Analyse des réseaux sociaux et web sémantique: un état de l'art

Emetteur	Guillaume Erétéo (guillaume.ereteo@orange-ftgroup.com)
Contributeurs	Guillaume Erétéo, Fabien Gandon, Michel Buffa, Patrick Grohan
Relecteurs	Talel Abdessalem
Date de livraison prévue	T0+6: 2009/08/01
Date de livraison	2009/07/20
Workpackage	T3. Social management of shared knowledge representations
Delivrable	T3.2 Analyse des réseaux sociaux et web sémantique: un état de l'art
Référence	ISICIL-DOC-EA1-SNAetWS-20090720
Version	0.2
Destinataires	Membres ISICIL

Projet ISICIL :
Intégration Sémantique de l'Information
par des Communautés d'Intelligence en Ligne

Appel ANR CONTINT 2008

ANR-08-CORD-011-05

Sommaire

A. Objet du document

Ce document constitue le premier résultat d'une thèse réalisée conjointement au sein du laboratoire BIZZ/MUSE et rattachée à l'objet de recherche Health Care and Vertical Application et au sein de l'équipe Edelweiss de l'INRIA de Sophia Antipolis.

Cette thèse constitue aussi une contribution au projet ANR ISICIL qui a pour thème l'Intégration Sémantique de l'Information par des Communautés d'Intelligence en Ligne dont l'un des objectifs est de montrer que non seulement les approches WEB2.0 peuvent bénéficier des apports des plateformes du WEB sémantique mais qu'elles peuvent réellement être améliorées grâce à l'introduction, dans les interactions avec un utilisateur, d'un comportement intelligent produit par des inférences additionnelles. Grâce aux résultats attendus par le projet, on se propose de jeter un pont entre le WEB 2.0 et le WEB sémantique, d'adopter la modélisation ontologique pour représenter des réseaux sociaux, et de fournir une meilleure utilisabilité du WEB 2.0 par des entreprises.

Dans ce cadre d'étude, notre travail de recherche se focalise sur l'utilisation de modèles ontologiques pour représenter et analyser les réseaux sociaux. Un des objectifs scientifiques est d'améliorer l'analyse des réseaux sociaux en réalisant des inférences sur des graphes représentatifs de ces réseaux grâce à l'utilisation d'ontologies dédiées. Cette nouvelle fonctionnalité va permettre dans un premier temps de détecter plus facilement des communautés d'intérêts et dans un second temps, grâce à la conception d'algorithmes adaptés permettant de suivre l'activité de ces communautés, de concevoir des services à valeur ajoutée grâce aux connaissances acquises dans l'étape d'analyse.

L'état de l'art présenté concerne les techniques classiques d'analyse des réseaux sociaux et l'utilisation des technologies du web sémantique pour modéliser les interactions en ligne. La première partie présente la démarche qui consiste à utiliser le modèle de graphe pour représenter un réseau social. On rappelle un certain nombre de définitions qui formalisent les notions manipulées par la théorie des graphes. Dans la seconde partie on présente, d'une part, un certain nombre d'indicateurs (densité, centralité, cycle) destinés à caractériser la structure d'un réseau social et d'autre part un ensemble d'algorithmes qui peuvent être hiérarchiques (agglomératifs ou séparatifs) ou non hiérarchiques (à base d'heuristiques) et qui vont permettre de découper le graphe en un certain nombre de clusters. Dans la troisième partie, on s'attache à fournir une manière de représenter sémantiquement un réseau social au travers d'un ensemble d'ontologies telles que SIOC,[1] FOAF[2] , SKOS[3] et SKOT[4]. La dernière partie présente une architecture permettant

[1] SIOC Semantically-Interlinked Online Communities
[2] FOAF Friend of a Friend
[3] SKOS Simple Knowledge Organisation System
[4] SKOT Social Semantic Cloud of Tags

d'exploiter le meilleur de ces deux approches en s'orientant vers une analyse sémantique des réseaux sociaux.

B. Analyses des réseaux sociaux et web sémantique : un état de l'art

Les interactions des utilisateurs au travers des usages du web 2.0 amènent la communauté scientifique à réfléchir sur les moyens de capter ces usages pour y appliquer les techniques d'analyse des réseaux sociaux. Les applications bien connues à l'origine de l'émergence du web 2.0 sont les blogs, les wikis (ex : wikipedia), les services de social bookmarking (ex : del.ico.us), les sites de partages de médias (ex : youtube, flickr) et bien sûr les sites de réseaux sociaux (ex : facebook, LinkedIn). Ces applications ont considérablement accru la participation, les interactions et le partage entre les utilisateurs du web. L'analyse et la compréhension de tels réseaux sociaux suscitent de vifs intérêts au sein de plusieurs communautés scientifiques.

Le web sémantique fournit des formalismes pour la représentation sémantique des personnes et de leurs usages sur le web. L'ontologie FOAF décrit "les personnes, les liens entre elles, ce qu'elles créent et ce qu'elles font". L'ontologie SIOC décrit "l'information contenue explicitement et implicitement dans les moyens de communication d'internet" comme, par exemple, les blogs. Gruber propose une ontologie des folksonomies [Gruber 2005] et l'ontologie SCOT est un moyen de "représenter la structure et la sémantique des données du social tagging afin de les partager et de les réutiliser". Les ontologies SKOS (représentation de thésaurus et autres ressources linguistiques) et MOAT [Passant et al 2008] (désambigüisation des tags) sont quant à elles souvent utilisées pour modéliser la signification des tags.

En regard de ces moyens de représentation il existe un certain nombre de propositions d'utilisation des méthodes d'analyse des réseaux sociaux pour extraire des informations, comme la construction de réseaux d'accointances ou la détection de communautés d'intérêt. La plupart de ces méthodes d'analyses sont basées sur la théorie des graphes. Par exemple, [Mika 2005] exploite les folksonomies en utilisant la théorie des graphes afin d'identifier des champs sémantiques et des communautés d'intérêt. L'approche de [Paolillo et al 2006] utilise une base d'annotations FOAF pour identifier des communautés d'intérêt. D'autres chercheurs [Anyanwu et al 2007] [Kochut et al 2007] [Alkhateeb et al 2007] [Corby 2008] ont étendu des outils SPARQLafin d'extraire des chemins entre des ressources sémantiquement liées dans les graphes RDF, fournissant ainsi une base pour une représentation et une analyse sémantique d'un réseau social.

1. Représentation d'un réseau social

La première personne à avoir représenté un réseau social est Jacob Levy Moreno au début des années 1930 [Moreno, 1933]. Son objectif étant de visualiser graphiquement un réseau social, il a représenté les personnes par des points et une relation entre deux personnes

par des flèches. Cette représentation est depuis désignée par le terme sociogramme, mais on parlait également de toiles en raison de leur aspect en toile d'araignée. Cette forme de visualisation, aussi peu innovante qu'elle puisse paraître de nos jours, fut un premier outil d'identification rapide des caractéristiques d'un réseau social. Moreno a ainsi introduit le concept d'étoile pour désigner les personnes ayant le plus de relations dans un réseau social, en référence à l'étoile formée par un point et ses connections.

Les mathématiciens ont rapidement fait le rapprochement entre les représentations sociogrammes et la théorie des graphes au sens mathématique. [Scott 2000] passe en revue l'évolution de la représentation des réseaux sociaux . Au milieu du vingtième siècle, Cartwright et Harary sont les premiers à avoir appliqué la théorie des graphes à l'analyse des réseaux sociaux. Le graphe est devenu par la suite la représentation adoptée par toutes les sciences manipulant l'analyse des réseaux sociaux, dont la sociologie, les mathématiques et l'informatique. Les définitions suivantes listes quelques notions manipulées par la théorie des graphes pour les réseaux sociaux:

- Un **sommet** est l'unité de base d'un réseau, il en représente une ressource. Dans un réseau social on parle d'acteur. Le terme nœud est également utilisé pour désigner un sommet.
- Une **arête** est une connexion entre deux sommets. On parle également d'arc ou de lien.
- Une hyperarête (hyperedge) est une arête qui connecte 2 ou plusieurs sommets.
- Une arête est **orientée** si elle ne s'utilise que dans une seule direction. Inversement, on parle d'arête **non orientée** pour une arête qui s'utilise dans les deux directions.
- Une arête est **pondérée** lorsqu'on lui attribue un poids.
- Une arête est étiquetée lorsqu'on lui attribue un label.
- Un **graphe** est défini par un ensemble de sommets et un ensemble d'arêtes.
- Un hypergraphe est défine par un ensemble de sommets et un ensemble d'hyperarête. [Berge 1985]
- Un **graphe orienté** désigne un graphe avec des arêtes orientées.
- Un **graphe pondéré** désigne un graphe avec des arêtes pondérées.
- Un **graphe étiquetté** désigne un graphe avec des arêtes étiqutée.
- Un **graphe multipartite** désigne un graphe avec des sommets de types différents.
- Le **degré** d'un sommet est le nombre de ses arêtes adjacentes.
- Un **chemin** est une séquence d'arêtes qui relie deux sommets.
- Un chemin orienté est une séquence d'arêtes qui relie deux sommets en respectant l'orientation du parcours à chaque arrête.
- Une **géodésique** est l'un des plus courts chemins entre deux sommets donnés.
- Le **diamètre** d'un graphe est le plus long chemin géodésique de ce graphe.
- Un graphe est **complet** lorsqu'il existe une arête entre toute paire de sommets.
- Un graphe est dit **connexe** lorsqu'il existe un chemin entre toute paire de sommets.

Nous utiliserons la notation suivante pour la suite de ce document :

- Nous notons un graphe $G = (V, E)$ avec V l'ensemble des sommets, E l'ensemble des arêtes, $n=|V|$ et le nombre de sommets et $m=|E|$ et le nombre d'arrêtes.
- Un sous graphe de G est noté $G' = (V', E')$ avec $V' \subset V$, $E' \subset E$ et restreint à des arrêtes reliant des sommets de V', $n'=|V'|$ et $m'=|E'|$.

- v_i désigne le i^{ème} sommet.
- (v_i,v_j) désigne une arête entre les sommets v_i et v_j.
- Le degré d'un sommet v_i est noté k_i.
- d_{ij} représente la longueur d'une géodésique entre les sommets v_i et v_j. La moyenne des géodésiques est notée l.

Les graphes non orientés sont adaptés pour les réseaux sociaux avec des relations non orientés. Les graphes orientés sont adaptés pour représenter des relations non symétriques comme les réseaux de confiance par exemple. Les graphes pondérés sont adaptés aux réseaux sociaux qui contiennent différents niveau d'intensités dans les relations. Les graphes étiquetés permettent de représenter différents types de relations.Les graphes multipartites sont adaptés pour des réseaux sociaux incluant différent types de ressources manipulées par les acteurs et qui sont le support d'intéractions.

Nous prendrons comme exemple, le célèbre réseau d'amis du club de karaté de Zachary en 1977, représenté par un graphe non orienté, non pondéré et non étiqueté*(Figure 1*). Ce club a été scindé en deux clubs, les membres du premier sont représentés par des sommets ronds et blancs, les membres du deuxième sont représentés par des sommets carrés et grisés.

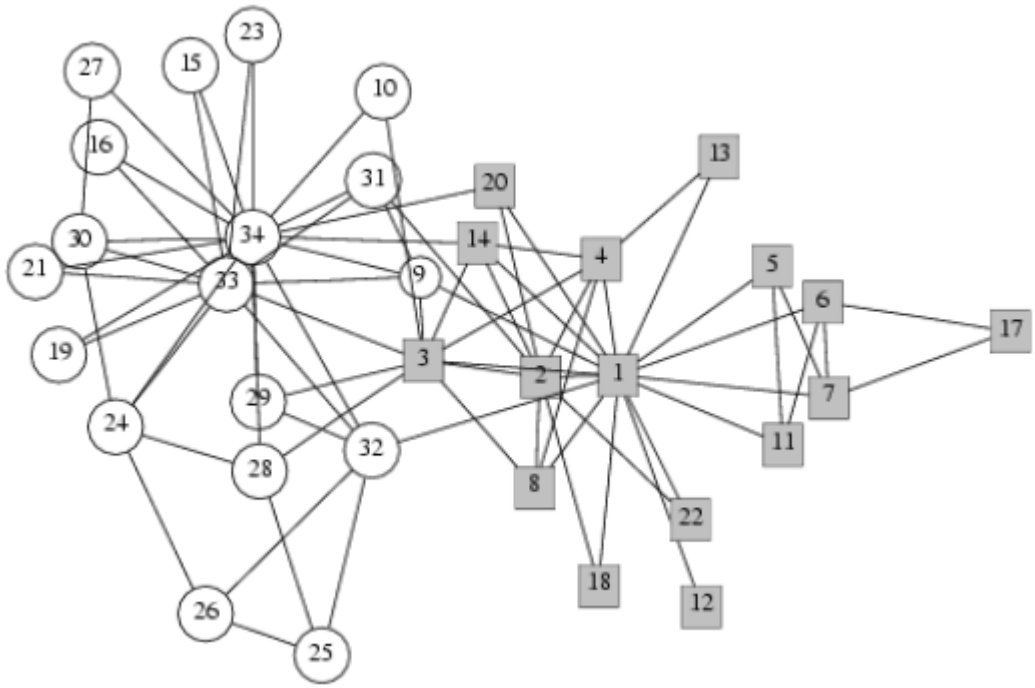


Figure 1 : Le club de karaté de Zachary s'est divisé en deux clubs, les membres du premier club sont représentés par des ronds blancs et les membres du second par des carrés grisés.

La matrice est l'objet mathématique le plus utilisé pour manipuler ces concepts, mais des approches ensemblistes ont aussi était proposées [Scott 2000].

On distingue deux types de matrices dans un réseau social, les matrices d'incidence (figure 2) et les matrices d'adjacence. On parle de matrice d'adjacence lorsqu'on a les mêmes ressources en ligne et en colonne, on obtient ainsi une matrice carrée avec la ligne *i* et la colonne *i* représentant la même

ressource. Un graphe peut ainsi être représenté sous la forme d'une matrice *M* à *n* lignes et *n* colonnes représentant un tableau. Chaque case de ce tableau est notée a_{ij} avec *i* et *j* les numéros respectifs de ligne et de colonne de la case. La valeur contenue dans la case a_{ij} est le poids de la relation entre les ressources v_i et v_j (égal à 1 dans le cas d'un graphe non pondéré), 0 correspond à une absence de relation.

Les matrices d'incidence contiennent deux types de ressources, les lignes représentent un type et les colonnes un autre type. Une matrice d'incidence est convertible en deux matrices d'adjacence représentant chacune les ressources des lignes et des colonnes (figures 3 et 4), les valeurs des cases contiennent les points communs entre les ressources correspondantes dans la matrice d'incidence, a_{ii} n'ayant pas de valeur.

	Projet1	Projet2	Projet3	Projet4
Employé1	1	1	1	0
Employe2	1	0	0	0
Employe3	1	1	1	1
Employe4	0	0	1	1

Figure 2: Exemple de matrice d'incidence indiquant sur quel projet travaille chaque employé

	Employe1	Employe2	Employe3	Employe4
Employe1	-	1	3	1
Employe2	1	-	1	0
Employe3	3	1	-	2
Employe4	1	0	2	-

Figure 3: Matrice d'adjacence des employés déduite de la figure 2, chaque case représente le nombre de projets partagés entre les employés correspondants

	Projet 1	Projet 2	Projet 3	Projet 4
Projet 1	-	2	2	1

Projet 2	2	-	2	1
Projet 3	2	2	-	2
Projet 4	1	1	2	-

Figure 4: Matrice d'adjacence des projets déduite de la figure 2, chaque case représente le nombre d'employés partagés entre les projets correspondants

	V_1	V_2	V_3	V_4	V_5	V_6	V_7	…
V_1	-	1	1	1	1	1	1	…
V_2	1	-	1	1	0	0	0	…
V_3	1	1	-	1	0	0	0	…
V_4	1	1	1	-	0	0	0	…
V_5	1	0	0	0	-	0	1	…
V_6	1	0	0	0	0	-	1	…
V_7	1	0	0	0	1	1	-	…
…	…	…	…	…	…	…	…	…

Figure 5 : Extrait de la matrice d'adjacence du réseau social du club de karaté de Zachary, chaque case précise s'il existe une arête entre les deux sommets (valeur 1) ou pas (valeur 0)

La figure 5 permet de visualiser la matrice d'adjacence du club de karaté de ZAKARY (figure 1)

Un graphe peut être également représenté par une matrice de Laplace qui se différencie par la valeur contenue dans ses cases (k_i désigne le degré du nœud v_i) :

$$a_{ij} = \begin{cases} k_i & si \;\; i = j \\ -1 & si \;\; i \neq j \;\; et \;\; (v_i, v_j) \in E \\ 0 & autrement \end{cases}$$

2. Indicateurs et Algorithmes

a) Indicateurs

La Densité indique la quantité de liens au sein d'un réseau et permet de définir la cohésion d'un réseau social. Selon [Scott 2000] cette mesure peut-être utilisée dans l'optique d'une analyse socio-centrée ou égocentrée Une analyse centrée sur l'individu consiste à mesurer la densité des liens autour d'un nœud donné. Une telle analyse montre notamment l'influence du nœud analysé sur la densité du sous graphe auquel il appartient avec ses voisins. Une analyse socio-centrée considère la densité sur l'ensemble du graphe et mesure la contrainte du réseau sur ses membres. Le calcul de la densité est relatif au nombre maximal de lignes que peut contenir un graphe. Or, ce nombre maximal est lui-même fonction de la taille du graphe, ainsi toute comparaison de densité entre graphes ne fournit aucun résultat significatif. [Scott 2000] proposent une approche intéressante dans le calcul du nombre maximal de connexions dans un réseau social. En effet, la gestion de relations sociales est consommatrice en temps, ainsi le temps limite le nombre de contacts qu'une personne peut conserver et plus un réseau social est grand, moins la densité est élevée. [Dunbar 1998] argumente le coût cognitif inhérent à l'entretien de relations sociales. La densité varie également en fonction du type de relations considérées dans un réseau social, un réseau basé sur des relations amoureuses est beaucoup moins dense qu'un réseau de relations professionnelles notamment en raison des caractéristiques des liens (ex : nature exclusive, différence de temps ou de ressources requis pour l'entretien, etc.). Ainsi le typage des relations dans un réseau social permettrait de paramétrer la densité, par exemple une densité est maximale pour un sommet ayant une relation, dès lors qu'on considère le sous graphe d'une relation exclusive.

La centralité d'un réseau social a été largement discutée. La problématique est de définir ce qui rend un nœud plus central qu'un autre, on parle alors de centralité locale. Plusieurs approches ont été considérées. [Freeman, 1979] reprend l'ensemble de ces approches et en extrait trois principales.

La première approche appelée **centralité de degré** [Nieminem 1974], considère comme centraux les nœuds qui possèdent les degrés les plus élevés du graphe. En effet, ces nœuds suscitent un grand intérêt, sont très visibles, et ont un potentiel élevé à faire circuler l'information, par leur forte connectivité aux autres éléments du réseau. [Scott J. 2000] propose d'étendre la notion de degré à des distances variables, en considérant par exemple tous les voisins à une distance inférieure ou égale à deux.

La **centralité d'intermédiarité** [Freeman, 1979] se concentre sur la capacité d'un nœud à servir d'intermédiaire dans un graphe. Un nœud situé sur un chemin géodésique possède une position stratégique dans la cohésion d'un réseau et dans la circulation de l'information, d'autant plus si ce chemin est unique. Par exemple, un nœud situé sur l'unique chemin reliant deux ensembles

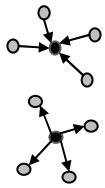
connectés de nœuds possède un fort contrôle sur la communication de ces deux groupes. Plus un nœud est intermédiaire, plus le réseau est dépendant de lui et plus il a de pouvoir.

Enfin, la **centralité de proximité** [Freeman, 1979] mesure la centralité d'un nœud en se basant sur la taille des chemins qui le lient aux autres nœuds. Cette mesure représente la capacité d'un nœud à se connecter rapidement avec les autres nœuds du réseau.

Dans le réseau social du club de karaté de Zachary (figure 1), les sommets 1, 33 et 34 possèdent des degrés bien supérieurs au reste du réseau et sont les plus centraux en termes de centralité de degré et de proximité. Toutefois on constate que les sommets 3, 9, 14, 20, 31 et 32 sont les plus centraux en termes d'intermédiarité, leur absence ou la rupture de leurs liens avec un des deux clubs couperait le réseau en deux groupes

[Freeman, 1979] explicite comment évaluer le caractère centralisé de la structure d'un réseau social. Cette mesure est basée sur les 3 approches explicitées précédemment. La **centralité globale**, ou centralisation, d'un réseau social est calculée à partir des centralités locales des sommets. L'indice de centralité locale choisi détermine le sens de la centralité globale. Le calcul de la centralisation dépend de la définition de centralité locale que l'on considère, à savoir si on considère la centralité comme le contrôle, l'indépendance ou l'activité. En considérant une centralité locale de degré, le calcul de la centralité globale permet d'établir les points dominants, les centres d'intérêts, dans un réseau social, à savoir une activité concentrée autour de certaines ressources. Une mesure de la centralisation d'un réseau social, à partir des centralités locales d'intermédiarité, fournit un indice de la dépendance de l'efficacité de ce réseau par rapport à certains nœuds. Enfin une mesure de la centralité globale d'un réseau, basée sur une centralité locale de proximité, permet de mesurer la performance de la communication dans ce réseau, notamment pour la circulation d'informations.

Pour chacun de ces indices de calcul de centralité locale et globale, Freeman propose une méthode de calcul dépendante de la taille du réseau social et une mesure indépendante permettant de comparer des réseaux sociaux.



Toutefois, [Freeman, 1979] ne considère que les graphes non orientés. Or dans un réseau social, l'orientation des relations contient à elle seule beaucoup de sémantique. Par exemple, pour analyser la propagation d'informations dans un réseau, l'orientation des arcs est primordiale, pour acheminer une information d'un point A à un point B, les chemins allant uniquement de B à A ne sont pas à prendre en compte.

La prise en compte de la direction des relations nous amène à *la notion de prestige*, qui à partir de l'orientation des arcs d'un sommet montre son positionnement par rapport à ses voisins. On détermine deux types de prestiges suivant que l'on considère les arcs entrants ou sortants. Un arc entrant est considéré comme support pour le nœud cible alors qu'un arc sortant représente une

influence de la part de ce nœud. Les trois mesures de centralité évoquées précédemment sont donc nuancées si l'on prend en compte l'orientation des arcs.

La centralité de degré mesurera le support ou l'influence de l'activité des nœuds.

La notion de centralité d'intermédiarité reste la même, mais son calcul est légèrement modifié car l'orientation des arcs doit être considérée pour prendre en compte le sens de circulation de l'information.

La centralité de proximité évalue la capacité d'un nœud à atteindre un autre nœud ou à être atteint par un autre nœud.

[Scott 2000] aborde une approche intéressante en argumentant qu'un calcul de centralité d'un sommet doit prendre en compte la centralité des sommets adjacents. En effet, un point proche d'un point ayant une centralité élevée profite d'une partie de l'avantage offert par cette position. La centralité d'un sommet est ainsi égale à la somme de ses connections, pondérée par la centralité de chacun des sommets correspondants.

D'autres approches se sont concentrées sur la **centralité égocentrée**, qui détermine l'influence d'un nœud par rapport à son voisinage. Cette approche est considérée plus en profondeur par [Everett et Borgatti 2005] qui démontre une corrélation entre la centralité et l'égo-centralité d'un sommet.

En relation avec la centralité locale d'intermédiarité, [Burt 1992] introduit la notion de **trou structural** qu'il définit comme une séparation entre deux contacts non-redondants. Des contacts sont redondants lorsqu'ils sont en contact direct ou qu'ils appartiennent à un même sous-groupe de contacts. Il argumente qu'un trou structural possède un bénéfice informationnel. Les trous structuraux offrent deux atouts majeurs aux personnes contrôlant ces trous. Tout d'abord, ils offrent un bénéfice informationnel, en permettant un accès rapide à des informations non redondantes. L'information entre contacts redondants est généralement partagée, l'apport de nouvelles informations dans un groupe cohérent provient donc de l'extérieur et les trous structuraux sont les canaux de circulation de cette information. Ainsi, les contacts les plus proches des trous structuraux sont mieux informés et plus rapidement. Ensuite les personnes qui contrôlent les trous structuraux possèdent un avantage sur le contrôle de cette information et peuvent en tirer le meilleur profit par leur pouvoir d'intermédiarité. Dans [Burt 2004], Burt démontre que les personnes proches des trous structuraux sont les plus susceptibles d'avoir des "bonnes idées", grâce au bénéfice informationnel apportés par les trous structuraux.

L'ensemble de ces notions nous amène à la **résistance d'un réseau social** au retrait de sommets ou d'arêtes (départ d'une ressource, suppression d'une relation). [Newman 2003] nous offre un aperçu des travaux concernant cette notion. Nous avons vu précédemment que la mesure de la centralisation d'un réseau montre la dépendance d'un réseau par rapport à ses sommets. Cette dépendance peut également être mesurée par l'impact du retrait d'un sommet ou d'une arête sur la connectivité du réseau. En effet, le retrait d'un nœud ou d'une arête stratégique, par exemple un nœud ayant une forte centralité d'intermédiarité ou de proximité, peut augmenter la longueur du

plus court chemin entre de nombreux autres nœuds voir scinder un réseau en deux ou plusieurs réseaux non reliés. Cette mesure s'effectue sur deux types de retraits possibles, des retraits aléatoires et des retraits ciblés. En général, les structures des réseaux sociaux sont assez résistantes à des retraits aléatoires de sommets ou d'arêtes alors qu'un retrait ciblé peut affecter sérieusement ces structures. Par exemple, le retrait d'un pont entre deux groupes de sommets fortement connectés réduit considérablement voire coupe la communication entre ces deux groupes. [Holme et al 2002] rappellent l'ensemble des stratégies possibles d'attaque de réseaux ciblées sur les sommets stratégiques et étend ces stratégies à des attaques basées sur les arêtes.

L'extension de ces stratégies aux arêtes a amené [Holme et al 2002] à étendre les notions de degré et d'intermédiarité des arêtes. Le degré d'une arête est relatif au degré des sommets (min, max, somme ou produit) qu'elle relie alors que l'intermédiarité d'une arête est tout comme l'intermédiarité d'un sommet relative aux chemins géodésiques sur lesquels elle se trouve. L'adaptation de la définition de degré et de l'intermédiarité des sommets aux arêtes est alors utilisée pour appliquer la centralité aux arêtes. Ainsi, les stratégies d'attaques énumérées dans cet article consistent à retirer itérativement les nœuds (resp. arêtes) les plus centraux en termes de degré ou d'intermédiarité, en recalculant ou non les centralités à chaque itération.

Détection de communautés

Nous avons parlé de groupes, de réseaux de contacts redondants, il est maintenant nécessaire de définir la notion de **cohésion dans un groupe** qui a aussi été largement discutée et qui est fortement liée aux notions précédentes. Par exemple, la détection de communautés permet, entre autres, de détecter les communautés non connectées et donc les trous structuraux. En connaissant les groupes fortement connectés, on peut aussi facilement déduire les sommets les plus intermédiaires.

En plus de son lien étroit avec les notions précédemment mentionnées et tout particulièrement la centralité d'intermédiarité, la détection de communauté suscite d'autres intérêts. Dans un réseau social, la détection des communautés permet de déterminer la répartition des acteurs et des activités. Dans l'élaboration de sa théorie sur les trous structuraux, Burt définie la **contrainte de réseau** qui est une mesure de la redondance des contacts d'une personne. Plus les contacts d'une personne sont reliés entre eux, plus le comportement de cette personne est contraint par le réseau. Cette notion se rapproche de la notion de **fermeture de réseau**, argumentée par [Coleman 1988], qu'il définit comme un réseau dense où tous les nœuds sont connectés de manière à connaître l'information détenue par chacun. [Burt 2001] explique comment la redondance des contacts facilite à la fois la sanction et la confiance. En effet, au sein d'un tel réseau, ou sous-réseau, les erreurs d'une personne se propagent rapidement jusqu'à ses contacts directs, augmentant ainsi la probabilité de sanction envers cette personne. Une sanction possible est notamment l'isolement dans le réseau, par la perte de confiance. La facilitation de la sanction tend à éviter la diffusion de mauvaises informations et les mauvais comportements, diminuant ainsi le risque d'accorder sa

confiance à tort. De plus les chemins entre les personnes étant réduits, la perte de qualité dans la transmission de l'information est minimisée. Dans un but éducatif ou en entreprise, l'analyse du réseau social formé par un ensemble de personnes permet de former des groupes de travail productifs et d'améliorer la communication.

[Scott 2000] identifie trois structures principales de groupes fortement connectés: **les composants, les cliques et les cycles**. La première structure abordée par Scott est le **composant.** Un composant est un ensemble de nœuds connectés entre eux par un ou plusieurs chemins avec aucun lien vers d'autres nœuds à l'extérieur du composant. Un composant fort est un composant dont les chemins ne contiennent pas de changement de direction. Un composant faible ne tient pas compte des directions des connexions, seule la présence de liens est prise en compte.

Ensuite, [Scott 2000] traite les **cliques** et les différentes variantes proposées. Une clique est un sous-graphe complet d'un réseau, à savoir un ensemble de nœuds deux à deux connectés. Cette définition manque de souplesse et quelques définitions en proposent des variantes. Une **n-clique** est un ensemble de nœuds reliés entre eux par des chemins de longueur maximale *n*. Toutefois les chemins reliant les sommets d'une n-clique peuvent contenir des sommets exclus de cette clique. Un **n-clan** est une restriction de la définition de n-clique, c'est un ensemble de nœuds tous reliés entre eux par des chemins de longueur maximale *n* et formant un sous graphe d'un diamètre inférieur ou égal à *n*. La figure 6 illustre la différence entre une n-clique et un n-clan. Un **k-plex** est un graphe dont tous les sommets sont reliés à tous les autres sommets sauf *k*.

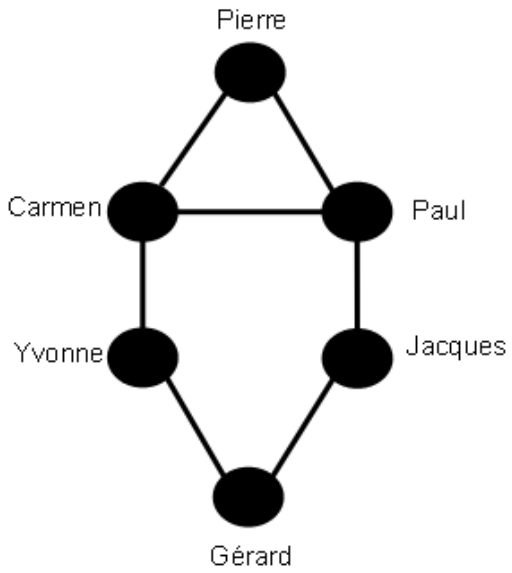


Figure 6 : Pierre, Paul, Jacques, Carmen et Yvonne forment une 2-clique et un 3-clan. L'unique géodésique entre Yvonne et Jacques est de longueur 2 et passe par Gérard.

Enfin la dernière structure que mentionne [Scott 2000] est le **cycle**. Un cycle est un chemin qui revient à son point d'origine. Encore une fois, un cycle fort est un chemin qui ne contient pas de changements de direction alors que la définition d'un cycle faible le permet. Les cycles de longueur

trois sont appelés **triades**. Les réseaux sociaux ont une forte tendance au clustering, à savoir que deux sommets reliés à un même nœud ont une forte probabilité d'être liés entre eux. Cette tendance au clustering est évaluée par un **coefficient de clustering** qui est pour un réseau donné le rapport du nombre de triades sur le nombre maximum de triades possibles pour ce réseau soit :

$$\frac{3\times|TRIADES|}{|TRIPLETS|}$$

avec |TRIADES| et |TRIPLETS| les nombres de triades et de triplets de sommets connectés du réseau. Les triplets connectés du réseau sont les nœuds contenus sur les chemins de longueur deux. Le coefficient de clustering d'un sommet est de la même manière défini par :

$$Ci = \frac{|TRIADESi|}{|TRIPLETSi|}$$

avec $|TRIADES_i|$ et $|TRIPLETS_i|$ le nombre de triades et de triplés connectés contenant le sommet i. On peut ainsi calculer alternativement le coefficient de clustering du réseau à partir des valeurs locales:

$$\frac{1}{n}\sum_i Ci\,.$$

Toujours en relation avec la notion de cycle, Scott introduit les **composants cycliques**. Un composant cyclique est constitué de cycles qui ne se chevauchent pas et qui sont reliés entre eux par des ponts.

Nous noterons également les **LS-SET** qui sont des sous-ensembles de sommets *S* tels que tout sous-ensemble propre de *S* (sous ensemble de S différent de S) a plus de liens vers son complément dans S que vers l'extérieur de S.

Ces définitions sont toutefois trop théoriques et ne correspondent pas à la structure des communautés contenues dans les réseaux sociaux réels. Par exemples, dans le réseau social du club de karaté de Zachary, on distingue clairement de manière visuelle deux groupes, et aucun ne possède strictement les propriétés mentionnées précédemment. De ce fait des notions plus larges ont été prises en compte pour la détection de communautés dans les réseaux sociaux. Ces notions sont abordées dans la partie algorithmique.

Structure d'un réseau social

[Newman 2003] et [Mika 2007] rappellent les caractéristiques relatives à la structure des réseaux sociaux. La principale caractéristique est l'effet de **petit monde** issu de la célèbre expérience de [Milgram 1967]. Ainsi toute personne dans un réseau social est connectée à toute autre personne par un chemin de courte distance. Le plus court chemin entre deux sommets dans un réseau social de taille *n* est de l'ordre de *log(n)*. Ainsi lorsque la taille du réseau augmente, la longueur des plus courts chemins n'augmente que très peu. De plus les membres de ce réseau possèdent la faculté de

trouver facilement ces plus courts chemins [Newman 2003]. Une autre caractéristique est issue de la tendance de l'homme à se socialiser en groupe ce qui donne aux réseaux sociaux une forte **tendance au clustering** et une structure en communautés. Si un sommet A est connecté à un sommet B et que ce sommet B est connecté à un sommet C, alors A et C ont une forte probabilité d'être également connectés, on parle aussi de transitivité. On arrive ainsi à une structure en communauté, à savoir des groupes de sommets avec une forte densité d'arêtes et reliés entre eux par des ponts. Cette socialisation s'effectue avec une tendance à l'affiliation entre des nœuds ayant des propriétés quasi-équivalentes. On constate également que **la distribution des degrés suit une loi de puissance**, à savoir que plus on considère un degré élevé, plus le nombre de sommets qui ont ce degré dans un même réseau est faible. La figure 7 montre la répartition des degrés dans le réseau social du club de karaté du club de Zachari (figure 1).

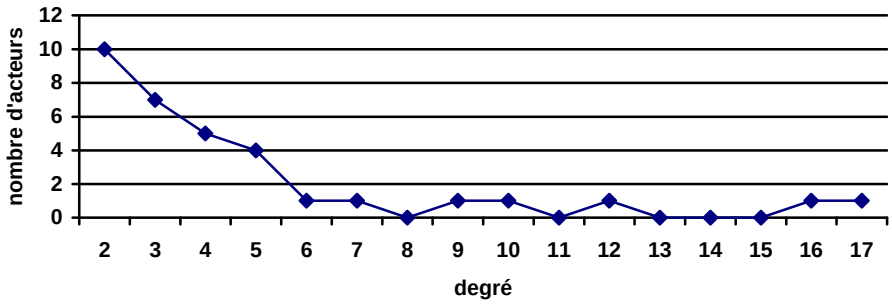


Figure 7: répartition des degrés du club de karaté de Zachary

b) Algorithmes

Nous avons vu précédemment que les principaux indices fournissant des informations importantes sur la *structure* et l'*aspect fonctionnel* d'un réseau social sont la **centralité**, la **répartition des degrés**, la **circulation/qualité de l'information**, la **résistance** du réseau et la **détection des communautés**. L'évaluation de ces indices passe tout d'abord par le calcul des paramètres de base que sont: le degré d'un nœud, les géodésiques, la densité, la détection des clusters. En effet, les calculs qui permettent d'évaluer la centralité sont liés au degré et aux géodésiques. La répartition des degrés est par définition dépendante du calcul du degré des nœuds, à l'instar du calcul du diamètre et des géodésiques.

Les Algorithmes de clustering

Les algorithmes de clustering sont utilisés afin de détecter ces communautés afin d'obtenir une vue globale d'un réseau social.

Algorithmes hiérarchiques

Un premier ensemble d'algorithmes regroupe les algorithmes hiérarchiques. Tout d'abord ils attribuent un poids à chaque paire de sommets ou aux arêtes. Ce poids représente la connectivité de cette paire dans la structure du réseau. Ensuite ils construisent un arbre dont les nœuds sont des groupes de sommets plus ou moins proches. Les nœuds les plus profonds de l'arbre représentent les groupes de sommets les plus proches. Ainsi, plus on remonte dans l'arbre plus on considère de grandes communautés, la racine représentant le réseau complet. Il existe deux catégories, les algorithmes agglomératifs et les algorithmes séparatifs. Ils se distinguent dans la construction de l'arbre et dans la logique d'attribution des poids aux arêtes.

Algorithmes agglomératifs

Dans ces algorithmes, on retrouve trois principaux critères d'attribution des poids aux paires de sommets. Le premier critère d'attribution de poids, est le nombre de chemins qui passent par ces nœuds. Les deux autres critères sont des variantes, les chemins considérés n'ont pas de nœud en commun pour un et pas d'arêtes en commun pour l'autre. Une fois ces poids attribués, ils regroupent itérativement les sommets en considérant les poids par ordre décroissant, jusqu'à avoir considéré tous les poids.

Le principal défaut de ces algorithmes est qu'ils excluent dans la plupart des cas les membres périphériques, plus isolés de leur communauté.

[Donetti et Munoz 2004] utilisent les vecteurs propres de la matrice de Laplace du graphe pour mesurer les similarités entre les sommets, cet algorithme fonctionne en temps $O(n^3)$.

L'algorithme netwalk [Zhou et Lipowsky 2004] est lui "basé sur le temps moyen d'atteinte d'un sommet par des marches aléatoires" pour mesurer la similarité entre les sommets. Sa complexité en temps est de $O(n^3)$.

Algorithmes séparatifs

Ces algorithmes construisent l'arbre de manière inverse. Le poids attribué à chaque arête représente son caractère séparatif entre ses extrémités. L'arbre est construit à partir du graphe entier, en retirant itérativement les arêtes par poids décroissant.

L'algorithme le plus connu est celui de [Girvan and Newman 2002] qui établit les poids des arêtes en fonction de leur intermédiarité, ainsi les nœuds "les plus intermédiaires" sont retirés en premier. Cette technique fournit de très bonnes coupes d'un réseau et est adaptée à la structure d'un réseau social. Toutefois, cet algorithme nécessite le calcul des centralités d'intermédiarité couteux en temps, et possède une complexité en $O(m^2.n)$ avec m le nombre d'arêtes et n le nombre de sommets. Il n'est donc exploitable que sur des petits réseaux. [Bothorel et Bouklit 2008] adapte cet algorithme pour les hypergraphes.

[Fortunato et al 2004] utilisent eux une notion plus stricte de la centralité, offrant un meilleur découpage mais de faibles performances en temps, $O(m^3.n)$.

[Radicchi et al 2004] étendent la notion de coefficient de clustering des sommets aux arêtes et propose un algorithme qui retire les arêtes ayant les coefficients les plus faibles. Le coefficient de clustering d'une arête correspond au nombre de cycles, d'une longueur donnée, auxquels appartient cette arête sur le nombre de cycle possibles en fonction des degrés des extrémités.

Algorithmes à base d'heuristiques

Un certain nombre d'algorithmes non hiérarchiques ont été proposés, ils sont basés sur des heuristiques liées à la structure en communauté des réseaux.

Newman propose un algorithme efficace [Newman 2004] pour des réseaux de grande taille avec une complexité en $O(n.log^2(n))$. Cet algorithme fournit une coupe du graphe optimisant une fonction de modularité :

$$Q = \sum_j (e_{ij} - a_i{}^2)$$

avec e_{ij} la part d'arêtes du réseau qui relie des sommets des groupes *i* et *j* et $a_i = \sum_j e_{ij}$. En d'autres termes, la modularité est, pour un découpage en communautés donné, la différence entre la part d'arêtes intra-communautaires du réseau analysé et la même valeur avec une répartition aléatoire des arêtes. Les valeurs négatives sont ramenées à 0 et la valeur maximale est 1. Cette fonction de modularité est la différence entre le nombre d'arêtes dans un groupe et le nombre d'arêtes attendues en se basant sur la probabilité d'avoir une arête entre chaque sommet. Dans [Newman 2008], il généralise la notion de modularité aux graphes orientés et propose une approche alternative de cet algorithme. [Djidev 2007] réduit le problème du calcul de modularité à celui de coupe minimale pondérée et propose un algorithme en $O(n.log(n)+m)$. [Barber 2007] propose une définition de la modularité pour les graphes bipartites. Enfin [Chen et al 2009] propose une variante qui optimise le degré moyen entrant à l'intérieur de la communauté et minimise le degré sortant des nœuds frontières.

[Wu 2004] fait l'analogie entre un graphe et un réseau électrique et fournit ainsi un algorithme basé sur la simulation de répartition d'un courant électrique. Cette méthode fournie un résultat en temps linéaire en pratique mais impose une contrainte forte qui est de connaître le nombre de clusters à l'avance.

Plusieurs algorithmes s'appuient sur les parcours aléatoires dans un graphe. Dans cette catégorie, l'algorithme de [Pons et al 2005] est le plus performant en temps ($O(n^2.log(n))$ en pratique) mais plus couteux en espace $O(n^2)$, il est basé sur l'hypothèse qu'un parcours aléatoire dans un graphe tend à se retrouver "piégé" dans les parties du graphe fortement connectées correspondant à des communautés. Nous noterons également le plus connu, Markov Cluster Algorithme, qui fonctionne quand à lui en temps $O(n^3)$. [Pons et al 2005] propose un aperçu plus large sur cette approche.

L'algorithme de [Capocci et al 2004] basé sur une analyse spectrale de la matrice d'adjacence, qui prend en considération l'orientation et la pondération des arcs. Cette solution a une complexité de $O(n^2)$ en temps.

L'algortihme par propagation de label de [Raghavan et al 2007] est l'algorithme le plus performant en pratique, mais avec une terminaison non déterministe. Tous les nœuds se voient attribuer un label initial représentant la communauté auquel ils appartiennent. A chaque étape chaque nœud change son label en prenant le plus réparti dans son voisinage. Ce processus itératif amène en pratique à un consensus avec un label unique pour chaque communauté.

Les algorithmes mentionnés précédemment sont les plus utilisés. Toutefois, d'autres algorithmes sont également décrits dans [Danon 2005] [Newman 2004 bis] [Girvan et Newman 2004].

La plupart des algorithmes de clustering, ne considèrent que des graphes non-étiquetés, non orientés et ils fournissent tous des clusters non-recouvrants. En ignorant l'orientation des arêtes nous en perdons toute la signification, alors que la notion de prestige, précédemment abordée, nous en montre la richesse. Le typage des liens dans un réseau social apporte lui aussi beaucoup de sémantique, tout comme le typage des sommets qui permet de décrire un réseau social multipartite. De plus une personne est susceptible d'appartenir à plusieurs communautés, avec des degrés d'implication différents. Ces algorithmes ne lui attribueront qu'une appartenance à la communauté dont elle est le plus proche.

Partant de cette dernière hypothèse, [Pissard 2008] propose l'algorithme FOCAL (Fast Overlapping Clustering ALgorithm) qui restitue des communautés recouvrantes. Son approche est intéressante car elle tient compte des caractéristiques structurelles des réseaux sociaux (petits mondes, transitivité) et des communautés. Toutefois il pose une hypothèse forte liée à son cadre d'application qui considère des communautés de tailles homogènes. L'algorithme SCAN [Xu et al 2007] permet aussi de détecter des communautés recouvrantes. En se basant sur l'idée de base que la structure communautaire d'un nœud est définie par ses voisins, cet algorithme forme des communautés en déterminant un score minimum de similarité structurel entre un nœud et ses voisins.

Le tableau 1 synthétise les catégories et performances des algorithmes précédemment mentionnés.

Type d'algorithme	Reférence	Complexité en temps	Taille des graphes	Caractéristiques de graphe pris en compte
Hiérarchiques agglomératifs	[Donetti et Munoz 2004]	$O(n^3)$	10^3 sommets	Non-typés Non-orientés

				Non-pondéré
	[Zhou et Lipowsky 2004]	$O(n^3)$	10^4 sommets	Non-types Non-orientés Non-pondéré
Hiérarchiques Sépartifs	[Girvan et Newman 2002]	$O(m^2.n)$ pour un graphe non-pondéré $O(m^2.n.log(n))$ pour un graphe pondéré.	10^4 sommets	Non-types Non-orientés Pondérés
	[Radicchi et al 2004]	$O(n^2)$	10^4 sommets	Non-types Non-orientés Non-pondérés
A base d'heuristique	[Newman 2004]	$O(n.log^2(n))$	10^5 sommets	Non-types Non-pondéré, Non-orientés
	[Newman 2008]	$O(n.log^2(n))$	10^5 sommets	Non-types Non-pondérés, orientés
	[Djidev 2007]	$O(n.log(n)+m)$	10^5 sommets	Non-types Non-pondéré, Non-orientés
	[Wu 2004]	$O(n+m)$	10^5 sommets	Non-types Non-orientés
	[Pons et al 2005]	$O(m.n^2)$ dans le pire des cas et	10^4 sommets	Non-types Non-orientés

		O(n².log(n)) en moyenne		Non-pondérés
	[Capocci et al 2004]	*O(n²)*	10^4 sommets	Non-typés, Orientés, pondérés
	[Raghavan et al 2007]	*Terminaison non déterministe*	10^6 sommets	Non-typés, Non Orientés, Non pondérés

Tableau 1 : Catégories et performances des algorithmes de détection de communautés.

Validation d'un découpage en communautés

[Bolshakova et Azuaje 2003] proposent trois indices permettant d'évaluer la qualité d'un découpage en cluster d'un graphe. **L'indice de Silhouette** mesure les propriétés d'isolation et d'hétérogénéité des clusters obtenus. **L'indice de Dunn** et **l'indice de Davies-Bouldin**, calculent le nombre de clusters denses et séparés, ils permettent de déterminer la qualité du nombre de clusters obtenus.

Dans [Girvan et Newman 2004], une approche différente est proposée: le calcul de la **modularité**. Plus le résultat du calcul est proche de 1 plus le découpage est précis. La modularité est actuellement la mesure de référence pour évaluer la qualité d'un découpage en communautés. Dans [Gustafsson et al 2006], une comparaison est effectuée entre la modularité et l'indice de Silhouette et la modularité est mise en avant comme plus pertinente.

[Rattigan 2007] propose quant à lui deux indices complémentaires pour mesurer la qualité d'un découpage en communautés. Ces deux indices sont la proportion d'arêtes intercommunautaires et la proportion d'arêtes intra-communautaires. Ils sont tous les deux compris entre 0 et 1. Un bon découpage en communautés possède un faible taux d'arêtes intercommunautaires et un taux élevé d'arêtes intra-communautaires.

Calcul de la centralité

La centralité permet de détecter les positions stratégiques dans un réseau social. Plusieurs méthodes d'évaluation de la centralité ont été proposées en fonction du critère choisi pour considérer un nœud comme plus central qu'un autre. Ces méthodes sont rappelées dans cette partie avant de rentrer plus en détail sur les algorithmes proposés pour calculer la centralité d'intermédiarité.

[Freeman 1979] propose 2 méthodes de calcul pour chacun des trois indicateurs de centralité locale (degré, intermédiarité, proximité) qu'il présente, une mesure dépendante de la taille du réseau et une mesure indépendante. La première mesure est intéressante pour mesurer l'influence de l'activité d'un nœud dans un réseau alors que la deuxième, indépendante de la taille du réseau, offre un indicateur de comparaison entre des nœuds de différents réseaux. Le fait de s'affranchir de la taille d'un réseau dans un indice permet également de comparer différents résultats locaux issus d'un même réseau, notamment pour comparer différents types de liens et donc différents types de réseaux dans un graphe multipartite. De plus, cela fournit une méthode générique de calcul de centralité globale, basée sur la centralité locale choisie.

La centralité de degré locale d'un nœud est tout simplement son degré.

La méthode de calcul de la centralité d'intermédiarité locale d'un nœud consiste à effectuer la somme des valeurs d'intermédiarité de ce nœud pour chaque couple de nœud du réseau. La valeur d'intermédiarité d'un nœud *A* pour un couple de nœud *B* et *C*, est le rapport du nombre de chemins géodésiques entre *B* et *C* contenant *A* sur le nombre total de chemins géodésiques entre *B* et *C*.

Le calcul de la centralité locale de proximité consiste à effectuer la somme des distances d'un nœud aux autres nœuds du graphe. Cette mesure est plutôt une mesure de "décentralité", à savoir que les nœuds qui obtiennent un score plus élevé sont les moins centraux. Ainsi pour faire un parallèle avec les deux méthodes précédentes, il est opportun de mesurer la centralité de proximité en considérant l'inverse de la somme des distances du nœud aux autres nœuds.

Pour rendre indépendantes ces mesures de la taille du réseau, Freeman propose dans les 3 cas de diviser le résultat obtenu par la valeur maximale possible. La valeur maximale est atteinte à chaque fois par le point central dans un réseau en étoile. Ainsi pour un réseau de taille *n*, la valeur maximale de la centralité de degré est *n-1* et la valeur maximale d'intermédiarité est $(n^2 -3n +2)/2$. Pour le calcul de la centralité de proximité, la somme minimale des distances est *n-1*, ainsi la valeur maximale de la centralité de proximité d'un nœud est le rapport de *n-1* sur la somme des distances avec les autres nœuds du réseau.

Enfin Freeman fournit une formule de calcul de la centralité globale d'un réseau adaptable pour chacun des 3 indices de centralité locale exposés. Le principe est de mesurer l'écart entre la valeur de centralité la plus élevée par rapport à celle des autres nœuds du graphe.

Les définitions précédentes mettent en avant la complexité de calcul de chacun de ces trois indices. Le calcul de la centralité de degré est bien évidemment trivial. Par contre les calculs de centralité d'intermédiarité et de proximité sont bien plus complexes en raison de leur dépendance au calcul des géodésiques. Toutefois la propriété de petit monde des réseaux sociaux crée un lien étroit entre la centralité de degré d'un sommet et sa centralité de proximité. De plus l'indice de centralité le plus significatif est l'intermédiarité qui met en avant les individus les plus influents dans un réseau. L'intermédiarité est ainsi l'indice de centralité le plus considéré dans la littérature. L'ensemble des

travaux mentionnés ci-dessous traitent principalement cet indice, mais certaines des notions et méthodes de calcul fournies s'appliquent également pour la mesure des autres indices.

Algorithmes exacts

Plusieurs algorithmes de calcul d'intermédiarité exacts ont été proposés. Ils sont applicables sur des réseaux de petites tailles, de l'ordre de 10^5 sommets pour le plus performant. Ces algorithmes proposent pour la plupart une version pour les graphes pondérés et non pondérés. Les principaux sont basés sur le calcul des géodésiques dans un premier temps puis sur les sommes des géodésiques où se trouve un sommet, et ce pour chaque sommet [Douglas et Borgatti 1994][Brandes 2001] [Newman 2001]. Les autres sont basés sur une répartition optimale du flot d'information dans le réseau entre les différents chemins possibles [Freeman et Borgatti 1991]. [Latora et Marchiori 2004] proposent une approche qui combine les deux premières. L'algorithme exact le plus performant est celui décrit dans [Brandes 2001], il offre un résultat en $O\ (n+m)$ en espace et en temps $O\ (nm)$ et $O(nm+log^2(n))$, respectivement pour des graphes non pondérés et pondérés. Cet algorithme s'appuie sur un ensemble de lemmes permettant de ne considérer que les calculs indispensables et de réduire ainsi la complexité des méthodes optimales basées sur le calcul des géodésiques. Par exemple, si v_s se trouve sur une **géodésique de** v_r à v_t ,alors $d_{rt} <= d_{rs} + d_{st}$.

Nous noterons l'article de [White et Borgati 1994] qui prend en considération l'orientation des arcs pour le calcul de la centralité d'intermédiarité.

Ulrik Brandes dans [Brandes 2008], effectue un tour d'horizon des variantes proposées pour le calcul de l'intermédiarité. Ces variantes portent notamment sur le niveau d'importance des différents sommets d'un chemin, sur la longueur des chemins à considérer, l'intermédiarité entre les groupes de sommets, l'intermédiarité des arêtes ou encore l'intermédiarité entre des sommets de différents types. Il adapte l'algorithme de [Brandes 2001] pour chacune des variantes discutées.

La prise en considération de différents types est faiblement traitée dans la littérature. Nous notons principalement [Flom et al 2004] (Brandes se base sur son approche), qui traite l'intermédiarité entre des sommets de deux types différents, c'est-à-dire des graphes bi-partites. L'approche de [Everett et Borgati 1999] adapte les principaux concepts de centralité des sommets aux groupes de sommets. Les critères d'appartenance d'un nœud à un groupe de sommets sont très variés et Everett et Borgati fournissent notamment des exemples basés sur le sexe et l'âge. Or, on pourrait considérer tout simplement les nœuds d'un type donné comme critère d'appartenance à un groupe, et considérer leur approche comme une solution au problème de centralité pour les graphes multipartites.

[Everett et Borgatti 2005] fournit une méthode de calcul de la centralité d'intermédiarité égocentrique, à savoir l'intermédiarité d'un nœud donné par rapport au réseau formé par son

voisinage direct. Cette mesure permet d'extraire les sommets les plus influents par rapport à leur voisinage direct.

[Bothorel et Bouklit 2008] propose un algorithme de calcul de la centralité d'intermédiarité pour les hypergraphes.

Algorithmes approchés

Plusieurs autres algorithmes, proposent des estimations de la centralité d'intermédiarité [Radicchi et al 2004][Brandes et Pich 2007][Bader et al 2007][Geisberger et al 2008], fournissant des résultats un peu moins précis mais avec de bien meilleures performances, les rendant utilisables pour des réseaux de l'ordre de 10^6 sommets. La qualité de ces derniers algorithmes dépend de leur technique d'échantillonnage. [Brandes et Pich 2007][Bader et al 2007][Geisberger et al 2008] proposent des approximations à partir d'un échantillon de sommets répartis dans le réseau.

Algorithmes parallèles

Enfin [Bader et Madduri 2006] et [Santos et al 2006] fournissent des contributions majeures en terme de performance avec des algorithmes parallèles du calcul de la centralité d'intermédiarité permettant de traiter des réseaux sociaux de l'ordre du million de sommets avec un résultat exact pour l'un et une approximation pour l'autre. L'algorithme de [Santos et al 2006] est tout particulièrement intéressant par son approche incrémentale qui fournit à tout moment un résultat approximatif de plus en plus précis avec un calcul réparti correspondant bien aux contraintes du web. L'algorithme de [Bader et Madduri 2006] fournit un résultat exact en parallélisant l'algorithme de [Brandes 2001].

Le tableau 2 synthétise les catégories et performances des algorithmes de calcul des centralités d'intermédiarité.

Référence	Exact	parallèle	Complexité	Taille des graphes	Incrémental	Type de graphe considéré
[Newman 2001]	Oui	Non	*O(n.m)* et *O(n.m.log(n))* respectivement pour des graphes non pondérés et pondérés	10^5 sommets	Non	Pondéré Non typés Non orientés
[Brandes 2001]	Oui	Non	*O(n.m)* et *O(n.m + n².log(n))* respectivement pour des graphes non pondérés et	10^5 sommets	Non	Pondéré Non typés Non orientés

			pondérés			
[Geisberger et al 2008] [Brandes et Pich 2007]	Non	Non	~[Brandes 2004] mais approximation à partir de k noeuds.	10^6 sommets	Oui	Graphes pondérés Non typés Non orientés
[Bader et Madduri 2006]	Oui	oui	*O(n.m)* et *O(n.m + n².log(n))* respectivement pour des graphes non pondérés et pondérés	10^6 sommets	Non	Graphes pondérés Non typés Non orientés
[Santos et al 2006]	Non	Oui	Non estimé	10^5	oui	Graphes pondérés Non typés Non orientés

Tableau 2: Catégories et performances des algorithmes de calcul des centralités d'intermédiarité.

Jeux de données couramment utilisés

La qualité et la performance des algorithmes utilisés sont évaluées sur plusieurs jeux de données. Ces jeux de données sont générés ou basés sur des réseaux réels. Concernant la génération de réseaux, trois méthodes principales sont utilisées, la génération de graphes aléatoires [Gilbert 1959], "preferential attachement" [Barabasi et Albert 1999] et "small world" de [Watts et Strogatz 1998]. La génération aléatoire de graphe produit des réseaux n'ayant aucune propriété d'un réseau social. Le modèle de [Watts et Strogatz] reproduit la propriété des petits mondes que l'on retrouve dans tous les graphes. [Barabasi et Albert 1999] fournit une solution permettant de générer un graphe possédant une structure proche de celle des réseaux sociaux, en fournissant notamment une répartition des degrés suivant une loi de puissance. Toutefois ces réseaux étant générés automatiquement, ils servent surtout de témoins et de point de comparaison entre les différentes méthodes. Plusieurs jeux de données réels reviennent alors régulièrement pour juger de l'efficacité et de la qualité d'un algorithme d'analyse de réseau social. Les tous premiers réseaux étudiés étaient construits à partir de questionnaires, en demandant par exemple à des personnes de citer des amis. Le réseau social du club de karaté de Zachary ne possède qu'une trentaine de nœud mais il est souvent utilisé comme preuve du bon fonctionnement d'un algorithme de clustering. Toutefois, l'amélioration de la complexité des algorithmes nécessite des réseaux de grandes tailles pour évaluer leurs performances, juger leur qualité et en observer les limites. L'extraction d'un sous-

ensemble du graphe du web formé par les hyperliens entre les pages est régulièrement utilisée, un crawl du web offre la possibilité d'obtenir des réseaux de très grandes tailles. Les articles scientifiques sont également beaucoup utilisés. On retrouve ainsi deux réseaux extraits à partir des articles scientifiques, le réseau de citation et le réseau de co-auteurs. La source principale servant d'extraction de ce type de réseaux est CiteSeer (http://citeseer.ist.psu.edu/).

c) Conclusion partielle

Nous avons abordé ici les principaux algorithmes de calcul de clustering et d'intermédiarité. Les algorithmes de clustering les plus appréciés pour leur découpage sont les algorithmes hiérarchiques séparatifs basés sur l'intermédiarité. Toutefois la complexité de calcul de l'intermédiarité est une limite éliminatoire pour utiliser ces algorithmes sur de larges réseaux sociaux tels que ceux du web qui contiennent plusieurs millions de sommets. Les approches telles que celles de [Newman 2004] sont donc privilégiées pour les très grands réseaux.

[Radicchi et al 2004] a ouvert la porte à l'utilisation de méthodes approximatives du calcul de la centralité d'intermédiarité pour le clustering. Ainsi, le calcul des centralités d'intermédiarité à partir d'échantillons de [Brandes et Pich 2007][Bader et al 2007][Geisberger et al 2008] sont des pistes intéressantes pour réduire le temps de calcul de l'algorithme de [Girvan et Newman 2002], tout en conservant la même complexité. Nous noterons tout particulièrement l'approche [Rattigan et al 2006] qui indexe la structure du graphe et optimise grandement les calculs de plus courts chemins et des centralités d'intermédiarité. Il utilise ensuite ces index pour optimiser deux algorithmes, dont celui de [Girvan et Newman 2002].

Certains de ces algorithmes mentionnés sont adaptables pour prendre en compte l'orientation, la pondération, l'étiquetage des arêtes et le typage des sommets. Ainsi [Brandes 2008] étend son algorithme [Brandes 2001] pour prendre en compte différentes caractéristiques de graphes pour calculer la centralité d'intermédiarité, ce qui ouvre désormais la porte à l'utilisation de ces différents algorithmes pour adapter [Girvan et Newman 2002].

Enfin nous avons vu sur quels réseaux la qualité et la performance de ces méthodes sont évaluées. Nous allons maintenant montrer que l'avènement du web 2.0 et l'émergence du web sémantique amènent à appliquer les méthodes d'analyse des réseaux sur de nouvelles traces générées par les usages du web.

3. Les réseaux sociaux en ligne

Le web fournit des outils de communications qui s'imposent toujours plus en tant qu'élément majeur des modes d'interaction de notre société. La communication est un élément essentiel de la socialisation et les interactions des utilisateurs du web au travers de leurs usages sont devenues des sources de choix pour extraire et analyser des réseaux sociaux de très grandes tailles (de l'ordre de 10^6 à 10^8 sommets). Les discussions électroniques et la structure en hyperliens du web était les

principales sources du web à disposition des chercheurs jusqu'à l'avènement du web 2.0. La popularité montante des outils collaboratifs du web 2.0 permet d'étudier de nouveaux réseaux avec des acteurs qui fournissent toujours plus d'informations sur eux-mêmes mais également sur les personnes avec qui ils interagissent. Ainsi [Mika 2007] distingue trois catégories de réseaux sociaux sur le web :

- Les réseaux sociaux inférés avec des techniques de web mining: citations entre pages personnels, pagerank, cooccurrence de noms.
- Les discussions électroniques: mails, chat, forum.
- Les applications sociales du web 2.0: outils de publication (wiki, blog, news), réseaux sociaux, sites de partage (contenu, produits, évènements, etc.) et jeux collaboratifs.

[Wellman 2001] argumente que les relations en ligne forment des réseaux sociaux virtuels représentatifs des réseaux sociaux réels. En effet ces réseaux virtuels sont créés à partir d'interactions initiées par des personnes physiques. Cet argument est confirmé par [Mika 2007], mais il souligne le caractère incomplet de ces réseaux sociaux en raison de l'absence en ligne de certaines composantes de la réalité. [Hendler et al 2008] montre que le web 2.0 et le web sémantique amplifient la connectivité des utilisateurs du web et rapprochent qualitativement les réseaux virtuels des réseaux réels.

Cette partie traite dans un premier temps de l'application des techniques d'analyse des réseaux sociaux précédemment évoquées aux réseaux sociaux du web, puis de l'apport du web sémantique à l'analyse des réseaux sociaux.

d) Web 1 et web 2

[Buffa 2008] " dresse l'historique des outils collaboratifs de l'époque précédant l'arrivée du web à nos jours". La "libéralisation" d'internet à la fin des années 80 a très rapidement été suivie "par la création du web par Tim Berners Lee" au début des années 90. Les moyens de communication synchrones et asynchrones proposés par ces technologies ont été massivement adoptés par les particuliers dans un premier temps et par les entreprises ensuite. Les sociologues se sont rapidement intéressés aux réseaux sociaux émergeant de ces nouveaux moyens de communication plus grands et plus faciles à reconstituer qu'à l'aide de questionnaires. L'explosion du volume de connaissance présent sur le web est à l'origine du web mining, discipline destinée à la découverte de cette connaissance sur le web, dont un cas d'application est l'extraction de réseaux sociaux. L'affranchissement des barrières géographiques proposées par internet a été vite perçu comme une aubaine pour la facilitation de la collaboration. Depuis le milieu des années 90 et l'apparition du premier wiki, créé par Ward Cunnigham, les logiciels sociaux n'ont cessé de proliférer sur le web jusqu'à donner aux internautes la possibilité d'améliorer grandement leur visibilité et devenir des acteurs importants dans le paysage du web et dans son développement.

Web mining

[Adamic et Adar 2003] propose une méthode d'extraction des réseaux d'amis des universités de Stanford et du MIT, à partir des pages personnelles des étudiants. Les étudiants de ces universités, au moment de l'étude, avaient pour usage de mettre des hyperliens de leur page personnelle vers la page personnelle de leurs amis. Ainsi, dans un premier temps, les auteurs démontrent que le graphe formé par la structure en hyperliens de ces pages possède les propriétés des réseaux sociaux : "small world", distribution des degrés en loi de puissance, et un taux de clustering élevé. Ensuite, un indice de similarité entre les pages personnelles est défini à partir de la cooccurrence d'éléments textuels et de la présence d'hyperliens entre les pages.

[Kautz et al 1997] [Mika 2005 bis] [Matsuo et al 2006] et [Jin et al 2007] se sont intéressés à l'extraction de réseaux sociaux à partir des cooccurrences de noms sur les pages web. Le principe de ces méthodes, consiste à mesurer la force d'une relation entre deux personnes en se basant sur les cooccurrences de leur nom. [Kautz et al 1997] et [Mika 2005 bis] utilisent le coefficient de Jaccard qui pour une paire de noms *X* et *Y* vaut $n_{X \cap Y}/(n_X+n_y)$ avec n_x et n_y le nombre de pages contenant repectivement les noms *X* et *Y*, et $n_{X \cap Y}$. le nombre de pages contenant à la fois *X* et *Y*. [Matsuo et al 2006] et [Jin et al 2007] utilisent le coefficient de recouvrement qui, avec la même notation, est défini ainsi : $n_{X \cap Y}/min(n_X,n_y)$. Le nombre de pages contenant un nom ou une cooccurrence de noms est obtenu par une requête à un moteur de recherche, Altavista pour [Kautz et al 1997] et Google pour les autres. Ces quatre articles proposent des méthodes d'extraction de réseaux sociaux très proches mais ils exploitent ces réseaux différemment. [Kautz et al 1997] propose un outil d'exploration de son réseau social pour la recherche d'experts. [Mika 2005 bis] et [Matsuo et al 2006] appliquent la cooccurrence entre des noms et des termes afin d'extraire des réseaux d'affiliation. [Mika 2005 bis] exploite ce réseau d'affiliation pour extraire et construire une ontologie légère des termes du web sémantique. [Matsuo et al 2006] propose un outil d'animation de communautés de chercheurs, POLYPHONET, qui extrait et exploite ce réseau d'affiliation. [Jin et al 2007] réapplique les techniques de [Matsuo et al 2006] pour extraire du web des réseaux d'artistes et de grandes firmes japonaises.

Les discussions synchrones et asynchrones

[Tyler et al 2003] construit un graphe d'interaction entre les personnes d'une entreprise à partir de l'analyse des entêtes des emails qui contiennent l'émetteur et le destinataire. Après avoir démontré que ce graphe possède les propriétés inhérentes aux réseaux sociaux il détermine des communautés de pratique en appliquant la méthode de [Wilkinson et Huberman 2002] basée sur l'algorithme de clustering de [Girvan et Newman 2002]. Le découpage en communautés et les personnes appartenant à ces communautés sont validés par des entretiens avec des membres de sept communautés choisies aléatoirement parmi les soixante six communautés détectées.

web 2.0

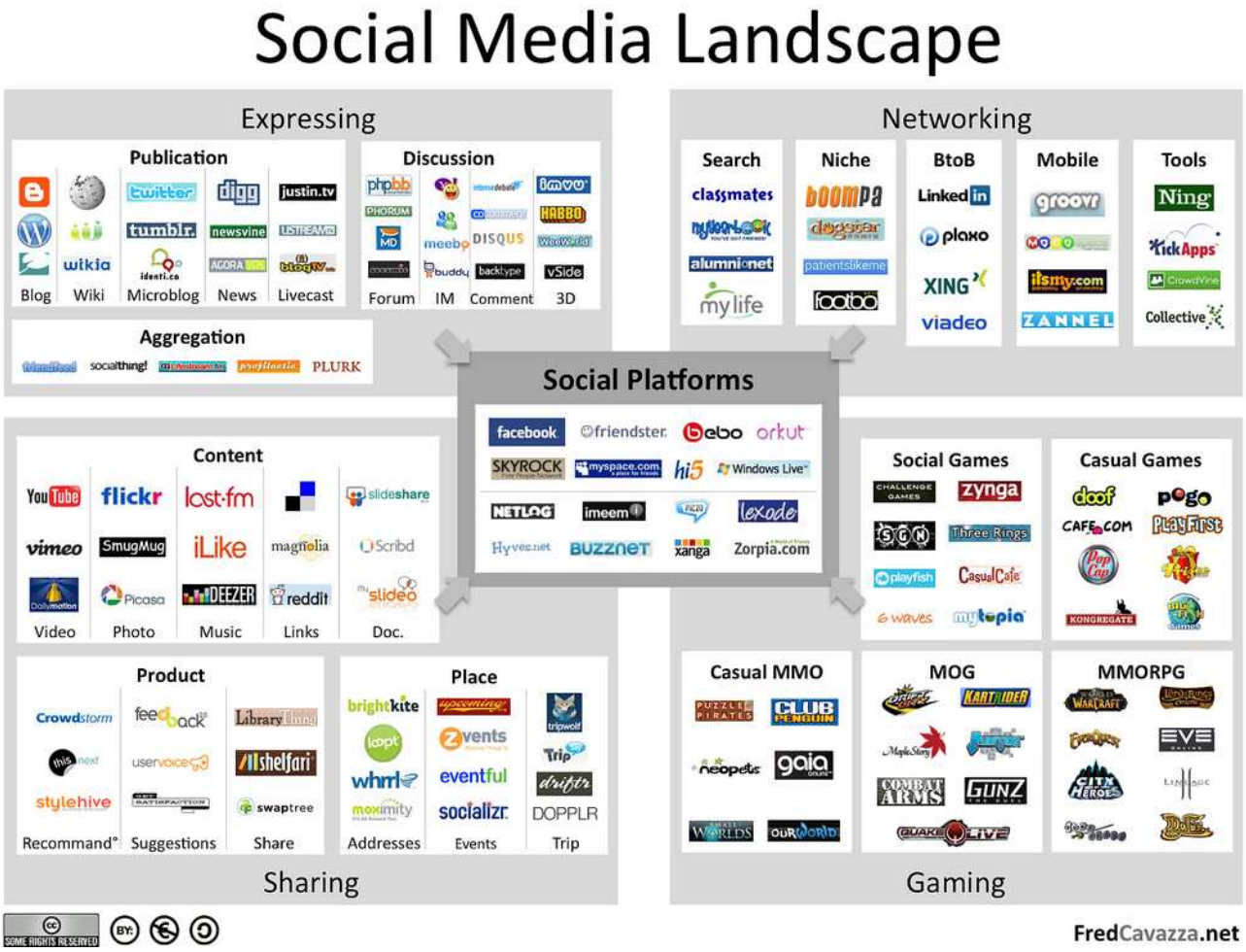


Figure 9: Panorama des médias sociaux proposé par Fred Cavazza [Cavazza 2009]

La figure 9 synthétise le panorama des médias sociaux proposé par Fred Cavazza sur son blog [Cavazza 2009]. Il décompose ces réseaux sociaux en 4 catégories principales, les outils d'expression pour publier, discuter et aggréger sa vie sociale, de réseautage pour rechercher, se connecter et interagir avec des personnes, de partage pour publier et s'échanger des resources, et des jeux en ligne basés sur la collaboration. Certaines plateformes sociales comme Facebook sont extensibles par API et permettent ainsi d'agréger ces différentes pratiques sociales avec des applications dédiées.

Le **social tagging**, qui consiste à classifier collaborativement des ressources en les annotant avec des tags, s'est imposé avec l'émergence du web 2.0 comme l'outil dominant de classification des ressources partagées en lignes (flickr, del.icio.us). [Mika 2005] modélise le social tagging avec un graphe tripartite, les sommets étant des utilisateurs, des tags ou des ressources annotées. Les arrêtes de ce graphe sont ternaires pour représenter l'association d'un tag à une ressource par un acteur. Il considère ensuite de plus près deux sous graphes bipartites. Le premier relie les acteurs aux concepts (tags). Ce graphe permet de déduire un réseau social d'affiliation, les liens sont entre

les acteurs ayant utilisé les mêmes concepts avec des poids représentant le nombre de concepts manipulés conjointement. On en déduit similairement un réseau de concepts, une arête entre deux concepts étant pondérée par le nombre d'utilisateurs utilisant ces deux concepts. Le deuxième sous graphe bipartite relie les concepts aux instances (ressources) et permet d'obtenir un réseau de concepts supplémentaires, un lien entre deux tags est pondéré par le nombre d'instances annotées avec ces deux tags. Ainsi à partir d'un crawl des flux RSS de del.icio.us, Peter Mika crée les graphes simples formés par les deux réseaux de concepts mentionnés et les normalise afin d'obtenir deux graphes de même taille. La densité et le coefficient de clustering moyen sont utilisés pour comparer la cohésion de ces deux réseaux. Il est ensuite démontré que les concepts ayant les coefficients de clustering les plus élevés sont les plus spécialisés. Inversement, les termes avec les coefficients de clustering les moins importants et une forte centralité d'intermédiarité sont les plus généraux. Enfin un algorithme de clustering, basé sur la définition de LS-SET, est appliqué en utilisant [UCINET 2002] afin de déterminer les centres d'intérêts des utilisateurs. [Bothorel et Bouklit 2008] modélise une folksonomie extraite à partir de flickr avec un hypergraphe. Ils proposent une généralisation de l'algorithme de détection de communautés de [Girvan and Newman 2002] pour générer des nuages de tags thématiques et "vérifier s'il apparaît un consensus ou des conflits dans l'utilisation des tags parmi les communautés".

Les sites de réseaux sociaux en ligne sont devenus des applications phares du web 2.0 et connaissent les plus fortes audiences du web. Parmi les premiers, on retrouve Friendster et Orkut, mais les plus connus et les plus visités aujourd'hui sont Facebook et Myspace. Ces sites permettent à leurs utilisateurs de maintenir en ligne leur réseau social réel. La grande audience de ses sites (plus de 100 million d'utilisateurs pour Myspace) et l'accès à leur réseau par API en font ainsi des sources de choix pour analyser des réseaux sociaux de très grandes tailles. En effet, les utilisateurs déclarent explicitement leurs relations, il n'est plus nécessaire d'établir des heuristiques sur leurs usages pour déterminer l'existence de relations entre deux personnes, la nature même de ces relations est fournie. L'un des problèmes les plus discutés ces derniers temps est l'interopérabilité de ces plateformes. Les "agrégateurs" proposent de centraliser le contenu de plusieurs réseaux sociaux. Toutefois ces plateformes sont obligées de manipuler différentes API et l'agrégation d'une nouvelle application nécessite l'apprentissage d'une nouvelle API. Pour palier à cette contrainte, l'initiative "google open social" propose l'interopérabilité entre les réseaux sociaux au travers d'une seule et unique API. La figure 9 représente le réseau social de Guillaume Erétéo sur facebook construit par l'application TouchGraph avec l'API de Facebook.

[Bonneau et al 2009] analyse le réseau facebook des étudiants de Stanford et Harvard à partir seulement des 8 amis affichés sur les profils publics. Ils montrent qu'un petit ensemble du réseau est suffisant pour analyser un réseau social et obtenir des informations essentielles telles que la couverture maximum, la centralité d'intermédiarité ou un découpage en communauté.

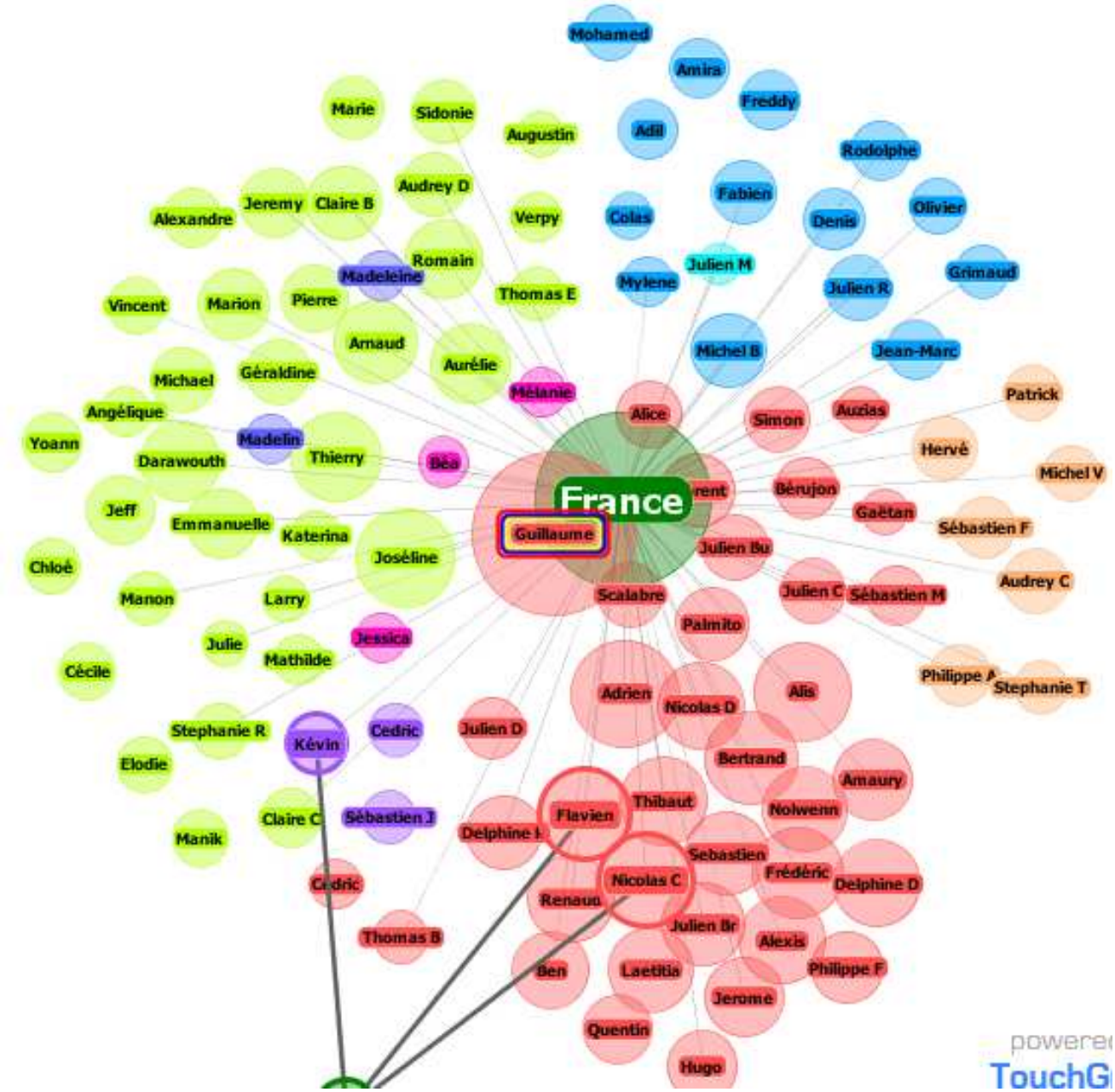


Figure 9: Le réseau social de Guillaume Erétéo extrait par l'application TouchGraph avec l'API facebook

e) Web sémantique

Le web sémantique offre la possibilité aux machines de comprendre et d'exploiter les ressources du web de manière interopérable. Pour cela le w3c propose des formalismes dotés d'une syntaxe XML permettant de modéliser les concepts du web, de les instancier et de les interroger. Les langages OWL (Ontology Web Language) et RDFS (Ressource Description Framework Schema) permettent de décrire une ontologie, "ensemble structuré des termes et concepts fondant le sens d'un champ d'informations" (http://fr.wikipedia.org/wiki/Ontologie_(informatique)). Le langage RDF (Ressource Description Framework) permet de décrire les ressources du web, identifiées par une URI, avec les

propriétés et concepts d'une ontologie, SPARQL en est le langage de requête. La structure de RDF est un modèle de graphe, sur lequel nous sommes intuitivement amenés à appliquer les techniques d'analyses des réseaux sociaux lorsque les ressources décrites incluent les internautes.

Représentation sémantique d'un réseau social

Avec le caractère toujours plus participatif du web, le paysage de la toile est désormais le produit de ses utilisateurs, devenus une des ressources majeures du web. En réponse à ce phénomène social, la communauté du web sémantique propose des modèles ontologiques pour représenter et exploiter les profils des utilisateurs, leurs usages et leur réseau social.

L'initiative la plus célèbre et la plus adoptée est l'ontologie FOAF, Friend Of A Friend. Cette ontologie décrit "les personnes, les liens entre elles et ce qu'elles créent et font". Tout d'abord un large ensemble de propriétés représentent la plupart des concepts nécessaires à la description d'un profil. Par exemple "family_name", "nick" et "interest" permettent respectivement de définir le nom de famille, le surnom et un intérêt d'une personne. Ensuite la propriété "knows" est utilisée pour connecter les profils entre eux et ainsi former le réseau social des profils FOAF. Enfin FOAF modélise les usages des utilisateurs avec des classes pour représenter les ressources manipulées (OnlineAccount, Document, Group…) et des propriétés pour les interactions des utilisateurs avec ces ressources (holdsOnlineAccount, weblog, member…).

Nous avons vu que si FOAF permet de décrire précisément les profils utilisateurs, la modélisation des relations entre utilisateurs et les usages est elle très large. Les bases proposées sont ainsi étendues par plusieurs ontologies. L'ontologie RELATIONSHIP[5] spécialise les relations dans le réseau social en proposant un ensemble de propriétés étendant la propriété "knows" de FOAF. RELATIONSHIP modélise un grand nombre de liens entre les personnes comme les relations familiales, amicales ou encore professionnelles. Les activités en lignes principalement modélisées dans l'ontologie FOAF par la classe "OnlineAccount" et la propriété "holdsOnlineAccount" sont spécialisées dans l'ontologie SIOC. SIOC décrit "l'information contenue explicitement et implicitement dans les moyens de communication d'internet". Pour cela, cette ontologie modélise les concepts issus des applications sociales du web, tels que les "Posts" des forums. SIOC réutilise au mieux les ontologies existantes et s'est presque imposée comme standard sémantique pour certaines applications dédiées, la plus connue étant le moteur de blog WordPress (http://wordpress.org). Ainsi, la gestion des propriétés des documents utilise l'ontologie du Dublin Core[6] qui fournit notamment les propriétés "title", "creator" et "subject". La gestion de l'articulation des concepts manipulés au travers des usages est également déléguée à l'ontologie spécialisée: SKOS. Cette dernière offre la possibilité de définir les labels associés à un concept avec les propriétés "prefLabel" et "altLabel", l'articulation entre ces concepts avec "narrower", "broader" et

[5] http://vocab.org/relationship/
[6] http://dublincore.org/

"related", mais aussi les liens avec les documents et la gestion des significations. La figure 10 illustre l'articulation des ontologies SIOC, FOAF et SKOS.

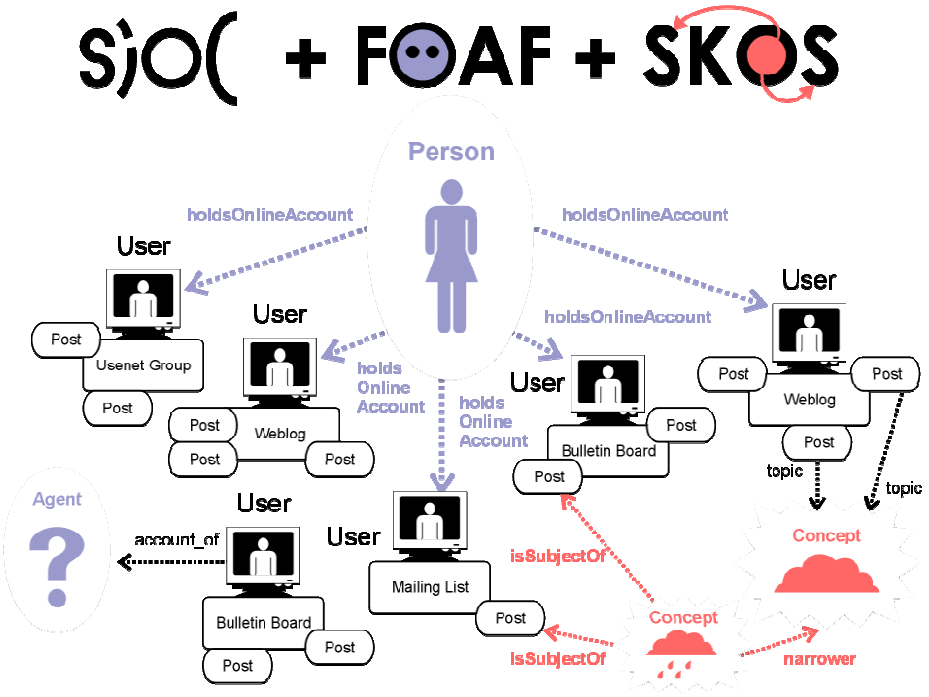


Figure 10 : Exemple d'articulation des ontologies SIOC, FOAF et SKOS

Le social tagging consiste à partager des ressources et à les classifier avec des annotations sous forme de tags. Le fruit du social tagging est une classification de ressources librement établie par les utilisateurs, appelée folksonomie. L'adoption massive de cette pratique par les utilisateurs du web2.0 et la classification proposée par les folksonomies ont amené la communauté du web sémantique à s'intéresser de près à ces usages. Ainsi [Gruber 2005] pose les bases d'une ontologie décrivant les concepts essentiels d'une folksonomie. Il définit tout particulièrement le noyau d'une folksonomie, à savoir l'action de "tagging" composée d'une ressource, d'un tag et d'un utilisateur. [Knerr 2007] s'appuie sur cette base pour proposer une ontologie qui prend notamment en compte la gestion de la vie privée et utilise FOAF pour modéliser les acteurs. L'ensemble des tags manipulés par une personne ou un groupe de personnes est appelé un nuage de tags. Le nuage de tags est l'une des alternatives pour naviguer au sein des ressources d'une folksonomie. L'ontologie SCOT [Kim et al 2007] s'intéresse de près à ces nuages de tags et commence à s'imposer comme moyen de "représenter la structure et la sémantique des données du social tagging afin de les partager et de les réutiliser". SCOT [Kim et al 2007] dans la suite de SIOC s'intègre parfaitement au sein du trio ontologique FOAF, SIOC et SKOS (figure 11). L'initiative MOAT [Passant et al 2008], Mining Of A Tag, complète cet ensemble ontologique en permettant de modéliser la signification des tags. Enfin [Limpens et al 2009] propose une ontologie pour modéliser les points de vues des utilisateurs sur la structuration des folksonomies en leur permettant de valider ou d'invalider des inférences algorithmiques de liens sémantiques.

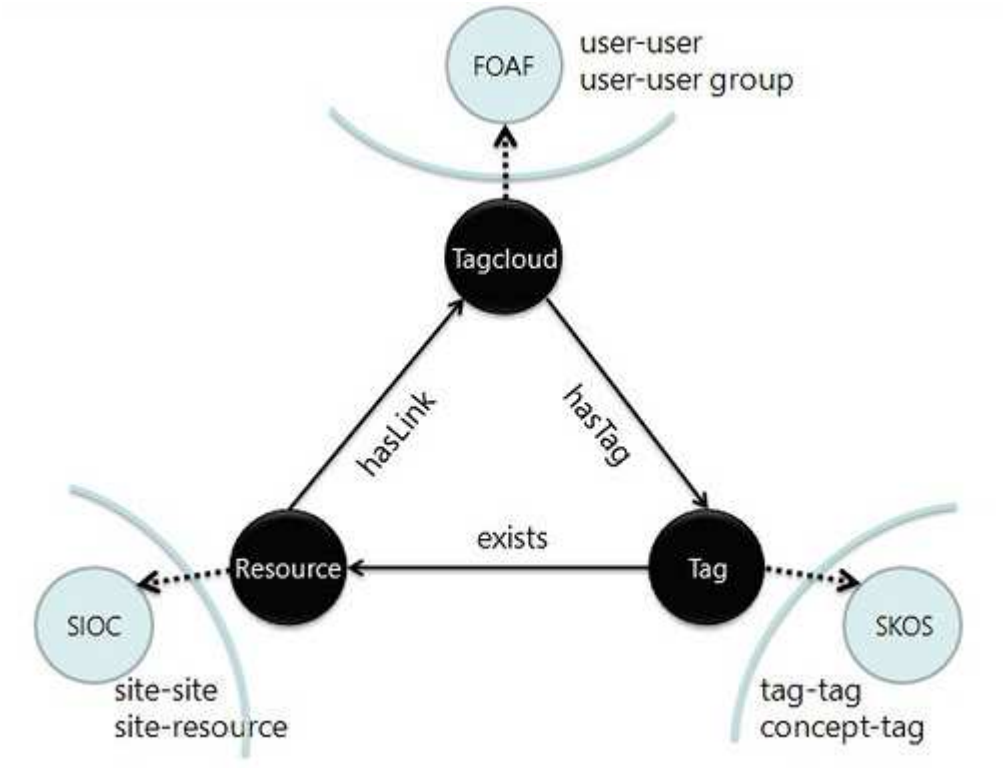


Figure 8 : Articulation de SCOT avec FOAF, SIOC et SKOS

Dans la représentation sémantique des personnes et des usages, il est important de mentionner les microformats. Comme l'argumente [Khare and Celik 2006], cette initiative est importante dans la marche en avant vers un web sémantique qui doit passer par une sémantique légère avant d'atteindre le but attendu par la communauté. Le principe des microformats est d'utiliser les attributs de HTML de manière consensuelle dans l'optique d'ajouter de la sémantique embarquée dans un document XHTML. Les règles mises en place permettent de s'abstenir de l'usage d'une ontologie et de mettre en place un mécanisme de sémantique légère, sans règles d'inférence ni relations de subsomption. On retrouve ainsi un ensemble de microformats (http://microformats.org/wiki) permettant de décrire des personnes, des ressources et des réseaux sociaux. Par exemple, le microformat hCard pour représenter une carte de visite (nom, couriel, adresse, etc.), hResume pour la publication de CV et "XFN" (XTML Friends Network) pour décrire un réseau de connaissances sont des microformats qui permettent de représenter les profils des personnes. De nombreux microformats sont destinés à la définition des ressources et usages du web: "hAtom" est utilisé pour la description des weblogs, "hCalendar" pour les évènements, "xfolk" pour les folksonomies, "votelink" pour les votes, "hReview" pour les revues sur les produits, "XMDP" pour les métadonnées d'une page, "adr" pour les adresses et "geo" pour la géo-localisation. Des micros formats sont aussi disponibles pour définir la nature d'un lien hypertext en utilisant l'attribut "rel" de la balise <a>: rel="tag" pour les tags, rel="enclosure" pour les fichiers attachés, rel="nofollow" pour les liens à ne pas prendre en compte pour les algorithmes d'indexation, rel="directory" pour les liens vers un répertoire, rel="licence" pour les licences et rel="home" pour désigner une page d'accueil. Grâce à leur facilité d'intégration, ces microformats sont largement

utilisés (http://microformats.org/wiki/implementations) notamment dans l'optique de la portabilité des données mais aussi pour une exploitation directe des informations (import d'une carte de visite dans son répertoire, ajout d'un évènement dans son agenda, visualisation sur une carte d'un lieu, etc.). [Adida 2008] propose une méthode pour augmenter la sémantique de ces microformats en les portant en RDFa afin de les relier à des ontologies existantes, telles que celles mentionnées précédemment.

Analyse de réseaux sociaux sémantiques

Plusieurs millions de profils FOAF sont en ligne sur le web. Le succès de FOAF est en grande partie dû à son adoption par des applications sociales ayant une forte audience. On retrouve notamment des fournisseurs de blogs (www.livejournal.net) et des sites de réseaux sociaux (www.tribe.net). Les liens entre les profils FOAF formés par la propriété "knows" définissent un réseau d'accointances. [Finin et al 2005] démontre que ce réseau possède des caractéristiques des réseaux sociaux comme la répartition des degrés suivant une loi de puissance et une structure en communautés. [Paolillo et al 2006] appliquent des techniques d'analyse des réseaux sociaux à une base d'annotations FOAF extraites par un crawl RDF des profils de LiveJournal. Ils construisent deux réseaux sociaux à partir des propriétés "knows" et "interest". Le premier est le réseau d'accointance formé par la propriété "knows" qui spécifie une relation en reliant des profils FOAF. Le deuxième réseau est le réseau d'intérêts extraits à partir de la propriété "interest" qui modélise les centres d'intérêts. Ces deux réseaux sont filtrés pour minimiser leur taille et les temps de calculs. Ainsi, le réseau "knows" est réduit aux 200 profils les plus connectés et le réseau "interest" prend en compte les 500 intérêts les plus mentionnés. Suite à ce filtrage, un clustering hiérarchique (l'algorithme utilisé n'est pas précisé) est appliqué pour extraire des groupes d'utilisateurs et des groupes d'intérêts. Les groupes d'intérêts sont ainsi concentrés en neuf groupes d'intérêts généraux tels que l'art, la sexualité ou la musique. Le résultat du clustering du réseau "knows" est visualisé à différents niveaux de coupe du dendrogramme obtenu afin de déterminer visuellement les indices de centralité et l'articulation des différents groupes. Ces deux réseaux sont ensuite fusionnés en un graphe bipartite pour déterminer les principaux centres d'intérêts de chaque groupe d'utilisateurs. [Goldbeck et al 2003] étend l'ontologie FOAF pour ajouter des propriétés relatives à la confiance afin de modéliser un réseau social de confiance. En se basant sur la notion de contrainte de réseau, un algorithme est ensuite proposé pour déterminer le risque pour une personne d'accorder sa confiance à une autre personne. La prolifération des profils FOAF et la décentralisation de leur production au sein des différents réseaux sociaux posent le problème de la multiplicité des profils pour une même personne. [Goldbeck et Rothstein 2008] transforme ce problème en atout pour le web sémantique avec une méthode de fusion des profils FOAF. Plusieurs propriétés de FOAF décrivent un courriel, un identifiant de messagerie ou une page web personnelle qui sont par nature unique à une personne. Deux profils partageant une valeur identique pour une de ces propriétés désignent donc la même personne et peuvent être fusionnés. Les personnes ayant des profils sur plusieurs sites de réseautage social deviennent ainsi des hubs entre les réseaux sociaux du web.

[San Martin et a 2009] étudie l'expressivité et la complexité de SPARQL. Ils montrent que RDF et SPARQL présentent toutes les caractéristiques pour l'échange, l'interopérabilité, la transformation et l'interrogation de données sociales sur le web. Toutefois ils montrent aussi que la version standard de SPARQL n'est pas assez expressive pour effectuer des requêtes "globales" sur un réseau social, nécessaires pour calculer la plupart des métriques de l'analyse des réseaux sociaux.

D'autres chercheurs ont quant à eux apporté des supports sémantiques à l'analyse des graphes formés par les bases d'annotations sémantiques au format RDF. [Anyanwu et al 2007] et [Kochut et al 2007] proposent des extensions de SPARQL, le langage de requête d'annotation RDF du W3C, afin d'extraire des chemins entre des ressources sémantiquement liées. [Anyanwu et al 2007] propose l'extension SPARQ2L qui permet d'imposer l'inclusion de certaines ressources dans les chemins extraits des contraintes de taille sur leur longueur. L'extension SPARQLeR de [Kochut et al 2007] permet de manipuler plus de caractéristiques sur les chemins :

- Contraintes sur la longueur des chemins.
- Possibilité d'imposer la présence d'une ressource sur les chemins.
- Prise en compte ou non de l'orientation des chemins qui sont par nature orientés dans les graphes RDF.
- Expressions régulières permettant de filtrer la séquence et type des ressources et propriétés contenues dans les chemins.
- Prise en compte du polymorphisme des ressources.

L'extension de [Kochut et al 2007] a été intégrée dans le moteur sémantique CORESE [Corby et al 2004] [Corby 2008], avec certaines modifications syntaxiques. [Ereteo et al 2009] propose un framework pour analyser des réseaux sociaux sémantiques en exploitant les extensions de SPARQL, implémentées dans CORESE.

[Tifous et al 2007] a ouvert la voie à de nouveaux algorithmes d'analyse des réseaux sociaux basés sur des définitions sémantiques des indices d'analyses de ces réseaux. Une ontologie des communautés de pratique est proposée en respect avec la définition de [Wenger 1998] qui extrait trois constituants essentiels à la définition d'un groupe d'individus comme communauté :

- **Un engagement mutuel** : tous les membres sont engagés dans un processus de partage et d'interaction de connaissances, de transmission de compétences, et d'entraide. Cet engagement mutuel est caractérisé par la réciprocité des relations, la confiance et l'ouverture.
- **Une entreprise commune** : une communauté possède une entreprise commune dont la signification dépasse celle d'un objectif ou d'un but. Il s'agit de l'ensemble des processus qui mènent à la constitution de produits communs.
- **Un répertoire partagé** : Un ensemble de ressources communes sont nécessaires à la vie de la communauté. Ces ressources servent de support dans la négociation, et la définition du sens, des choix à adopter. Ces ressources sont un vocabulaire propre à la communauté, des références (personnes, documents, sites…) ainsi que des lieux d'échange (lieu physique, forum, blog, …)

4. Analyse sémantique des réseaux sociaux

La disponibilité en ligne des données des réseaux sociaux sous différents formats, les efforts de modélisation sémantique associés et la structure en graphe du langage RDF nous amène à envisager une nouvelle conception de l'analyse des réseaux sociaux. Les approches actuelles des algorithmes d'analyse des réseaux sociaux sont basées sur des définitions et les caractéristiques des graphes représentant les réseaux sociaux. La sémantique des indicateurs mesurés n'est pas prise en compte. Par exemple, les algorithmes de détection de communautés utilisent des définitions reliées à des caractéristiques de graphes, aucune ne se base sur une définition sociologique d'une communauté. La modélisation ontologique de la définition de [Wenger 1998] par [Tifous et al 2007] peut servir d'exemple pour extraire sémantiquement des communautés.

Les données sociales décrites en RDF forment un graphe typé qui fournit une représentation plus puissante et plus riche des réseaux sociaux du web par rapport aux modèles de graphe classiques l'analyse des réseaux sociaux. Dans [Ereteo et al 2009] nous décrivons un framework basé sur ces représentations enrichies pour proposer une analyse sémantique des interactions en ligne. La Figure 11 illustre la pile d'abstraction que nous utilisons pour effectuer cette analyse.

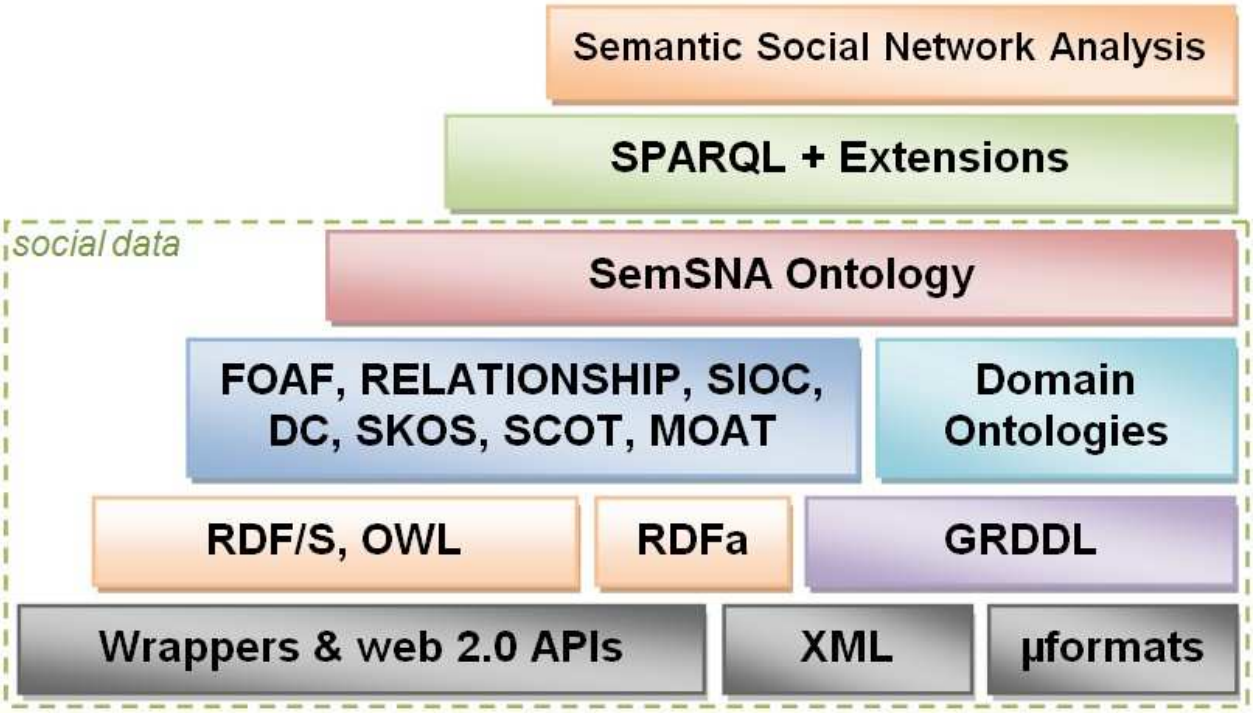


Figure 9: Pile d'abstraction d'une analyse sémantique des réseaux sociaux.

Nous représentons les données sociales en RDF en utilisant les ontologies présentées ainsi que des ontologies de domaines si nécessaire. Nous enrichissons ensuite ces données sociales avec des indicateurs issus de l'analyse des réseaux sociaux qui sont extraits avec des requêtes SPARQL. SemSNA est une ontologie qui décrit les concepts de l'analyse des réseaux sociaux (ex: la centralité). Cette ontologie permet (1) d'abstraire les ontologies construites à partir d'ontologies de domaine

pour appliquer nos outils sur des schémas existants; et (2) d'enrichir les données sociales avec de nouvelles annotations comme, par exemple, les indices de l'analyse des réseaux sociaux décrit précédemment. Ces annotations permettent d'accompagner plus efficacement le cycle de vie d'une analyse en ne calculant qu'une seule fois les indices coûteux et en les actualisant d'une manière incrémentale lorsque le réseau évolue dans le temps. Nous proposons des requêtes SPARQL paramétrables [Ereteo et al 2009] pour calculer les indices de l'analyse des réseaux sociaux et les paramétrer en fonction de la sémantique des liens sociaux considérés. Pour ce faire, nous utilisons le moteur de recherche sémantique CORESE [Corby et al 2004] qui propose des extensions puissantes de SPARQL telle que l'extraction de chemin dans des graphes typés [Corby 2008]. Cette approche permet d'interroger directement le graphe social en tenant compte de la diversité des liens sociaux sans passer par des représentations intermédiaires.

5. Conclusion et discussion

La sociologie possède aujourd'hui de nombreuses réponses sur la socialisation de l'homme. On retrouve ainsi des motifs récurrents dans les réseaux sociaux tels que le phénomène des petits mondes, une structure en communauté et la répartition des degrés suivant une loi de puissances. Comprendre, améliorer ou exploiter le cycle de vie d'un réseau social repose sur un ensemble d'indicateurs majeurs, globaux ou locaux, relatifs à ces paterns. Les indicateurs globaux permettent de comprendre la structure générale du réseau social comme la densité et l'organisation des groupes d'acteurs. Les indicateurs locaux indiquent les positions stratégiques et les acteurs influents d'un réseau social. Une manipulation conjointe et intelligente de ces deux types d'indicateurs permet d'améliorer l'échange d'informations, la créativité ou l'indépendance du fonctionnement d'un réseau par rapport à ses acteurs. Une analyse égocentrique permet d'un autre coté à un acteur d'adapter ses actions par rapport à son entourage direct, en analysant par exemple sa contrainte de réseau ou les risques d'accorder sa confiance, et d'avoir un meilleur accès à l'information.

La taille croissante des premiers réseaux analysés a rapidement apporté des limites aux calculs de certains de ces indices. Si les calculs de densité, de degré ou encore de coefficient de clustering sont triviaux et rapides, ce n'est pas le cas de la centralité d'intermédiarité et de la détection de communautés, riches en informations. Un calcul de centralité d'intermédiarité exacte est contraint par le calcul des géodésiques pour chaque sommet soit une complexité minimale de $O(n.m)$ et donc une limitation à 10^5 sommets pour un temps de calcul raisonnable. Heureusement des approximations de bonne qualité à partir d'un petit pourcentage de sommets offrent de bonnes performances et permettent d'estimer la centralité d'intermédiarité pour 10^6 sommets. L'évaluation d'un découpage en communauté pose deux problèmes principaux, la définition même d'une communauté et la complexité de calcul en fonction de la définition choisie. Certaines définitions sont liées à des caractéristiques particulières des graphes telles que les cliques mais sont bien loin des réalités sociales. D'autres sont liées aux caractéristiques des réseaux sociaux et à l'interprétation d'indices révélateurs de coupes dans le graphe. Les définitions des indices des

réseaux sociaux ne considèrent que peu de sémantique dans les relations sociales. L'orientation et l'intensité d'une relation sont prises en compte dans certaines définitions mais augmentent considérablement la complexité de calcul de la plupart des indices. Les réseaux sociaux contenant plusieurs types de ressources sont en général modélisés par plusieurs graphes simples afin d'éviter l'explosion des complexités de calcul au sein d'un graphe multipartite.

Le web étant devenu un élément de communication majeur de notre civilisation, les interactions massives au sein des outils collaboratifs du web 2.0 sont devenues des sources privilégiées d'extraction de réseaux sociaux pour les sociologues. Les premiers réseaux sociaux du web étaient extraits à partir d'hypothèses basées sur la cooccurrence de noms dans des pages web ou encore les liens entre les pages personnelles. Le web 2.0 a effectué un pas supplémentaire dans la socialisation du web en fournissant toujours plus d'interactions entre les internautes, et réservant même une présence en ligne privilégiée pour les réseaux sociaux réels, au travers de plateformes dédiées à la socialisation. En modélisation sémantiquement les personnes, leurs usages en ligne et les ressources qu'ils manipulent, la communauté du web sémantique ouvre la voie à une approche sémantique de l'analyse des réseaux sociaux. Certains travaux s'orientent déjà dans ce sens en fournissant des outils d'analyse des graphes du web sémantique. L'avènement du web sémantique est un pas supplémentaire pour la qualité de représentation en ligne des réseaux sociaux réels, est-ce aussi une porte ouverte à la sémantisation de leur analyse?

C. References

[Adamic et Adar 2003] L. A. Adamic et E. Adar. Friends ans Neighbors on the web. Social Networks, vol 25, p211-230 (2003).

[Adida 2008] B. Adida : hGRDDL: Bridging micorformats and RDFa. Special Issue of the Journal of Web Semantics on Semantic Web and Web 2.0, Volume 6, Edited by Mark Greaves and Peter Mika, Elsevier, p 61-69 (2008)

[Alkhateeb et al 2007] F. Alkhateeb, J.F. Baget, J. Euzenat, RDF with Regular Expressions INRIA RR-6191, http://hal.inria.fr/inria-00144922/en. (2007)

[Anyanwu et al., 2007] M. Anyanwu, A. Maduko, A. Sheth, SPARQL2L: Towards Support for Subgraph Extraction Queries in RDF Databases, Proc. WWW2007. (2007)

[Bader et Madduri 2006] D. A. Bader, K. Madduri, Parallel algorithms for evaluating centrality in real-world networks. ICPP2006 (2006).

[Bader et al 2007] D. A. Bader, S. Kintali, K. Madduri, and M. Mihail. Approximating betweenness centrality. WAW2007 (2007)

[Barabasi et al, 1999] A. Barabasi and R. Albert. Emergence of scaling in random networks. Science, 286:509-512. (1999).

[Barber 2007] M.J. Barber: Modularity and Community Detection in Bipartite Network. Phys. Rev. E, 76, 036106. (2007).

[Berge 1985] C. Berge. Graphs and Hypergraphs. Elsevier Science Ltd. (1985)

[Bolshakova et Azuaje 2003] N. Bolshakova et F. Azuaje. Cluster validation techniques for genome expression data. Signal processing, 83:825-833. (2003).

[Bonacich 1987] P. Bonacich, Power and centrality: A family of measures. American Journal of Sociology, 92, 1170-1182. (1987).

[Bonneau et al 2009] J. Bonneau, J. Anderson, F. Stajano, R. Anderson: Eight Friends are Enough: Social Graph Approximation Via Public Listings. SocialNets 2009: The Second ACM Workshop on Social Network System. (2009).

[Borgatti 2005] S. P. Borgatti, 2005. Centrality and network flow. Social networks 27 p: 55-71. (2005)

[Bothorel et Bouklit 2008] C. Bothorel et M. Bouklit. An algorithm for detecting communities in folksonomy hypergraphs. 8th International Conference on Innovative Internet Community Systems I2CS 2008 June 16-18, 2008, Schoelcher, Martinique Sponsored by IEEE.

[Bothorel et Bouklit 2008] C. Bothorel et M. Bouklit. Détection de structures de communauté dans les hyper-réseaux d'interactions. AlgoTel'2008, 10èmes Rencontres Francophones sur les Aspects Algorithmiques de Télécommunications, Saint-Malo, 3 - 16 mai 2008.

[Brandes 2001] U. Brandes, A faster algorithm for betweenness centrality. J. Math. Socio 25(2): 163-177 (2001).

[Brandes et Pich 2007] U. Brandes et C. Pich, Centrality estimation in large networks. Journal of Bifurcation and Chaos in Applied Sciences and Engineering 17(7): 2303-2318 (2007).

[Brandes 2008] U. Brandes, On variants of shortest-path betweenness centrality and their generic computation. Social Networks 30 (2): 136-145.

[Buffa et al, 2008] M. Buffa, F. Gandon, G. Ereteo, P. Sander et C. Faron, SweetWiki: A semantic wiki, Special Issue of the Journal of Web Semantics on Semantic Web and Web 2.0, Volume 6, Issue 1, February 2008, Edited by Mark Greaves and Peter Mika, Elsevier, Pages 84-97. (2008).

[Burt 1992] R. S. Burt, *Structural holes. The Social Structure of Competition*, Cambridge, Harvard University Press. (1992)

[Burt 2001] R. S. Burt, Structural Holes versus Network Closure as Social Capital. N. Lin, K. Cook, R. S. Burt: Social Capital: Theory and research. Aldine de Gruyter: 31-56 (2001)

[Burt 2004] R. S. Burt, Structural Holes and Good Ideas. American Journal of Sociology 100(2): 339-399 (2004)

[Cavazza 2009] F. Cavazza. Social Media Lanscape Redux. http://www.fredcavazza.net/2009/04/10/social-media-landscape-redux/

[Chen et al 2009] J. Chen, O. R. Zaiane and R. Goebel, Detecting Communities in Social Networks using Max-Min Modularity, SIAM International Conference on Data Mining (SDM'09), Sparks, Nevada, USA, April 30- May 2, 2009

[Coleman 1988] J. S. Coleman. Social capital in the creation of human capital. The American journal of sociology, Vol 94, Supplement: Organizations and Institutions: Sociological and Economic Approaches to the Analysis of Social Structure. (1988)

[Corby et al 2004] C. Corby, R. Dieng-Kuntz et C. Faron-Zucker, querying the semantic web with the corese search engine. ECAI/PAIS2004 (2004)

[Corby, 2008] Graph Path in SPARQL, Olivier Corby, INRIA, March 2008 http://www-sop.inria.fr/edelweiss/software/corese/v2_4_0/manual/next.php

[Danon 2005] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. Journal of statical Mechanics: Theory and Experiment, 2005(09):P09008. (2005).

[Donetti et Munoz 2004] L. Donetti et M. A. Munoz. Detecting communities: a new systematic and efficient algorithm. Journal of statical mechanics, 2004(10):10012, 2004.

[Ereteo et al 2009] G. Erétéo, F. Gandon., O. Corby, M. Buffa: Semantic Social Network Analysis. Web Science 2009. (2009)

[Everett et Borgatti 1999] M. G. Everett, S. P. Borgatti, 1999. The centrality of groups and classes. Journal of Mathematical Sociology 23 (3), 181 – 201.

[Everett et Borgatti 2005] M. G. Everett, S. P. Borgatti, 2005. Ego network betweenness. Social Networks 1, 215-239.

[Finin et al 2005] T. Finin, L. Ding et L. Zou, Social networking on the semantic web. Learning organization journal 5 (12): 418-435. (2005)

[Flom et al 2004] P. L. Flom, S. R. Friedman, S. Strauss, A. Neaigus. A new measure of linkage between two sub-networks. Connections 26(1): 62-70 (2004)

[Freeman, 1979] L.C. Freeman, Centrality in social networks: Conceptual Clarification. Social Networks. 1, 215-239. (1979).

[Freeman et Borgatti 1991] L. C. Freeman, S. P. Borgatti, Centrality in valued graphs: A mesure of betweenness based on network flow. Social Networks 13: 141-154 (1991).

[Fortunato et al 2004] S. Fortunato, V. Latora, and M. Marchiori. Method to find community structures based on information centrality. Phys. Rev. E 70(5): 056104 (2004)

[Geisberg et al 2008] R. Geisberg, P. Sanders et D. Scultes, Better approximation of betweenness centrality. ALENEX08 (2008).

[Girvan and Newman 2002] M. Girvan and M. E. J. Newman, Community structure in social and biological networks. PNAS 99 (12): 7821-7826 (2002)

[Girvan and Newman 2004] M. Girvan and M. E. J. Newman, Finding and evaluating community structure in networks. Phys. Rev. E, 69:026113. (2004)

[Golbeck et al, 2003] J. Golbeck, B. Parsia, and J. Hendler, Trust network on the semantic web. Proceedings of cooperative information agents (2003).

[Goldbeck et Rothstein 2008] J. Goldbeck et M. Rothstein, Linking social Networks on the web with FOAF. Proceedings of the twenty-third conference on artificial intelligence, AAA08. (2008).

[Gruber 2005] T. Gruber, Ontology of folksonomy: A mash-up of apples and oranges. In Conference on Metadata and Semantics Research MTSR2005 (2005).

[Gustafsson et al 2006] M. Gustafsson, M. Hörnquist et A. Lombardi. Comparison and validation of community structures in complex networks. Phys 367: 559-576 (2006).

[Hendler et Goldbeck 2008] J. Hendler et J.r Goldbeck, Metcalfe's law, web 2.0 and the Semantic Web. Journal of Web semantic 6(1): 14-20, 2008

[Holme et al 2002] P. Holme, B. J. Kim, C. N. Yoon et S. K. Han, Attack vulnerability of complex networks, Phys. Rev. E 65, 056109 (2002).

[Jin et al 2007] Y. Jin, Y. Matsuo, et M. Ishizuka. Extracting a Social Network among Entities by Web mining. ESWC 2007. (2007).

[Kautz et al 1997] H. Kautz, B. Selman, et M. Shah. The hidden Web. AI magazine, Vol. 18, No. 2, pp. 27-35. (1997).

[Khare and Celik 2006] R. Khare et T. Celik, Microformats: a pragmatic path to the Semantic Web. Proceedings of the 15th international conference on World Wide Web.

[Kim et al 2007] H. Kim, S. Yang, S. Song, J. G. Breslin et H. Kim, Tag Mediated Society with SCOT Ontology. ISWC2007. (2007).

[Knerr 2007] T. Knerr, Tagging Ontology – Towards a Common Ontology for Folksonomies. http://tagont.googlecode.com/files/TagOntPaper.pdf (2007)

[Kochut et al 2007] K. J. Kochut et M. Janik, SPARQLeR: Extended SPARQL for Semantic Association Discovery, Proc. European Semantic Web Conference, ESWC'2007, Innsbruck, Austria (2007).

[Latora et Marchiori 2004] V. Latora et M. Marchiori, A measure of centrality based on the network efficiency. Phy 9(6): 188 (2004)

[Limpens et al 2009] F. Limpens, F. Gandon et M. Buffa: Sémantique des Folksonomies: structuration collaborative et assistée. IC2009. (2009)

[Matsuo et al 2006] Y. Matsuo, M. Hamasaki, H. Takeda, T. Nishimura, K. Hasida et M. Ishizuka. POLYPHONET: An advanced social network extraction system. In proceedings WWW 2006 (2006).

[Mika, 2005] P. Mika, Ontologies are us: A unified model of social networks and semantics., in The Semantic Web. Proceedings of the 4th International Semantic Web Conference, ISWC 2005, Galway, Ireland, November 6-10, volume 3729 of Lecture Notes in Computer Science, p. 522–536: Springer.

[Mika 2005 bis] P. Mika, Flink: Semantic Web Technology for the Extraction and Analysis of Social Networks. Web Semantics: Science, Services and Agents on the World Wide Web, Vol. 3, No. 2-3., pp. 211-223. (2005).

[Milgram 1967] S. Milgram. The Small World Problem. Psychology Today, 1(1): 61 – 67. (1967).

[Moreno 1933] J.L. Moreno, Emotions mapped by new geography, New York Times (1933).

[Newman 2001] M. E. J. Newman. Scientific collaboration networks. Shortests paths weighted networks, and centrality. Phys Rev 64: 016132 (2001)

[Newman 2003] M. E. J. Newman, The structure and function of complex networks. SIAM Review 45, 167-256 (2003).

[Newman 2003 bis] M. E. J. Newman. A measure of betweeness centrality based on random walks. Cond-mat/0309045 (2003)

[Newman 2004] M. E. J. Newman, Fast algorithm for detecting community in networks. Phys. Rev. E 69, 066133 (2004).

[Newman 2004 bis] M. E. J. Newman, Detecting community structure in networks. Eur. Phys. J. B, 38:321-330. (2004).

[Newman 2008] E. A. Leicht, M. E. J. Newman, Community structure in directed networks, Phys. Rev. Lett. 100, 118703 (2008).

[Nieminem 1974] N. J., On Centrality in a graph". Scandinavian Journal of Psychology 15:322-336.

[Paolillo et al 2006] J. C. Paolillo and E. Wright, Social Network Analysis on the Semantic Web: Techniques and Challenges for Visualizing FOAF, in Book Visualizing the semantic WebXml-based Internet And Information (2006).

[Pissard 2008] N. Pissard. "Etude des interactions sociales mediates: methodologies, algrithmes, services". Thèse de doctorat.

[Passant et al 2008] A. Passant, P. Laublet. Meaning Of A Tag: A Collaborative Approach to Bridge the Gap Between Tagging and Linked Data. LDOW2008. (2008).

[Pons et al 2005] Pa. Pons and M. Latapy. Computing communities in large networks using random walks. ISCIS2005 (2005)

[Radicchi et al 2004] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. Proceedings of national Academy sciences USA 101, p: 2658-2663 (2004)

[Raghavan et al 2007] R.N. Raghavan, R. Albert, S. Kumara: Near Linear Time Algorithm to Detect Community Structures in Large Scale Network. Phys. Rev. E, 76, 036106. (2007)

[Rattigan et al 2006] M. J. Rattigan, M. Maier, D. Jensen. Using structure indices for efficient approximation of network properties. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 357-366. (2006)

[Rattigan et al 2007] M. J. Rattigan, M. Maier, D. Jensen. Graph clustering with network structure indices. International Conference on Machine Learning (2007).

[San Martin et al 2009] M., San Martin, C., Gutierrez: Representing, Querying and Transforming Social Networks with RDF / SPARQL. ESWC09. (2009).

[Santos et al 2006] E. E., Santos, L. Pan, D. Arendt, M. Pittkin: An Effective Anytime Anywhere Parallel Approach for Centrality Measurements in Social Network Analysis. IEEE2006 (2006)

[Scott 2000] J. Scott, Social network analysis, a handbook. Deuxième edition, Edition Sage. (2000).

[Tifous 2007] A. Tifous, A. E. Ghali, R. Dieng-Kuntz, A. Giboin, C. Evangelou, G. Vidou, An Ontology for Supporting a Community of Pratice. K-CAP'07 (2007).

[Tyler et al 2003] J. R. Tyler, D. M. Wilkinson, et B. A. Huberman. Email as spectroscopy: automated discovery of community structure within organizations. International Conference on Communities and Technologies p 81-96, Deventer, The Netherlands, 2003.

[UCINET 2002] S.P. Borgatti, M.G. Everett et L.C. Freeman. Ucinet for Windows: Software for Social Network Analysis. Harvard, MA: Analytic Technologies. (2002).

[Wellman 2001] B. Wellman. Computer Networks As Social Networks. Science 293, 2031-34 (2001).

[Wenger 1998] E. Wenger, Communities of Practice: Learning as a Social System Thinker (1998)

[White et Borgatti 1994] D. R. White et S. P. Borgatti, Betweenness centrality measures for directed graphs. Social Networks 16, p 335 - 346 (1994).

[Wilkinson et Huberman 2003] D. M. Wilkinson et B. A. Huberman. A method for finding communities of related genes. In proceedings of the national Academy of sciences. (2003)

[Wu 2004] Fang Wu and Bernando A. Huberman, Finding communities in linear time: a physics approach. Hp Labs (2004)

[Xu et al 2007] X. Xu, N. Yuruk, Z. Feng and T. A. J. Schweiger. SCAN: a Structural Clustering Algorithm for Networks. In KDD, pages 824.833. (2007)

[Zhou et Lipowsky 2004] H. Zhou et R. Lipowsky. Network browniam motion: A new method to measure vertex-vertex proximity and to identify communities and subcommunities. International conference on computational science, p: 1062-1069 (2004).