de nœuds. Soient :

– a le nombre de pairs placés dans une même communauté selon U et R
– b le nombre de pairs placés en même communauté selon U et en différents communauté selon R
– c le nombre de pairs placés en même communauté selon R et en différents communauté selon U
– d le nombre de pairs placées en différentes communautés selon U et selon R.

La somme $a + d$ donne le nombre d'accords entre les deux partitions, tandis que $b + c$ donne le nombre des désaccords. L'indice de rand est simplement définie par [5] :

$$Rand(U, R) = \frac{a+d}{\binom{n}{2}} \tag{27}$$

Une version ajustée de cette indice, appelée ARI [6] est proposée dans [Hubert and Arabie, 1985] afin d'avoir une mesure dont l'espérance est nulle pour des partitions aléatoires. L'indice ARI est donnée par :

$$ARI(U, R) = \frac{\binom{n}{2}(a+d) - [(a+b)(a+c)+(c+d)(b+d)]}{(\binom{n}{2})^2 - [(a+b)(a+c)+(c+d)(b+d)]} \tag{28}$$

Mesures basées sur l'information mutuelle L'information mutuelle mesure le degré de dépendance entre deux variables aléatoires X et Y et est donnée par la formule générale :

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \tag{29}$$

où $H(X)$ est l'entropie de Shanon de la variable X et $H(X, Y)$ est l'entropie conjointe des deux variables X et Y. Rappelons que l'entropie d'une variable X est mesurée par :

$$H(X) = -\sum_{i=1}^{n_x} p(x_i) log(p(x_i))$$

et l'entropie conjointe de deux variables X, et Y est donnée par :

$$H(X, Y) = -\sum_{i=1}^{n_x}\sum_{j=1}^{n_y} p(x_i, y_j) log(p(x_i, y_j))$$

5. rappel : $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
6. Adjusted Rand Index

où $\{x_i\}$ (reps. $\{y_j\}$) est l'ensemble de n_x (reps. n_y) valeurs possibles de X (reps. Y). $p(x_i)$ est la probabilité pour X d'avoir la valeur x_i. et $p(x_i, y_j)$ est la probabilité pour que conjointement X ait la valeur x_i et Y ait la valeur y_j.

En assimilant une partition U à une variable aléatoire, on peut écrire son entropie sous la forme :

$$H(U) = -\sum_{i=1}^{|U|} p(u_i) log(p(u_i))$$

La probabilité $p(u_i)$ d'un nœud de V d'appartenir à la communauté u_i est égale à : $\frac{\|u_i\|}{n}$. En substituant dans l'expression 29, on obtient l'expression de l'information mutuelle entre deux partitions :

$$IM(U, R) = \sum_{u \in U} \sum_{r \in R} p(u, r) log(\frac{p(u, r)}{p(u)p(r)}) \tag{30}$$

Une version normalisée de l'information mutuelle est introduite dans [Strehl and Ghosh, 2003] afin d'obtenir une mesure entre 0 et 1. L'information mutuelle normalisée (NMI) est donnée par

$$NMI(U, V) = \frac{I(U, V)}{\sqrt{H(U)H(V)}} \tag{31}$$

Une version ajustée de la mesure NMI est récemment introduite dans [Nguyen et al., 2009]. Dans [Meila, 2003], une mesure similaire est introduite, appelée variation de l'information (IV) et est donnée par :

$$VI(U, V) = H(U) + H(V) - 2 \times IM(U, V)$$

Les mesures de comparaison de partitions basées sur la théorie de l'information sont assez corrélées entre elles. En pratique, la mesure NMI reste la plus utilisée dans la littérature scientifique.

Distance d'édition Le principe est de calculer le coût minimale en terme de changement d'appartenance communautaire des nœuds pour obtenir la partition R en partant d'une partition U. Nous supposons que les deux partitions ont le même nombre de communautés (nous ajoutons des communautés vides s'il n'y en a pas le même nombre). On définit une *association* $\delta : U \to R$ comme une bijection entre les communautés de U et celles de R. Ensuite, nous calculons le coût correspondant au nombre de mouvements

nécessaires pour faire coïncider chaque communauté $u_i \in U$ et son image $\delta(u_i) \in R$. Une fonction de coût peut être $\parallel u_i \setminus \delta(u_i) \parallel$. Le coût associé à une fonction d'association δ_x est alors donnée par :

$$cost(\delta_x(U,R)) = \sum_{u_i \in U} \parallel u_i \setminus \delta_x(u_i) \parallel$$

La distance d'édition entre les deux partitions est alors donnée par le coût minimale qu'on peut trouver :

$$Distance(U,R) = \min_{\delta_x} cost(\delta_x(U,R)) \tag{32}$$

L'ensemble des mesure présentées ici, sont des mesures issues du domaine de la classification non-supervisée. Ces mesures, appliquées au problème d'évaluation de la détection de communautés ignorent par construction une partie importante de l'information disponible, à savoir la topologie du réseau. Des travaux récents ont tenté d'aborder ce problème. Dans [Orman et al., 2012], les auteurs proposent d'utiliser de façon conjointe les mesures traditionnelles et différentes propriétés topologiques. Cependant, ils reconnaissent eux-mêmes que l'utilisation de ces dernières n'est pas simple, car leur quantification prend la forme de plusieurs séries numériques, difficiles à comparer.

Dans [Labatut, 2012], une adaptation de la mesure de pureté est proposée afin de prendre en compte la topologie du réseau lors de la comparaison de deux partitions. L'idée de base est de définir une pureté nodale qui mesure la pureté pour un nœud x pour une partition U par rapport à une partition R. Cette pureté est définie comme suit :

$$purity(x,U,R) = \delta_{argmax_j \|u(x) \cap r_j\|, r(u)}$$

où $delta_{x,y}$ est l'indice de Kronecker, $u(x)$ (reps. $r(x)$) est la communauté dans U (reps. R) auquel appartient le nœud x. Cette pureté nodale a donc une valeur binaire : 1 si $r(x)$ est la communauté majoritaire dans $u(x)$, et 0 sinon. La pureté d'une communauté par rapport à une partition est reformulée pour être la moyenne de la pureté nodale des nœuds dans la communauté :

$$purity(u_i,R) = \frac{1}{\parallel u_i \parallel} \sum_{x \in u_i} purity(x,U,R)$$

Cette nouvelle pureté peut être utilisée dans l'équation 26 pour avoir la pureté d'une partition U par rapport à une autre R. Une généralisation de ce schéma est à faire pour ajuster l'ensemble des autres mesures de comparaison de partitions.

3.2 Mesures topologiques pour l'évaluation de communautés

Nous distinguons ici deux types de mesures topologiques :

- Les mesures globales qui évaluent la qualité d'une partition. Le critère de la modularité, présenté dans la section 2.2.3 est le critère le plus employé pour mesurer la qualité intrinsèque d'une partition.
- Les mesures centrées sur la qualité individuelle des communautés formant une partition. En effet, beaucoup de mesures de qualité d'une communauté isolée ont été introduites dans le cadre de la résolution du problème d'identification de la communauté d'un nœud ou ce qui est appelé aussi la détection de communauté locale [Chen et al., 2009, Zhang and Wu, 2012, Ngonmang et al., 2012]. Dans ce contexte, la qualité d'une partition est donnée par la moyenne des qualités des communautés qui la composent.

$$Q(\mathcal{C}) = \frac{\sum_i f(S_i)}{|\mathcal{C}|} \tag{33}$$

où $f()$ est une fonction de qualité d'une communauté.

Dans [Leskovec et al., 2010] trois familles de fonctions de qualité topologique d'une communauté sont identifiées :

- Fonctions basées sur la connectivité interne.
- Fonctions basées sur la connectivité externe.
- Fonctions hybrides

La tableau 4 résume l'essentiel de ces fonctions de qualité. La notation employé est la suivante :

- n_c : est le nombre de nœuds dans la communauté c
- m_c : le nombre de liens dans la communauté c.
- b_c : le nombre de liens sortant de la communauté c.
- d^m est le médian de degrés des nœuds dans V

3.3 Evaluation guidée par une tâche

La rareté des réseaux de grands tailles pour lesquelles une partition de référence est connue (voir 3), les limitations des critères topologiques d'évaluation des communautés (voir 19) et les limitations des modèles générateurs de réseaux artificiels de benchmark, sont quelques facteurs qui ont motivé la recherche de nouvelles approches pour l'évaluation des partitions détectées par les différentes algorithmes de détection de communautés. L'évaluation guidée par une tâche semble être une alternative prometteuse. Le principe est simple : Soit T une tâche où la détection de com-

TABLE 4: Mesures topologiques d'évaluation d'une communauté c

Mesure	Type	Formule	Description
Densité interne	int.	$\frac{2\times m_c}{n_c\times(n_c-1)}$	Densité du sous-graphe induit par la comunauté
Degrés moyen	int.	$\frac{2\times m_c}{n_c}$	Moyen de degrés internes
FOMD	int.	$\frac{\lvert\{u:u\in c,\lvert(u,v),v\in c\rvert>d^m\}\rvert}{n_c}$	Le pourcentage des nœuds interns ayant une degré > la medianne des degrés
TPR	interne	$\frac{\lvert\{u\in c\}:\exists v,w\in c:(u,v),(w,v),(u,w)\in E\rvert}{n_c}$	Taux d'implication dans des traingles
Expansion	Ext.	$\frac{b_c}{n_c}$	Nombre de liens sortant per nœud
Taux de coupe	Ext.	$\frac{b_c}{n_c\times(N-n_c)}$	Taux de liens sortant sur les lines sortants possibles
Conductance	Hybride	$\frac{b_c}{2m_c+b_c}$	La fraction de liens sortant
MAX-ODF	Hybride	$\max_{u\in c}\frac{\lvert\{(u,v)\in E,v\notin c\}\rvert}{d_u}$	Le max per nœud de liens sortant
AVG-ODF	Hybride	$\frac{1}{n_c}\times\sum_{u\in c}\frac{\lvert\{(u,v)\in E,v\notin c\}\rvert}{d_u}$	

munautés peut être appliquée. Soit $per(T, Algo^x_{com})$ un indicateur de performance de l'exécution de la tâche T en utilisant l'algorithme de détection de communautés $Algo^x_{com}$. Nous pouvons comparer les performances des deux algorithmes différents en fonction des indicateurs $per(T, Algo^x_{com})$ et $per(T, Algo^y_{com})$.

Dans [Papadopoulos et al., 2012], les auteurs proposent d'utiliser la tâche de recommandation de tags dans des folksonomies. Dans [Yakoubi and Kanawati, 2012, Yakoubi and Kanawati, 2013] la tâche de classification non-supervisée de données *non relationnelles* est employée (voir figure 16). Afin de classifier les données, l'approche com-

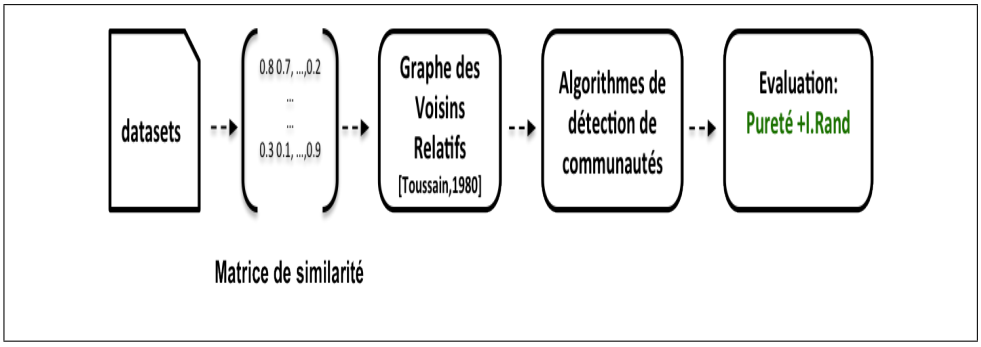


FIGURE 16: Evaluation des algorithmes de détection de communautés sur une tâche de clustering de données

mence par structurer les données sous forme de graphe de voisinage [Toussaint and Bhattacharya, 1981] définie à l'aide d'une fonction de distance appropriée. Les algorithmes de détection de communautés peuvent alors être appliqués sur ce graphe pour identifier des clusters.

4 Approches de détection de communautés dans des graphes multiplexes

4.1 Problématique

Nous définissons un graphe multiplex, désigné aussi par le termes graphe multi-couches [7] ou encore graphe multi-relationnel, comme un graphe composé d'un ensemble de nœuds de même type, reliés par différents types de relations. Une présentation usuelle (voir figure 17) est de représenter un tel réseau sous forme de réseau multi-couches. Chaque couche contient le même ensemble de nœuds V. Mais chaque couche corresponds à un type de relation différente. Par exemple, dans le cas d'un réseau de co-achat, les nœuds représentent les clients. Chaque couche représente les relations de co-achat d'un genre particulier de produits. Dans le cas de réseaux bibliographiques, on peut définir un multiplex où les nœuds sont les auteurs et chaque couche corresponds à une relation différente : co-publication, co-citation, co-cités, co-participation à une conférence [Davis et al., 2011, Kanawati, 2013a].

D'une manière formelle, nous définissons un réseau multiplex structuré en α couches par $G =< V, E_1, \ldots, E_\alpha >$. Chaque couche du multiplex est décrite par une matrice d'adjacence $A_G^{[\alpha]}$. Le tableau 5 donne les principales

7. Multi-slice en anglais

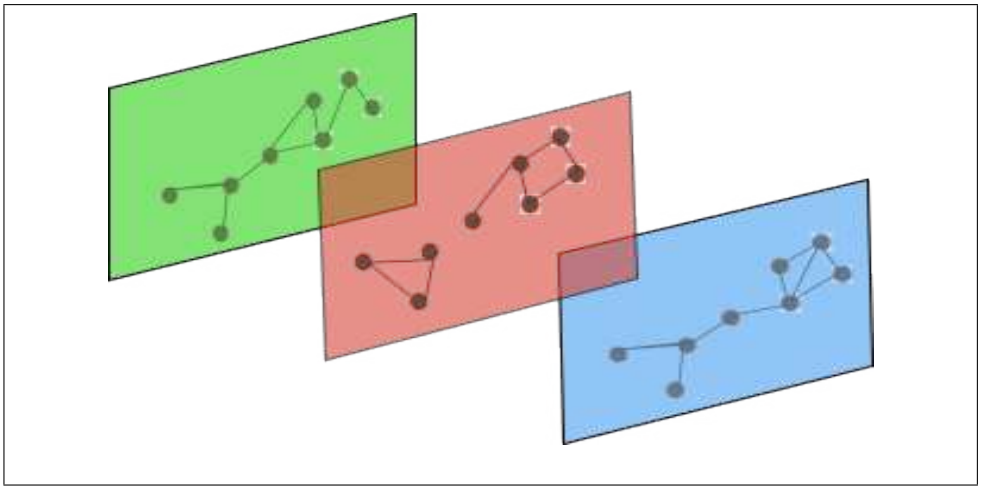


FIGURE 17: Illustration d'un réseau multiplex

TABLE 5: Notations utilisées pour les réseaux multiplexes

Notation	Description
$A^{[k]}$	Matrice d'adjacence de la couche k
$d_i^{[k]}$	degré de nœud i dans la couche k
$m^{[k]}$	nombre de liens dans la couche k
C_{ij}^{kl}	Le poids de lien inter-couches k et l entre le nœuds i et le nœud j

notations que nous utilisons dans la suite de ce travail.

Par analogie à la définition commune d'une communauté dans un graphe simple, nous définissions une communauté multiplexe comme un sous-graphe *dense dans le réseau multiplex* qui est faiblement connecté aux autres communautés dans le graphe. La notion de densité dans un graphe multiplex reste une notion floue : Lequel des graphes illustrés à la figure 18 est le plus dense dans le réseau multiplex ?

Les différents travaux abordant le problème de détection de communautés dans les réseaux multiplexes adoptent des définitions différentes. Trois principales approches peuvent être identifiées dans la mise en œuvre des travaux existant :

Agrégation des couches : Le principe est de transformer un réseau multiplex en un réseau simple en appliquant une stratégie d'intégration des différentes couches. Les algorithmes de détection de communautés

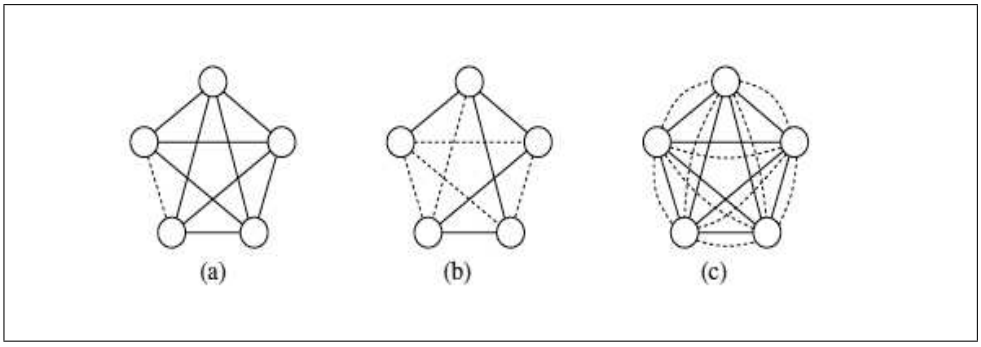


FIGURE 18: Sous-graphes denses dans un réseau multiplex ? - Exemple tiré de [Berlingerio et al., 2011]

classiques peuvent alors être appliqués sur le réseau résultant.

Agrégation des partitions : L'idée de base est de détecter de structures communautaires dans chaque couche du réseau multiplex en isolation de l'autre. Ensuite une stratégie de fusion des partitions retrouvées de chacune des couches peut être appliquée.

Exploration simultanée des couches L'idée est d'intégrer la prise en compte de la nature multiplex du réseau dans le processus même de la détection des communautés.

4.2 Agrégation des couches

Soit $G_M =< V, E_1, \ldots, E_\alpha >$ un graphe multiplex défini sur α couches. $A^{[k]}$ est la matrice d'adjacence L'approche consiste à transformer ce graphe en graphe simple pondéré $G =< V, E, W >$ où W est une matrice de poids sur les liens $e \in E$. L'idée est que la matrice de poids préserve autant que possible les informations contenues dans le graphe multiplex d'origine. Différentes fonctions de calcul des poids sont proposées. Nous citons dans la suite les plus utilisées :

Pondération binaire : Le principe est de lier deux nœuds dans le graphe simple si il existe au moins un lien entre ces deux mêmes nœuds dans au moins une des α couches [Berlingerio et al., 2011, Suthers et al., 2013]. Plus formellement nous avons :

$$w_{ij} = \begin{cases} 1 & \text{si } \exists 1 \leq i \leq \alpha : (i,j) \in E_i \\ 0 & \text{sinon} \end{cases} \tag{34}$$

Pondération selon la fréquence Dans [Tang and Liu, 2010] on propose de pondérer les liens (i, j) dans E par la moyenne des poids du lien dans l'ensemble des couches. Formellement ;

$$w_{ij} = \frac{1}{\alpha} \sum_{k=1}^{\alpha} A_{ij}^{[k]} \tag{35}$$

Un autre schéma de pondération similaire est celui proposé dans [Berlingerio et al., 2011] où le poids d'un lien est calculé par la redondance du lien dans les différentes couches :

$$w_{ij} = \| \{d : A_{ij}^{[d]} \neq 0\} \| \tag{36}$$

Pondération par un mesure de similarité Une manière plus générale est de pondérer un lien (i, j) dans un graphe simple représentant un graphe multiplex, par une similarité multiplex. On peut pratiquement, utiliser les mêmes techniques employées pour le calcul de versions temporelles des mesures de similarité dyadiques dans un graphe dynamique [Potgieter et al., 2009]. En effet, l'historique de l'évolution d'un graphe sur β pas de temps peut être assimilée à un graphe multiplex de β couches. La différence est que le temps induit un ordre sur les couches contrairement au réseau multiplex général où aucune ordre ne peut être définie sur les couches. Dans [Berlingerio et al., 2011] les auteurs proposent l'utilisation du coefficient de clustering d'un lien potentiel (i, j) comme une mesure de similarité pour la pondération du graphe G.

Combinaison linéaire Dans [Cai et al., 2005] les auteurs avancent l'idée que les couches d'un multiplex peuvent avoir des contributions variables à la composition des communautés et que l'intégration de couches peut être faite par une combinaison linéaire des matrices d'adjacence des couches. Nous avons :

$$A = \sum_{k=1}^{\alpha} w_k A^{[k]} \tag{37}$$

où A est la matrice d'adjacence du graphe simple résultat. Les auteurs proposent d'apprendre les poids w_k à utiliser à partir des contraintes que l'utilisateur peut fournir sur l'appartenance communautaire de certains couples de nœuds : des couples qui doivent être dans une même communauté et d'autres qui doivent être placés dans des communautés différentes.

Toutes ces approches d'intégration de couches induisent une perte de l'information de la multiplicité des liens. Il n'est pas possible de retrouver le réseau multiplex à partir du réseau simple résultat de l'agrégation. L'avantage évidente, est la simplicité de la mise en œuvre et la possibilité d'utiliser une large choix d'algorithmes de détection de communautés développés pour les réseaux simples. La seule condition sur l'algorithme à utiliser est d'être capable de prendre en compte les poids de liens lors du calcul des communautés.

4.3 Agrégation de partitions

A l'opposé des approches d'agrégation de couches, l'agrégation de partitions consiste à appliquer un algorithme de détection de communautés à chacune des α couches. On obtient alors α partitions différentes de l'ensemble des nœuds V. Ces partitions peuvent être agrégées en une seule partition en utilisant les techniques classiques de clustering d'ensemble [mlX, 2010, Strehl and Ghosh, 2003, Topchy et al., 2005, Goder and Filkov, 2008].

Une autre approche similaire est l'approche de calcul de cœurs de communautés développée dans [Seifi, 2012]. L'approche consiste à construire une matrice $\mathcal{F}$ de dimension $n \times n$ où chaque élément F_{ij} donne la fréquence de l'association des nœuds i et j dans une même communauté dans l'ensemble des partitions $P^{[k]}$. Un graphe G^{β} de fréquence de co-association est construit sur l'ensemble de nœuds V où deux nœuds sont liés par un lien si $F_{ij} \geq \beta$. Les composantes connexes du graphe G^{β} sont déclarés des cœurs de communautés du graphe d'origine G. L'approche est initialement conçue pour le calcul de cœure de communautés en utilisant des algorithmes non stables de détection de communautés mais elle peut naturellement être étendue à la fusion de partitions différentes obtenue sur différentes couches d'un multiplex. Cependant, l'approche donne de meilleurs résultats si le nombre de couches est assez grande afin de pouvoir obtenir des fréquences de co-affectation significatives.

4.4 Exploration simultanée des couches

Très peu de travaux existant ont abordé le problème d'exploration directe de l'ensemble des couches pour la détection de communautés multiplexes. Quelques approches cependant ont tenté de généraliser des approches de détection de communautés dans les graphes simples aux graphes multiplexes. Un des premiers travaux dans ce contexte est le travail de Tang et.al. [Tang and Liu, 2010] dans lequel les auteurs, et grâce au modèle unifié

(voir section 2.2.4) qu'ils proposent, ont pu identifier de nouvelles approches d'agrégation autres que les deux approches *triviales* d'agrégation des couches et d'agrégation des partitions. En effet, selon le modèle unifié illustré à la figure 11, l'agrégation peut se faire au niveau des matrices d'utilités ou encore au niveau des matrices d'indication d'appartenance communautaire. L'inconvénient de cette approche reste l'emploi en dernière étape d'un schéma de clustering utilisant l'algorithme K-means qui nécessite d'avoir le nombre de communautés à trouver. L'approche est adéquate pour des graphes de taille intermédiaire mais peu adaptée pour les très grandes graphes.

Le rôle prépondérant que la modularité et son optimisation ont joué dans le contexte de détection des communautés dans des graphes simples a tout naturellement motivé des travaux de généralisation de la modularité au cas de réseaux multiplexes. Une nouvelle formule de la modularité multiplexe est ainsi proposée dans [Mucha et al., 2010]. Cette nouvelle modularité est donnée par la formule suivante :

$$Q_{multiplex}(P) = \frac{1}{2\mu} \sum_{c \in P} \sum_{\substack{i,j \in c \\ k,l:1 \to \alpha}} \left(\left(A_{ij}^{[s]} - \lambda_k \frac{d_i^{[k]} d_j^{[k]}}{2m^{[k]}} \right) \delta_{kl} + \delta_{ij} C_{ij}^{kl} \right) \quad (38)$$

où $\mu = \sum_{\substack{j \in V \\ k,l:1 \to \alpha}} m^{[k]} + C_{jkl}$ est un facteur de normalisation, et λ_k est un facteur de résolution comme introduite pour la modularité multi-résolution [Reichardt and Bornholdt, 2006] (voir formule 23). Noter que dans notre cas (i.e. multiplex multi-couches) les seuls liens inter-couches sont les liens implicites reliant un nœud i à lui même dans les autres couches. Par conséquent nous avons : $C_{ij}^{kl} = 0 \ \forall i \neq j$.

Cette nouvelle modularité permet d'étendre l'applicabilité des algorithmes d'optimisation développées pour les graphes simples au cas des graphes multiplexe. Récemment, une version inspirée de l'algorithme de Louvain a été proposée en utilisant la modularité multiplex [Carchiolo et al., 2011]. Cependant, l'optimisation de la modularité multiplex est aussi susceptible d'avoir les mêmes inconvénients et les mêmes limitations des approches d'optimisation de la modularité.

Dans un travail très récent [Battiston et al., 2013], les auteurs proposent de généraliser les mesures classiques dans les graphes simples au cas des graphes multiplexes. L'approche proposée est assez radicale dans le sens où l'implication de plus d'une couche est requise pour qualifier un nœud ou un lien. Nous prenons le cas simple de la définition du degré d'un nœud pour

illustrer l'approche empruntée pour la définition de mesures multiplexes. Soit $d_i^{[tot]} = \sum_{i=2}^{\alpha} d_i^{[i]}$ le degré total d'un nœud i dans le réseau multiplex. On définit le degré multiplex du nœud i par :

$$d_i^{multiplex} = -\sum_{k=1}^{\alpha} \frac{d_i^{[k]}}{d_i^{[tot]}} log\left(\frac{d_i^{[k]}}{d_i^{[tot]}}\right) \tag{39}$$

En utilisant une fonction d'entropie, les auteurs de [Battiston et al., 2013] mettent l'accent sur l'importance de l'implication d'un nœud dans plus d'une couche du multiplex. Par exemple, le degré multiplex d'un nœud i dont tous les liens adjacents sont dans une même couche est égale à 0. Le degré sera maximale si le nœud a le même nombre de liens dans chacune des couches. Ce schéma de redéfinition des mesures topologiques pour les graphes multiplexes permet d'envisager d'appliquer d'autres types d'algorithmes de détection de communautés, notamment les algorithmes centrés graines et les algorithmes centrés diffusion qui ont montré des performances intéressantes lors de l'application sur des réseaux simples.

4.5 Critères d'évaluation

Dans la section 3, nous avons passé en revue les principales approches d'évaluation des communautés détectées dans les graphes simples. Très peu de travaux ont abordé cette épineuse question dans le cas des graphes multiplexes. En effet, à notre connaissance, nous n'avons pas de graphes réels ni de générateurs de graphes artificiels qui peuvent nous donner des réseaux multiplexes de benchmrak.

Au niveau des indicateurs topologiques, la modularité multiplexe peut être une mesure globale. Or, les limitations de l'optimisation de la modularité limite aussi l'intérêt de cette mesure pour l'évaluation. Un indicateur topologique de qualité d'une communauté multiplexe a été proposé dans [Berlingerio et al., 2011]. L'indicateur, appelé *mesure de redondance* calcule la moyenne de la redondance de chaque lien intra-communauté dans l'ensemble des couches du multiplex. L'intuition est que les liens intra-communauté doivent être des liens récurrents dans les différentes couches. Le calcul de cet indicateur est fait comme suit. Soient :

- P l'ensemble de couples (u, v) qui sont directement connectés dans une couche au moins.
- $\bar{\bar{P}}$ l'ensemble de couples (u, v) qui sont directement connectés dans deux couches au moins.

– $P_c \subset P$ l'ensemble de liens dans la communauté c
– $\bar{\bar{P}}_c \subset \bar{\bar{P}}$ le sous-ensemble de $\bar{\bar{P}}$ et qui sont aussi dans c.

La redondance d'une communauté c est alors donnée par :

$$\rho(c) = \sum_{(u,v)\in\bar{\bar{P}}_c} \frac{\parallel \{k : \exists A_{uv}^{[k]} \neq 0\} \parallel}{\alpha\times \parallel P_c \parallel} \tag{40}$$

La qualité d'un partition multiplexe peut être donnée par :

$$\rho(\mathcal{P}) = \frac{1}{\parallel \mathcal{P} \parallel} \sum_{c\in\mathcal{P}} \rho(c) \tag{41}$$

5 Conclusion

La codage de traces des activités interactives sous forme de graphes permet d'appliquer la très riche *boîte à outils* d'approches d'exploration, d'analyse et de fouille de réseaux complexes développé aujourd'hui dans une communauté scientifique pluridisciplinaire. Un problème central dans l'étude des réseaux complexes est celui de la détection des communautés : des sous-graphes denses faiblement connectés entre eux. La découverte de la structure communautaire d'un graphe permet d'enrichir nos connaissance sur la structure interne des schéma des interactions mais aussi nous renseigner sur les possibilités d'évolution du graphe. Dans le cadre des systèmes de recommandation, la détection de communautés peut être vue comme une généralisation du principe du filtrage collaboratif où nous pouvons exploiter les caractéristiques de chaque communauté afin de mieux cibler les recommandations aux niveaux individuelles. Les mêmes techniques peuvent être utilisées aussi pour caractériser de groupements (i.e clusters) de produits. La littérature scientifique concernant la détection de communautés est très abondante. Il existe un nombre impressionnant d'approches différentes développées dans différentes disciplines. Dans ce travail, nous avons présenté un bref survol des principales approches de détection de communautés dans les graphes simples. Nous avons aussi pointé la difficulté de l'évaluation et de la caractérisation des communautés détectées par les différents algorithmes. Une approche qui nous semble prometteur est celle de l'évaluation guidée par des tâches. La tâche de recommandation est une tâche de choix pour ce type d'évaluation.

Or, dans beaucoup de cas réels qui nous intéressent, les graphes d'interaction ont une structure multiplexe : Le graphe peut être ramené à un graphe

multi-couches défini sur le même ensemble de nœuds mais où chaque couche encode une relation différente. Cette nouvelle modélisation permet de capter plus d'informations sur les interactions étudiées comme c'est le cas par exemple de réseau d'évaluation de produite, les réseaux de co-achats mais aussi des relations différentes dans les réseaux sociaux dans lesquelles les utilisateurs sont impliqués (relation d'amitiés, relation familiales, ..., etc). Dans le cas des interactions dans les sites de partage sociale et de microblogging (ex. Twitter), la problématique d'analyse de graphes multiplexes est naturellement posée. Nous avons centré notre étude ici sur les techniques de détection de communautés dans les graphes multiplexes. Le concept est encore flou et davantage d'études de caractérisation de communautés multiplexes est nécessaire. La pauvre palette d'outils d'évaluation propres aux communautés multiplexes en témoigne. En effet, la majeur partie des approches existantes consiste à transformer le problème, d'une manière ou d'aune autre, en problème de détection de communautés dans des graphes simples. Ceci est fait soit par agrégation des couches du réseau multiplex soit par agrégation des partitions obtenues sur chacune des couches séparément. Très récemment des approches qu'on peut qualifier de natives ont été proposées pour la détection de communautés multiplexes. Ces approches sont principalement des généralisations des approches déjà développées pour les graphes simples et partagent naturellement les inconvénients de leurs homologues. C'est notamment le cas des approches, pourtant rapides sur des très grands graphes, de l'optimisation gloutonne de la modularité. La proposition récente de mesures topologiques propres aux graphes multiplexes ouvre la voie à l'application d'autres approches intéressantes pour la détection de communautés et qui ont fait leurs preuves pour les graphes simples, notamment les approches centrées graines et les approches centrées propagation.

Références

[mlX, 2010] (2010). Clustering ensembles. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning*, page 180. Springer.

[Adamic and Adar, 2003] Adamic, L. and Adar, E. (2003). Friends and neighbors on the Web. *Social Networks*, 25(3) :211–230.

[Aggarwal and Reddy, 2014] Aggarwal, C. C. and Reddy, C. K., editors (2014). *Data Clustering : Algorithms and Applications*. CRC Press.

[Aidouni et al., 2009] Aidouni, F., Latapy, M., and Magnien, C. (2009). Ten weeks in the life of an edonkey server. In *IPDPS*, pages 1–5. IEEE.

[Archambault and Grudin, 2012] Archambault, A. and Grudin, J. (2012). A longitudinal study of facebook, linkedin, & twitter use. In Konstan, J. A., Chi, E. H., and Höök, K., editors, *CHI*, pages 2741–2750. ACM.

[Aynaud and Guillaume, 2010] Aynaud, T. and Guillaume, J.-L. (2010). Static community detection algorithms for evolving networks. In *WiOpt*, pages 513–519. IEEE.

[Baeza-Yates, 2007] Baeza-Yates, R. A. (2007). Graphs from search engine queries. In *SOFSEM (1)*, pages 1–8.

[Bajec, 2011] Bajec, M. (2011). Robust network community detection using balanced propagation. *Nature*.

[Barabási and Albert, 1999] Barabási, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286 :509.

[Barthelemy, 2010] Barthelemy, M. (2010). Spatial networks. *CoRR*, abs/1010.0302.

[Bastian et al., 2009] Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi : An open source software for exploring and manipulating networks. In Adar, E., Hurst, M., Finin, T., Glance, N. S., Nicolov, N., and Tseng, B. L., editors, *ICWSM*. The AAAI Press.

[Battiston et al., 2013] Battiston, F., Nicosia, V., and Latora, V. (2013). Metrics for the analysis of multiplex networks. *CoRR*, abs/1308.3182.

[Benchettara et al., 2010] Benchettara, N., Kanawati, R., and Rouveirol, C. (2010). A supervised machine learning link prediction approach for academic collaboration recommendation. In Amatriain, X., Torrens, M., Resnick, P., and Zanker, M., editors, *RecSys*, pages 253–256. ACM.

[Berlingerio et al., 2011] Berlingerio, M., Coscia, M., and Giannotti, F. (2011). Finding and characterizing communities in multidimensional networks. In *ASONAM*, pages 490–494. IEEE Computer Society.

[Blondel et al., 2008] Blondel, V. D., Guillaume, J.-l., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. pages 1–12.

[Bouveyron and Chipman, 2007] Bouveyron, C. and Chipman, H. A. (2007). Visualization and classification of graph-structured data : the case of the enron dataset. In *IJCNN*, pages 1506–1511. IEEE.

[Brandes et al., 2008] Brandes, U., Delling, D., Gaertler, M., Görke, R., Hoefer, M., Nikoloski, Z., and Wagner, D. (2008). On modularity clustering. *IEEE Trans. Knowl. Data Eng.*, 20(2) :172–188.

[Cai et al., 2005] Cai, D., Shao, Z., He, X., Yan, X., and Han, J. (2005). Mining hidden community in heterogeneous social networks. In *ACM-SIGKDD Workshop on Link Discovery : Issues, Approaches and Applications (LinkKDD'05)*, Chicago, IL.

[Cai et al., 2011] Cai, Y., Shi, C., Dong, Y., Ke, Q., and Wu, B. (2011). A novel genetic algorithm for overlapping community detection. In Tang, J., King, I., Chen, L., and Wang, J., editors, *ADMA (1)*, volume 7120 of *Lecture Notes in Computer Science*, pages 97–108. Springer.

[Carchiolo et al., 2011] Carchiolo, V., Longheu, A., Malgeri, M., and Mangioni, G. (2011). Communities unfolding in multislice networks. In da F. Costa, L., Evsukoff, A., Mangioni, G., and Menezes, R., editors, *Complex Networks ; Revised Selected Papers form Second International Workshop CompleNet'2010 Rio de Janeiro, Brazil*, pages 187–195.

[Chebotarev and Shamis, 1997] Chebotarev, P. and Shamis, E. V. (1997). The matrix-forest theorem and measuring relations in small social groups. *Automation and Remote Control*, 58 :1505.

[Chen et al., 2009] Chen, J., Zaïane, O. R., and Goebel, R. (2009). Local community identification in social networks. In [Memon and Alhajj, 2009], pages 237–242.

[Chevaleyre et al., 2007] Chevaleyre, Y., Endriss, U., Lang, J., and Maudet, N. (2007). A short introduction to computational social choice. *SOFSEM 2007 : Theory and Practice of Computer Science*, pages 51–69.

[Clauset, 2005] Clauset, A. (2005). Finding local community structure in networks. *Physical Review E*.

[Cordasco and Gargano, 2012] Cordasco, G. and Gargano, L. (2012). Label propagation algorithm : a semi-synchronous approach. *IJSNM*, 1(1) :3–26.

[Corlette and III, 2010] Corlette, D. and III, F. M. S. (2010). Link prediction applied to an open large-scale online social network. In *HT*, pages 135–140.

[Davis et al., 2011] Davis, D., Lichtenwalter, R., and Chawla, N. V. (2011). Multi-relational Link Prediction in Heterogeneous Information Networks. In *2011 International Conference on Advances in Social Networks Analysis and Mining*, pages 281–288. IEEE.

[Donetti and Munöz, 2004] Donetti, L. and Munöz, M. (2004). Detecting network communities : a new systematic and efficient algorithm. *Journal of Statistical Mechanics : Theory and Experiment*, 10.

[Duch and Arenas, 2005] Duch, J. and Arenas, A. (2005). Community detection in complex networks using extremal optimizition. *Physical Review E*, 72.

[Erdös and Rényi, 1959] Erdös, P. and Rényi, A. (1959). On random graphs. *Publ. Math. Debrecen*, 6 :290–297.

[Fortunato, 2010] Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486(3-5) :75–174.

[Fortunato and Barthélemy, 2007] Fortunato, S. and Barthélemy, M. (2007). Resolution limit in community detection. *Proceedings of the National Academy of Sciences,*, 104(1) :36.

[Fouss et al., 2007] Fouss, F., Pirotte, A., Renders, J.-M., and Sarens, M. (2007). Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommandation. *IEEE Transactions on knowledge and data engineering*, 19(3) :355–369.

[Goder and Filkov, 2008] Goder, A. and Filkov, V. (2008). Consensus clustering algorithms : Comparison and refinement. In Munro, J. I. and Wagner, D., editors, *ALENEX*, pages 109–117. SIAM.

[Good et al., 2010] Good, B. H., de Montjoye, Y.-A., and Clauset., A. (2010). The performance of modularity maximization in practical contexts. *Physical Review*, E(81) :046106.

[Gregor, 2010] Gregor, S. (2010). Finding overlapping communities in networks by label propagation. *New Journal of Physic*, 12 :103018.

[Guigourès et al., 2013] Guigourès, R., Boullé, M., and Rossi, F. (2013). Étude des corrélations spatio-temporelles des appels mobiles en france. In Vrain, C., Péninou, A., and Sèdes, F., editors, *EGC*, volume RNTI-E-24 of *Revue des Nouvelles Technologies de l'Information*, pages 437–448. Hermann-Éditions.

[Guimera et al., 2004] Guimera, R., Sales-Pardo, M., and Amaral, L. A. N. (2004). Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70 :025101.

[Gutierrez et al., 2013] Gutierrez, T., Krings, G., and Blondel, V. D. (2013). Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. *CoRR*, abs/1309.4496.

[Hernández and Navarro, 2012] Hernández, C. and Navarro, G. (2012). Compressed representation of web and social networks via dense subgraphs. In Calderón-Benavides, L., González-Caro, C. N., Chávez, E., and Ziviani, N., editors, *SPIRE*, volume 7608 of *Lecture Notes in Computer Science*, pages 264–276. Springer.

[Hubert and Arabie, 1985] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1) :192–218.

[Jaccard, 1901] Jaccard, P. (1901). Etude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Societe Vaudoise des Science Naturelles*, 37 :547.

[Kanawati, 2011] Kanawati, R. (2011). Licod : Leaders identification for community detection in complex networks. In *SocialCom/PASSAT*, pages 577–582. IEEE.

[Kanawati, 2013a] Kanawati, R. (2013a). Co-authorship link prediction in multiplex bibliographical networks. In *Multiplex network workshop - European conference on complex systems (ECCS'13)*.

[Kanawati, 2013b] Kanawati, R. (2013b). A complex network approach for evaluating query similarity metrics. In *International workshop on Web mining (Webi)*, Angers.

[Kanawati, 2013c] Kanawati, R. (2013c). Seed-centric approaches for community detection in complex interaction networks : A comparative review. In *Complexity in social systems : from data to models*, Cergy.

[Kappe et al., 2009] Kappe, F., Zaka, B., and Steurer, M. (2009). Automatically detecting points of interest and social networks from tracking positions of avatars in a virtual world. In [Memon and Alhajj, 2009], pages 89–94.

[Katz., 1953] Katz., L. (1953). A new status index derived from socimetric analysis. *Psychmetrika, 18(1)*, 18(1) :39–43.

[Khorasgani et al., 2010] Khorasgani, R. R., Chen, J., and Zaiane, O. R. (2010). Top leaders community detection approach in information networks. In *4th SNA-KDD Workshop on Social Network Mining and Analysis*, Washington D.C.

[Kolaczyk et al., 2009] Kolaczyk, E. D., Chua, D. B., and Barthelemy, M. (2009). Group betweenness and co-betweenness : Inter-related notions of coalition centrality. *Social Networks*, 31(3) :190–203.

[Labatut, 2012] Labatut, V. (2012). Une nouvelle mesure pour l'évaluation des méthodes de détection de communauté. In *Actes de 3ième Conférence sur les modèle e analyse des réseaux : approches mathématiques et informatiques (MARAMI'12)*.

[Lancichinetti and Fortunato, 2011] Lancichinetti, A. and Fortunato, S. (2011). Limits of modularity maximization in community detection. *CoRR*, abs/1107.1.

[Lancichinetti and Radicchi, 2008] Lancichinetti, A. and Radicchi, F. (2008). Benchmark graphs for testing community detection algorithms. *Physical Review E*, Physical Review E(4) :046110.

[Leicht et al., 2006] Leicht, E. A., Holme, P., and Newman, M. E. J. (2006). Vertex similarity in networks. *Phys. Rev. E*, 73 :026120.

[Leskovec et al., 2010] Leskovec, J., Lang, K. J., and Mahoney, M. W. (2010). Empirical comparison of algorithms for network community detection. In Rappa, M., Jones, P., Freire, J., and Chakrabarti, S., editors, *WWW*, pages 631–640. ACM.

[Li and Song, 2013] Li, J. and Song, Y. (2013). Community detection in complex networks using extended compact genetic algorithm. *Soft Comput.*, 17(6) :925–937.

[Lichtenwalter et al., 2010] Lichtenwalter, R. N., Dame, N., Lussier, J. T., and Chawla, N. V. (2010). New Perspectives and Methods in Link Prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM.

[Lü and Zhou, 2011] Lü, L. and Zhou, T. (2011). Link prediction in complex networks : A survey. *Physica A : Statistical Mechanics and its Applications*, 390(6) :1150–1170.

[Lusseau et al., 2003] Lusseau, D., Schneider, K., Boisseau, O. J., Haase, P., Slooten, E., and Dawson, S. M. (2003). The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations - can geographic isolation explain this unique trait ? *Behavioral Ecology and Sociobiology*, 54 :396–405.

[Meila, 2003] Meila, M. (2003). Comparing clusterings by the variation of information. In Schölkopf, B. and Warmuth, M. K., editors, *COLT*, volume 2777 of *Lecture Notes in Computer Science*, pages 173–187. Springer.

[Memon and Alhajj, 2009] Memon, N. and Alhajj, R., editors (2009). *2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2009, 20-22 July 2009, Athens, Greece*. IEEE Computer Society.

[Milgram and Travers, 1969] Milgram, S. and Travers, J. (1969). An experimental study of the small world problem. volume 32, pages 425–443.

[Mucha et al., 2010] Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980) :876–878.

[Newman, 2006] Newman, M. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E (Statistical, Nonlinear, and Soft Matter Physics)*, 74(3).

[Newman, 2001] Newman, M. E. J. (2001). Scientific collaboration networks. ii. shortest paths, weighted networks, and centrality. *Phys. Rev. E*, 64(1) :16132.

[Newman, 2004a] Newman, M. E. J. (2004a). Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Science of the United States (PNAS)*, 101 :5200–5205.

[Newman, 2004b] Newman, M. E. J. (2004b). Fast algorithm for detecting community structure in networks. *Physical Review E*, 69(6) :066133.

[Ngonmang et al., 2012] Ngonmang, B., Tchuente, M., and Viennet, E. (2012). Local community identification in social networks. *Parallel Processing Letters*, 22(1).

[Nguyen et al., 2009] Nguyen, X. V., Epps, J., and Bailey, J. (2009). Information theoretic measures for clusterings comparison : is a correction for chance necessary ? In Danyluk, A. P., Bottou, L., and Littman, M. L., editors, *ICML*, volume 382 of *ACM International Conference Proceeding Series*, page 135. ACM.

[Orman et al., 2012] Orman, G. K., Labatut, V., and Cherifi, H. (2012). Qualitative comparison of community detection algorithms. *CoRR*, abs/1207.3603.

[Papadopoulos et al., 2010] Papadopoulos, S., Kompatsiaris, Y., and Vakali, A. (2010). A graph-based clustering scheme for identifying related tags in folksonomies. In Pedersen, T. B., Mohania, M. K., and Tjoa, A. M., editors, *DaWak*, volume 6263 of *Lecture Notes in Computer Science*, pages 65–76. Springer.

[Papadopoulos et al., 2012] Papadopoulos, S., Kompatsiaris, Y., Vakali, A., and Spyridonos, P. (2012). Community detection in social media - performance and application considerations. *Data Min. Knowl. Discov.*, 24(3) :515–554.

[Papadopoulos et al., 2011] Papadopoulos, S., Vakali, A., and Kompatsiaris, Y. (2011). Community detection in collaborative tagging systems. In *Community-Built Databases*, pages 107–131.

[Pizzuti, 2012] Pizzuti, C. (2012). Boosting the detection of modular community structure with genetic algorithms and local search. In Ossowski, S. and Lecca, P., editors, *SAC*, pages 226–231. ACM.

[Pons and Latapy, 2006] Pons, P. and Latapy, M. (2006). Computing communities in large networks using random walks. *J. Graph Algorithms Appl.*, 10(2) :191–218.

[Potgieter et al., 2009] Potgieter, A., April, K., Cooke, R., and Osunmakinde, I. (2009). Temporality in link prediction : Understanding social complexity. *The Complexity Society*, 11 :69–83.

[Pujari, 2013] Pujari, M. (2013). Path betweenness centrality : A new topological measure for link prediction. In *Actes de la journée de Fouille de grandes graphes (JFGG'13)*, Saint-Etienne.

[Pujari and Kanawati, 2012] Pujari, M. and Kanawati, R. (2012). Tag recommendation by link prediction based on supervised machine learning. In *Sixth International AAAI Conference on Weblogs and Social Media (ICWSM'2012)*, pages 547–550, Dublin.

[Radicchi et al., 2004] Radicchi, F., Castellano, C., Cecconi, F., Loreto, V., and Parisi, D. (2004). Defining and identifying communities in networks. In *Proc. Natl. Acad. Sci. USA*, pages 2658–2663.

[Raghavan et al., 2007] Raghavan, U. N., Albert, R., and Kumara, S. (2007). Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76 :1–12.

[Rand, 1971] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66 :846ï¿½850.

[Ravasz et al., 2002] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 29 :155.

[Reichardt and Bornholdt, 2006] Reichardt, J. and Bornholdt, S. (2006). Statistical mechanics of community detection. *Physical Review E*, 74(1).

[Resnick et al., 1994] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994). Grouplens : An open architecture for collaborative filtering of netnews. In Smith, J. B., Smith, F. D., and Malone, T. W., editors, *CSCW*, pages 175–186. ACM.

[Roth et al., 2011] Roth, C., Kang, S. M., Batty, M., and Barthelemy, M. (2011). Long-time limit of world subway networks. *CoRR*, abs/1105.5294.

[Salton and McGill, 1983] Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. MuGraw-Hill.

[Schifanella et al., 2010] Schifanella, R., Barrat, A., Cattuto, C., Markines, B., and Menczer, F. (2010). Folks in folksonomies : Social link prediction from shared metadata. *CoRR*, abs/1003.2281.

[Seifi, 2012] Seifi, M. (2012). *Coeurs stables de communautés dans les graphes de terrain*. PhD thesis, Université Pierre et marie Curie (paris 6).

[Shah and Zaman, 2010] Shah, D. and Zaman, T. (2010). Community Detection in Networks : The Leader-Follower Algorithm. In *Workshop on Networks Across Disciplines in Theory and Applications, NIPS*.

[Shah and Sukthankar, 2011] Shah, F. and Sukthankar, G. (2011). Constructing social networks from unstructured group dialog in virtual worlds. In Salerno, J. J., Yang, S. J., Nau, D. S., and Chai, S.-K., editors, *SBP*, volume 6589 of *Lecture Notes in Computer Science*, pages 180–187. Springer.

[Shahabi and Kashani, 2007] Shahabi, C. and Kashani, F. B. (2007). Modelling p2p data networks under complex system theory. *IJCSE*, 3(2) :103–111.

[Sørensen, 1948] Sørensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biol. Skr*, 5.

[Strehl and Ghosh, 2003] Strehl, A. and Ghosh, J. (2003). Cluster ensembles : a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3 :583–617.

[Suthers et al., 2013] Suthers, D. D., Fusco, J., Schank, P. K., Chu, K.-H., and Schlager, M. S. (2013). Discovery of community structures in a heterogeneous professional online network. In *HICSS*, pages 3262–3271. IEEE.

[Szell and Thurner, 2012] Szell, M. and Thurner, S. (2012). Social dynamics in a large-scale online game. *Advances in Complex Systems*, 15(6).

[Tang and Liu, 2010] Tang, L. and Liu, H. (2010). *Community Detection and Mining in Social Media*. Synthesis Lectures on Data Mining and Knowledge Discovery. Morgan & Claypool Publishers.

[Tarissan et al., 2013] Tarissan, F., Quoitin, B., Mérindol, P., Donnet, B., Pansiot, J.-J., and Latapy, M. (2013). Towards a bipartite graph modeling of the internet topology. *Computer Networks*, 57(11) :2331–2347.

[Topchy et al., 2005] Topchy, A. P., Jain, A. K., and Punch, W. F. (2005). Clustering ensembles : Models of consensus and weak partitions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12) :1866–1881.

[Toussaint and Bhattacharya, 1981] Toussaint, G. and Bhattacharya, B. K. (1981). Optimal algorithms for computing the minimum distance between two finite planar sets. *Pattern Recognition Letters*, pages 79–82.

[Watts and Strogats, 1998] Watts, D. and Strogats, S. (1998). Collective dynamics of small world networks. *Nature*, 8(393) :440–442.

[White and Smyth, 2005] White, S. and Smyth, P. (2005). A spectral clustering approach to finding communities in graph. In *SDM*.

[Xie and Szymanski, 2011] Xie, J. and Szymanski, B. K. (2011). Community Detection Using A Neighborhood Strength Driven Label Propagation Algorithm. In *Procof IEEE .Network Science Workshop*.

[Yakoubi and Kanawati, 2012] Yakoubi, Z. and Kanawati, R. (2012). Classification non-supervisée par application d'un algorithme de détection de communautés dans les réseaux complexes. In *SFC'12 : XIX journée de la société Francaise de Classification*, page 4 pages.

[Yakoubi and Kanawati, 2013] Yakoubi, Z. and Kanawati, R. (2013). Leader-driven approach for community detection in complex network. In *proceedings of the international conference on intercations in complex systems*, Orléans.

[Yang and Leskovec, 2012] Yang, J. and Leskovec, J. (2012). Defining and evaluating network communities based on ground-truth. *CoRR*, abs/1205.6233.

[Zachary, 1977] Zachary, W. (1977). An information flow model for conflict and fission in small groups. *Journal of anthropological research*, page 452 :473.

[Zhang and Wu, 2012] Zhang, T. and Wu, B. (2012). A method for local community detection by finding core nodes. In *ASONAM*, pages 1171–1176. IEEE Computer Society.

[Zhou et al., 2009] Zhou, T., Lü, L., and Zhang, Y.-C. (2009). Predicting missing links via local information. *Eur. Phys. J. B*, 71 :623.