UNIVERSITÉ DU QUÉBEC À MONTRÉAL

MODÈLE DE SOCIALISATION AUTOMATIQUE : POUR LA CRÉATION DE COMMUNAUTÉS D'INTÉRÊT

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN INFORMATIQUE COGNITIVE

PAR

MÉLANIE LORD

JUIN 2013

REMERCIEMENTS

Ce document représente l'aboutissement d'un long processus de recherche au cours duquel j'ai côtoyé diverses personnes qui ont eu, de près ou de loin, une influence positive sur l'accomplissement de ce projet.

Je veux tout d'abord remercier Daniel Memmi et Pierre Poirier, professeurs à l'Université du Québec à Montréal, de m'avoir dirigée dans mes travaux de recherche. Ils ont su me guider et me conseiller de façon judicieuse, tout au long de cette thèse. Je les remercie aussi pour leur disponibilité, leur générosité et leur compréhension qui m'ont grandement motivée à mener à bien ce projet.

Aussi, je voudrais exprimer ma reconnaissance à Hafedh Mili, professeur à l'Université du Québec à Montréal, parce qu'il m'a accordé de son temps, à plusieurs reprises, en me permettant ainsi de profiter de son expérience et de ses remarques pertinentes.

Mes remerciements vont de plus à mes amis et à ma famille. En particulier, je remercie mon amie Ghizlane avec qui j'ai souvent discuté de ma thèse et que j'estime énormément. Je remercie mon amie Madeleine pour sa gentillesse, sa compréhension et son enthousiasme. Je remercie mes cousines Myriam et Valérie pour leur amitié et leurs encouragements continuels.

Enfin, je remercie mon père et ma mère pour leur amour, leur présence et leur soutien constant.

TABLES DES MATIÈRES

LISTE DES FIGURES

LISTE DE TABLEAUX

RÉSUMÉ

Le développement rapide des technologies comme Internet et le Web a remarquablement intensifié la prolifération et la décentralisation de l'information à travers les nombreux réseaux électroniques et communautés virtuelles qui ne cessent de croître. Cette effervescence pose problème lorsqu'il s'agit de trouver de l'information pertinente dans le cadre de besoins spécifiques. Par ailleurs, ces mêmes technologies sont aussi responsables de l'ouverture ainsi que de la décentralisation des communautés traditionnelle et, plus récemment, de la multiplication de divers médias sociaux, qui permettent aujourd'hui d'avoir accès à une grande variété d'individus, provenant de divers milieux sociaux.

En effet, nos contacts sociaux sont des atouts précieux lorsqu'il s'agit de trouver de l'information utile et pertinente. Les connaissances acquises par échanges sociaux ont l'avantage non négligeable d'avoir déjà été traitées cognitivement, d'avoir été intériorisées, élaguées, contextualisées, élaborées, etc. Dans un contexte où l'information surabonde, de pouvoir profiter de ce prétraitement de l'information, par l'intermédiaire de nos contacts sociaux, est un avantage majeur pour acquérir plus rapidement les connaissances qui nous sont vraiment utiles. Toutefois, que ce soit dans nos réseaux sociaux réels ou virtuels, cette grande accessibilité à de nombreux individus ne facilite pas la localisation de ceux qui sont vraiment pertinents et intéressants. C'est par le biais de la socialisation que nous entretenons et renouvelons nos réseaux de contacts personnels, mais il faut du temps, de la motivation et un certain talent pour trouver les bonnes « connexions ». C'est dans cette optique que nous proposons, dans le cadre de cette thèse, un modèle de socialisation automatique qui favorise la rencontre des individus intéressants et utiles les uns pour les autres en formant et maintenant automatiquement des communautés d'intérêt, au sein d'un réseau social dynamique.

Plus précisément, les règles d'évolution du modèle proposé s'inspirent des mécanismes de socialisation qu'on observe dans nos réseaux sociaux habituels, et donc, gèrent un réseau complètement décentralisé (contrairement aux sites de réseautage en ligne). D'un point de vue informatique, nous avons conçu notre modèle afin qu'il soit utilisable pour l'implémentation d'applications distribuées basées sur la formation de communautés d'intérêt au sein de réseaux sociaux virtuels. D'un point de vue sociocognitif, nous proposons aussi ces mécanismes de socialisation ainsi formalisés comme processus évolutifs des réseaux sociaux usuels, qui pourraient expliquer, dans une certaine mesure, les propriétés structurales typiques et récurrentes qu'on y retrouve.

Mots-clés : Modélisation sociale, Socialisation, Réseaux sociaux, Médias sociaux, Analyse de réseaux sociaux, Structures des réseaux sociaux, Systèmes collaboratifs, Systèmes décentralisés.

ABSTRACT

The rapid growth of technologies such as Internet and the Web has remarkably intensified the proliferation and the decentralization of information through electronic networks and virtual communities, which continue to grow. This profusion of information is problematic when it comes to locate relevant informations that meet specific needs. Moreover, these same technologies are also responsible for more open and decentralized societies and, more recently, for the proliferation of various social media, that provide access to a wide variety of individuals, from different social environments.

Indeed, our social contacts are valuable assets when it comes to find relevant information. Knowledge acquired by social exchanges has the advantage to have already been processed cognitively, to have been internalized, pruned, contextualized, developed, etc. In a context where information is overabundant, to benefit from this human filtering of information, through our social contacts, is a major advantage to more quickly acquire the knowledge that is really useful to us. However, in our social networks, whether real or virtual, this great access to a wide range of people does not facilitate the task of finding those that are really relevant and interesting. It is through socialization that we maintain and renew our networks of personal contacts, but it takes time, motivation and talent to find the "right connections". It is from this perspective that we propose, in this thesis, a model of automatic socialization that promotes interesting and useful encounters between individuals by automatically forming and maintaining communities of interest, within a dynamic social network.

More precisely, the evolution rules of our model are based on mechanisms of socialization that take place in our usual social networks, and therefore, manage a completely decentralized network (unlike common online social network). From a functionnal point of view, we have designed our model so that it can be implemented and used as a basis for distributed applications of social networks and communities of interest. Morevover, from a socio-cognitive point of view, we also propose these formalized socialization mechanisms as evolutionary social processes, which may explain, to some extent, typical and recurrent structural properties found in real social networks.

Keywords : Social modeling, Socialization, Social networks, Social media, Social network analysis, Social network structures, collaborative systems, decentralized systems.

CHAPITRE I

INTRODUCTION

1.1 Mise en contexte de la recherche

Ce projet de recherche, portant sur la formation et le maintien de communautés d'intérêt, se situe selon nous dans un contexte plus large que nous aimerions expliciter avant de décrire le travail effectivement réalisé. Ce contexte comporte plusieurs aspects différents, social, économique, cognitif, et technique, mais qui forment un système globalement cohérent. On peut distinguer en gros les aspects suivants :

Aspect social

Depuis deux siècles environ, nous sommes passés de petites communautés traditionnelles relativement fermées à des réseaux ouverts et flexibles, en constant remaniement. Plusieurs de nos interactions sociales prennent place dans ce nouveau cadre, notamment dans le domaine du travail. On ne peut pas comprendre le fonctionnement social réel si l'on ignore ce changement fondamental dans la structure sociale (Weber, 1956 ; Simmel, 1989).

Aspect économique

Les économies modernes incorporent une proportion croissante de connaissances, qui sont la condition de leur productivité et de leur croissance. Ces connaissances techniques ou portant sur la gestion sont maintenant le facteur prépondérant des évolutions économiques. L'acquisition et l'utilisation de connaissances sont un problème économique majeur pour nos sociétés (Foray, 2000).

Aspect cognitif

Information et connaissances ne sont pratiquement utiles que si elles sont cognitivement assimilées par les individus et organisations, c'est-à-dire filtrées, apprises, comprises et intégrées. Cette assimilation se fait le plus souvent au sein de relations sociales qui permettent d'accéder à l'information pertinente et facilitent son intégration cognitive. D'où la grande importance des relations sociales en jeu (Nonaka et Takeuchi, 1995).

Aspect technique

L'évolution socio-économique se poursuit en symbiose avec le développement accéléré de techniques de communication variées (du téléphone à Internet) qui facilitent la diffusion d'information et de connaissances. Mais cette diffusion ne peut être efficace que si elle respecte et favorise les relations sociales les plus appropriées à l'acquisition et l'assimilation de connaissances (voir Drucker, 1970).

On remarque la cohérence de ces différents aspects, qui se renforcent mutuellement depuis plus d'un siècle, et cela de manière qui s'accélère depuis quelques décennies avec le développement de plus en plus poussé des réseaux électroniques. Dans ce contexte historique, notre but est de proposer des techniques informatiques pour favoriser la construction de relations sociales utiles à la diffusion efficace d'information et de connaissances.

Plus précisément, nous allons montrer comment construire des réseaux sociaux virtuels regroupant automatiquement les individus ayant des intérêts similaires et donc, susceptibles d'échanger utilement de l'information et de collaborer à la résolution de problèmes. En partant d'une réflexion sur les mécanismes spontanés de socialisation dans la vie sociale habituelle, nous allons proposer des mécanismes automatiques de socialisation se voulant à la fois techniquement efficaces et socialement utiles. Mais pour cela, avant de parler des réalisations techniques, nous devons maintenant détailler davantage la problématique, notamment du point de vue cognitif.

1.1.1 Traitement sociocognitif de l'information

L'étude de la manière dont nous traitons l'information a donné lieu à un ensemble considérable de travaux, notamment en psychologie cognitive, qui nous a permis de mieux comprendre les capacités individuelles cognitives chez l'humain comme la perception, la mémoire, le raisonnement, l'apprentissage, la résolution de problèmes, le langage, etc. Plus récemment, cependant, on parle de cognition située (Brown et al., 1989 ; Robbins et Aydede, 2009) et de cognition distribuée ou collective (Salomon, 1993; Harnad, 2005). La cognition collective renvoie au fait que l'interaction entre plusieurs individus cognitifs peut générer de nouvelles connaissances, que la connaissance du groupe est plus que la somme des connaissances de chaque individu. On parle ici d'un savoir cumulatif, collectif et collaboratif. La cognition située, quant à elle, est l'idée que l'on ne peut pas séparer la connaissance de son contexte. Les individus réfléchissent et apprennent en situation : en accomplissant différentes activités, en discutant avec les autres, en utilisant une langue particulière, en baignant dans une culture spécifique, etc. La connaissance est déterminée tant par l'individu qui apprend que par l'environnement (physique, biologique, social, etc.) dans lequel il évolue.

On constate cependant que l'environnement dans lequel nous évoluons est en grande partie social. Parce que nous fonctionnons la plupart du temps en société, une grande partie des informations que nous recevons proviennent de notre milieu social et nous sont aussi indispensables pour fonctionner au sein même de ce milieu. La majorité de nos activités journalières sont de nature sociale. On échange constamment de l'information les uns avec les autres pour accomplir des tâches sociales complexes : planifier et coordonner des activités, résoudre des problèmes, collaborer sur des projets, faire des choix, innover, etc. De ce point de vue, on peut donc comprendre le traitement de l'information (créer, collecter, transmettre, appliquer, échanger, interpréter, enseigner, etc.) *comme une activité collaborative située socialement.*

D'autre part, le développement rapide des technologies (les ordinateurs, l'Internet, les réseaux électroniques, le Web, les appareils mobiles...) soulève de nouveaux défis quant au traitement de l'information. On assiste en effet à la prolifération croissante de données

électroniques de toutes sortes sur Internet et le Web. Cette effervescence pose un problème évident quant à la gestion de cette information surtout lorsqu'on considère que celle-ci est une ressource cruciale pour le développement économique, social et technologique de nos sociétés modernes (Foray, 2000). Cependant, ces mêmes technologies proposent aussi de nouveaux lieux de socialisation et de collaboration dont on peut tirer parti pour mieux canaliser ces ressources informationnelles.

1.1.2 Des communautés traditionnelles aux cybersociétés

La montée des technologies a intensifié le phénomène de reconfiguration graduelle des groupes sociaux traditionnels denses et cohésifs vers des réseaux sociaux plus ouverts et moins stables. Comme (Memmi, 2009) le fait remarquer, ce phénomène avait été observé, déjà vers la fin du 19e siècle, par des sociologues allemands comme (Tönnies, 1963), (Weber, 1956) et (Simmel, 1989) qui ont posé la distinction entre les communautés traditionnelles (*Gemeinschaft*) et les sociétés modernes (*Gesellschaft*) dans leurs travaux d'analyse sur les sociétés modernes et de la vie urbaine. Les communautés traditionnelles sont décrites comme des communautés personnelles locales assez petites, assez denses, très cohésives et très stables enracinées dans les villages et le voisinage. À l'opposé, on décrit les sociétés modernes comme des regroupements sociaux dans lesquels les relations entre les individus sont plutôt impersonnelles, souvent temporaires ou transitoires et fréquemment formées dans un but pratique. La distinction entre les relations durables observées dans les groupes traditionnels et ces liens temporaires qu'on retrouve dans les réseaux sociaux modernes a aussi été discutée, entre autres, par (Granovetter, 1973). Celui-ci explique, par exemple, que les liens faibles et superficiels, bien que ne permettant pas la formation de communautés fortes et solidaires, sont cependant très utiles pour la circulation de l'information.

Le développement des technologies a grandement favorisé ce passage graduel des communautés traditionnelles vers les sociétés modernes. Dans la deuxième moitié du 20e siècle, la prolifération de moyens de transport comme les automobiles, les avions, les autobus, le train ou des moyens de communication comme le téléphone a favorisé les

relations de longue distance et donc, la délocalisation progressive des communautés. On observe alors le passage graduel des structures sociales traditionnelles vers des structures sociales intermédiaires, un peu moins denses et moins cohésives basées sur les réseaux. Ces réseaux sont formés de petites communautés locales fortement structurées (*clusters*), comme la famille et les collègues de travail, qui se connectent entre eux à l'aide de connexions de longue distance, plutôt que de ne former qu'une seule grande composante très cohésive à la manière des groupes traditionnels.

On observe aussi que la venue d'Internet et de sa progéniture (les courriers électroniques qui ont été suivis par les messageries instantanées, les espaces de clavardage, les blogues, les wikis, etc.), et l'expansion de l'utilisation des téléphones cellulaires et autres appareils mobiles ont amplifié ce phénomène de délocalisation de lieu en lieu vers des réseaux individualisés de personne à personne (Wellman, 2001 ; Castells, 2001). Par exemple, les gens sont connectés par des téléphones cellulaires qui ne sont plus assujettis à des endroits fixes et les courriels peuvent être consultés n'importe où moyennant une connexion Internet.

De telles structures sociales plus grandes, plus ouvertes et plus flexibles préservent les avantages des structures intermédiaires : l'accès à une grande variété d'informations et d'individus provenant de divers milieux sociaux ainsi que l'accès rapide à de nouveaux contacts par les super connecteurs (acteurs pivots, individus qui sont reliés à un grand nombre d'autres individus). Dans ces réseaux, les liens sont plus spécialisés, et fournissent aux individus, différents types de ressources, dans une multitude de milieux sociaux. Cette individualisation des connexions signifie que *l'acquisition de ressources dépend maintenant essentiellement de la capacité et de la motivation des individus à maintenir et trouver les « bonnes connexions »*. Chacun doit développer et constamment entretenir son propre réseau de relations.

Plus récemment, les réseaux électroniques, l'informatique sociale et l'avènement du Web 2.0 ont favorisé l'apparition et la prolifération de diverses communautés virtuelles. Qu'on les appelle wikis, laboratoires à distance, jeux en ligne, forums, communautés électroniques, réseaux sociaux virtuels, cybersociétés, groupes Web, sites de réseautage en ligne ou sites

d'achats en ligne, ces cybercommunautés sont des lieux de rencontre qui fournissent, à différents degrés, divers moyens de communication et d'échange d'informations entre les individus, peu importe la distance géographique qui les sépare. Ce sont des lieux cyberspatiaux où l'on socialise pour s'entraider, collaborer, apprendre, jouer, échanger, etc. En raison de leur nature virtuelle et intemporelle, (Korzeny, 1978 ; Wallace, 1999) font remarquer que les communautés en ligne se forment généralement autour d'intérêts communs ou d'affinités telles que l'éducation, la provenance ethnique, la langue, les passe-temps, les croyances, et que la proximité géographique ou l'attirance physique, par exemple, sont des facteurs de regroupement négligeables, contrairement aux communautés traditionnelles. Les développements les plus récents de l'informatique sociale sont les sites de réseautage en ligne comme Facebook, Twitter, LinkedIn, etc. Ceux-ci, contrairement aux communautés virtuelles plus traditionnelles, consistent en des réseaux de personnes qui croissent et changent très rapidement.

Que le traitement de l'information soit une activité collective située socialement n'est pas un fait nouveau. Cependant, le développement rapide des nouvelles technologies, offrant de plus en plus de possibilités de socialisation, dans divers milieux cybernétiques, qui reformulent notre environnement social vers le réseautage actif et qui sont responsables de l'apparition de nouvelles façons de collaborer, de coordonner nos activités et d'interagir en temps réel, affecte nécessairement nos manières de gérer l'information.

1.1.3 La collaboration en ligne : exploiter les affinités

L'intérêt des cybercommunautés réside aussi dans le fait qu'elles génèrent de l'information électronique récupérable et exploitable. Par exemple, dans plusieurs de ces groupes sociaux, les utilisateurs peuvent publier explicitement un profil personnel décrivant leurs intérêts, leur nationalité, leur lieu de résidence, leur formation professionnelle, etc. C'est le cas de Facebook (http://www.facebook.com) ou LinkedIn (http://www.linkedin.com), par exemple. Il devient aussi possible de déduire les intérêts des individus de manière implicite en collectant de l'information sur leurs activités en ligne : le genre de sites Web qu'ils fréquentent, les produits qu'ils achètent ou le type de recherche qu'ils effectuent dans Google.

Dans le cas particulier et de plus en plus populaire des communautés en réseaux, il est souvent possible d'extraire de l'information concernant la structure (et parfois la nature) des relations entre les individus faisant partie du réseau : qui connaît qui, qui collabore avec qui, etc. Bref, les participants de ces nouveaux lieux virtuels laissent inévitablement des traces électroniques récupérables qui deviennent une source d'information en soi.

Dès lors, on remarque l'apparition de nouvelles techniques de gestion de l'information, basées sur la collaboration en ligne, qui tentent de récupérer ces traces électroniques dans le but d'améliorer, de personnaliser la recherche d'informations : séparer l'information pertinente de celle qui ne l'est pas, selon les besoins particuliers d'un utilisateur. En effet, l'accès qu'offrent les communautés virtuelles à toutes sortes d'individus, provenant de divers milieux, a propulsé le concept de collaboration en ligne. On peut désormais non seulement utiliser les réseaux électroniques pour rechercher de l'information utile, mais aussi pour rechercher des individus pertinents : des experts, dans un domaine particulier, qui pourraient nous fournir l'information recherchée ou nous aiguiller vers celle-ci ou bien tout simplement des pairs qui partagent nos intérêts avec qui l'on pourrait échanger des connaissances utiles. On parle alors de navigation sociale, de recherche d'informations collaborative, de filtrage collaboratif, de systèmes de recommandation collaboratifs, etc. Ces techniques diffèrent quelque peu dans leur utilisation, mais leur objectif reste le même : tirer parti de la collaboration entre les individus pour élaguer l'information et ne conserver que celle qui est pertinente et utile dans le cadre de besoins spécifiques. L'hypothèse sous-jacente de telles méthodes est que les individus qui ont des profils similaires peuvent vraisemblablement s'entraider en partageant leurs expériences et leurs connaissances.

Les communautés virtuelles deviennent ainsi des lieux informationnels pouvant être exploités en associant les utilisateurs ayant des profils similaires, en créant des liens d'affinité implicites entre les individus qui se ressemblent. Considérons, par exemple, la communauté des consommateurs en ligne sur un site comme Amazon (http://www.amazon.com). En cataloguant les utilisateurs selon ce qu'ils achètent, on peut les comparer et déduire les profils de consommation qui se ressemblent. Avec cette information, on peut alors proposer à un utilisateur, des produits ayant été achetés par des consommateurs similaires, en supposant que

ces produits seront intéressants pour cet utilisateur. C'est l'idée principale des systèmes de recommandations collaboratifs. Dans cet exemple, on construit les profils des utilisateurs selon les produits qu'ils achètent, mais dans d'autres contextes on utilisera d'autres types d'informations disponibles pour décrire les utilisateurs et déduire ceux qui présentent des affinités. Nous discuterons de ces méthodes au chapitre 2 (voir sect. 2.7).

Notons que nous utilisons le terme "liens implicites", ici, pour parler des relations qui ne sont pas formellement déclarées, mais simplement tacites, déduites par affinité (par similarité des profils). La section suivante porte plus particulièrement sur les communautés en réseau (les réseaux sociaux) dans lesquelles nous parlerons plutôt de "liens explicites".

1.1.4 Tirer profit des liens sociaux

Nous utilisons le terme "réseau social" pour désigner plus spécifiquement les communautés dans lesquelles les membres sont considérés comme des entités sociales entretenant des relations avec d'autres membres du réseau. Peu importe l'intensité ou la nature de ces relations (professionnelles, de collaboration, d'amitié, etc.), on parle ici de liens explicites, formellement énoncés (et non déduits par calculs). Dans ce type de communautés, chaque membre possède son propre sous-réseau de contacts personnels et la réunion de tous ces (plus ou moins grands) réseaux locaux forme la structure globale du réseau entier ; l'information structurelle, relationnelle, se trouve au niveau de chacun des membres.

Lorsqu'il s'agit d'un réseau social virtuel, comme LinkedIn (http://www.linkedin.com), par exemple, il est possible de conserver l'information concernant les relations entre tous les individus et d'avoir, à tout moment, une vue globale du réseau social. Ceci nécessite cependant un lieu centralisé pour le stockage de cette information, un ou plusieurs serveurs, par exemple, qui doivent constamment être entretenus et mis à jour.

Par contre, au sein de réseaux complètement décentralisés (sans gestion centrale de l'information), comme c'est le cas du Web ou de nos réseaux sociaux réels, par exemple, le réseau global est inconnu de tous. Dans les réseaux sociaux, par exemple, on connaît nos

relations et peut-être certaines des relations de nos relations, mais pas beaucoup plus loin. Comme la seule façon de découvrir le Web est de parcourir les pages, d'hyperlien en hyperlien, la seule manière de découvrir un réseau social est de parcourir les individus, en suivant les liens, de personne en personne. Dans le cas du Web, on appelle cette technique *crawling*, dans le cas des réseaux sociaux, on parle de socialisation.

Dans ce contexte décentralisé, on réalise que la connaissance des propriétés structurales typiques des réseaux est un atout précieux pour rendre ces parcours à l'aveugle plus efficaces en regard de la recherche d'informations spécifiques ou d'individus intéressants. On aimerait savoir si nos réseaux sociaux possèdent des propriétés structurales récurrentes sur lesquelles on pourrait se baser afin de concevoir, par exemple, de meilleurs algorithmes de parcours efficaces du réseau, lors de la recherche d'informations ou d'individus pertinents. De même que l'étude structurale du graphe du Web a donné lieu au fameux algorithme PageRank de Google (Page et al., 1998 ; Brin et Page, 1998), on s'intéresse aux propriétés structurales des réseaux sociaux dans l'espoir d'améliorer les systèmes de diffusion et de recherche d'informations. On retrouve ainsi, dans la littérature, un ensemble considérable de travaux portant sur l'analyse structurale des réseaux sociaux, sur la diffusion de l'information au sein de structures typiques, sur les mécanismes d'évolution des réseaux, sur la manière de localiser des individus particuliers au sein d'un réseau (navigation) en tirant parti des propriétés structurales de celui-ci, etc. Nous aborderons ces travaux plus en détail au chapitre suivant.

Les réseaux sociaux peuvent être vus comme un type particulier de réseaux d'informations lorsqu'on considère les individus qui les composent comme des sources de connaissances. Cependant, l'information acquise par échanges sociaux a l'avantage non négligeable d'avoir déjà été traitée cognitivement, d'avoir été intériorisée, élaguée, contextualisée, élaborée, etc. Dans un contexte où l'information surabonde, de pouvoir profiter de ce prétraitement de l'information, par l'intermédiaire de nos contacts sociaux, est un atout majeur pour acquérir plus rapidement les connaissances qui nous sont vraiment utiles. L'apparition et la prolifération des communautés virtuelles ont favorisé l'accès aux individus (et à leurs connaissances) par échanges sociaux, et ce, dans divers milieux.

Toutefois, l'utilisateur de telles communautés doit entretenir, s'il y a lieu, ses différents profils sur les différentes communautés auxquelles il appartient, il doit visiter ces communautés pour demeurer au fait de ce qui se passe et surtout, il doit socialiser activement pour entretenir et enrichir son réseau de contacts personnels. De même qu'il faut du temps pour perpétuer nos relations dans nos réseaux sociaux réels, il faut du temps et un certain talent pour trouver et maintenir les "bonnes connexions" dans les réseaux sociaux virtuels afin d'en tirer vraiment parti.

L'idée initiale de ce projet de recherche a été motivée par cet état de fait. Oui, nous avons maintenant de multiples possibilités de socialisation, et celle-ci est en effet un moyen efficace, si l'on s'en donne la peine, pour découvrir du contenu pertinent, pour parfaire son réseau de contacts par la rencontre de nouveaux individus intéressants et pour entretenir les relations qui nous sont utiles, mais la socialisation active prend du temps et de l'énergie. Serait-il possible, alors, d'informatiser les mécanismes de socialisation et concevoir un système qui socialiserait pour nous ? En effet, nos processus de socialisation dans les réseaux sociaux réels semblent assez efficaces quant à la découverte de connaissances/contacts utiles en contexte non centralisé. Nous pensons donc que ces mécanismes pourraient aussi fonctionner au sein de réseaux sociaux virtuels décentralisés.

Sur la base d'observations des mécanismes de socialisation dans nos réseaux sociaux réels, nous voulons donc concevoir un modèle de socialisation automatique capable de trouver et maintenir les "bonnes connexions" c'est-à-dire, capable de gérer et maintenir les communautés d'intérêt au sein d'un réseau social non centralisé ainsi qu'en constante évolution.

La section suivante décrit en détail les problématiques abordées dans cette thèse ainsi que nos objectifs de recherche quant à la résolution de ces problèmes.

1.2 Problématique et objectifs de recherche

Les informations que nous recevons proviennent souvent de nos connaissances et contacts. Parfois, lorsqu'on s'intéresse à un nouveau sujet par exemple ou que l'on commence à explorer un nouveau domaine de recherche, nos idées ne sont pas toujours claires sur l'information que l'on cherche, faute d'une bonne connaissance du domaine. On fait souvent appel à nos contacts sociaux, que l'on sait plus experts que nous sur le sujet, pour obtenir, par exemple, quelques pointeurs de recherche ou pour nous diriger vers d'autres personnes susceptibles de nous aider.

D'autre part, on socialise aussi sans être en quête d'informations spécifiques. De discuter, d'échanger tout simplement avec des individus qui partagent nos intérêts permet aussi la découverte spontanée de nouvelles connaissances utiles et pertinentes qui ouvre nos horizons vers d'autres possibilités, parfois insoupçonnées. Ces deux types d'acquisition de connaissances (ciblée et non ciblée) par socialisation se renforcent mutuellement : la recherche explicite de connaissances nous amène à socialiser, puis en socialisant on découvre d'autres avenues, qu'on cherche ensuite à approfondir en socialisant de nouveau, etc.

Le fait d'obtenir de l'information par l'intermédiaire d'une personne humaine est très avantageux au sens où l'information que l'on reçoit a préalablement été traitée cognitivement (choisie, intégrée, assimilée, triée, mise en contexte...). Les individus agissent comme des filtres de l'information lorsqu'ils partagent leurs connaissances. En ce sens, la socialisation est un processus collaboratif par lequel on réutilise le travail personnel de recherche, d'acquisition et d'élagage d'informations des uns pour augmenter les connaissances des autres. En effet, une information est utile lorsqu'elle répond à un besoin réel et d'autant plus utile lorsqu'elle est intégrée, comprise et contextualisée. C'est dans cette optique que (Nonaka et Takeuchi, 1995) font remarquer que la socialisation est effectivement un mécanisme efficace dans l'acquisition de connaissances.

Les individus sont donc des sources d'informations avantageuses et la socialisation est le moyen de les découvrir. Par exemple, lorsqu'on participe à des événements sociaux comme

des colloques ou des congrès, on fait de nouvelles rencontres. On discute d'abord avec une personne possiblement rencontrée au hasard, puis une autre et une autre pour tranquillement se grouper naturellement et échanger plus longuement avec les gens qui partagent nos intérêts. Lorsqu'on socialise de cette manière, on ne cherche pas nécessairement une information particulière, mais on se dirige naturellement vers les gens avec qui l'on a des affinités et c'est en discutant avec eux qu'on apprend souvent beaucoup de choses utiles et pertinentes. La socialisation favorise ainsi les *regroupements* d'individus qui ont des affinités en commun.

Il est étonnant de voir comment cette navigation à l'aveugle nous permet tout de même assez efficacement de cheminer, dans le réseau des individus présents, vers ceux qui partagent nos intérêts (s'ils existent). On passe d'une personne à une autre, sans connaître la structure globale des relations qui unissent les individus présents, sans chemin tracé d'avance et ça fonctionne. Nous pensons que cette réalité s'explique, du moins en partie, parce que la plupart du temps, les gens qui socialisent aiment partager leurs "connaissances". Ici, l'emploi du terme "connaissances" n'est pas un hasard. En effet, on peut parler de connaissances en terme de savoir (ce que l'on connaît), mais aussi en terme de contacts (les individus que l'on connaît). En socialisant, on s'échange non seulement du "savoir", mais aussi des "contacts" c.-à-d. qu'on se guide mutuellement les uns vers les autres. Ainsi, tout au long de ce *parcours social* dans le réseau des gens présents, on a très probablement rencontré des gens qui n'avaient pas tout à fait les mêmes intérêts que nous, mais qui nous ont recommandés à des individus susceptibles de nous intéresser. On fait de même pour les autres. Bref, lorsque l'on socialise, on navigue dans le réseau de personne en personne et l'on tend à se regrouper avec des individus qui partagent nos intérêts, par le biais de mécanismes de socialisation tels que les *rencontres* aléatoires et les *recommandations*.

De plus, lorsqu'on rencontre un individu intéressant (par hasard ou par recommandation), on risque aussi de rencontrer, par ricochet, les membres de sa communauté. La création d'un nouveau lien entre deux individus a pour effet de relier ensemble deux communautés dont certains membres sont susceptibles de partager des intérêts et d'ainsi bénéficier de cette seule rencontre. Dans cette optique, la socialisation est un processus collaboratif qui permet non

seulement de parfaire nos connaissances en réutilisant la connaissance des autres, mais aussi, d'élargir notre réseau personnel de contacts tout en augmentant aussi celui des autres.

En résumé, lorsqu'on socialise :

1. On fait des parcours de socialisation dans le réseau des gens présents.
2. On fait de nouvelles rencontres.
3. On se recommande les uns aux autres.
4. On se regroupe naturellement avec des individus partageant nos intérêts.

En nous inspirant des études sur la structure des réseaux sociaux ainsi que des travaux portant sur les systèmes collaboratifs, que nous aborderons au chapitre 2, *notre thèse est que l'implémentation de mécanismes de socialisation dans un réseau social virtuel complètement décentralisé peut favoriser les regroupements automatiques d'individus similaires ainsi que le maintien de ces groupes, au sein d'un réseau en constante évolution.*

Notre objectif principal est donc de concevoir un modèle de socialisation automatique, qui incorpore les mécanismes de rencontres aléatoires et de recommandation, par l'intermédiaire de parcours de socialisation dans le réseau. Nous attendons de ce modèle (1) qu'il puisse localiser automatiquement les individus semblables au sein du réseau, (2) qu'il les connecte entre eux sous forme de communautés d'intérêt et (3) qu'il maintienne et renouvelle ces communautés, dans la mesure du possible, tout au long de l'évolution du réseau.

Les problèmes que nous abordons ici doivent être compris dans un contexte complètement décentralisé, comme nos réseaux sociaux réels, où il n'existe aucun lieu de coordination centrale ou entrepôt de données global sur les individus et la structure du réseau complet. La seule information disponible se trouve au niveau des individus et se résume aux attributs propres à l'individu comme ses intérêts, par exemple ainsi que la liste de ses contacts personnels. Les relations sociales sont donc un atout majeur pour accéder à l'information utile et pertinente. Mais comment trouver ces "bonnes" connexions dans un réseau social dont on ne connaît qu'une infime partie et qui évolue constamment ? Plus généralement, comment

construire des réseaux dont les mécanismes d'évolution intrinsèques favorisent les associations entre individus de profils similaires ?

Pour concevoir un modèle de socialisation automatique qui rassemble les individus en communautés d'intrérêt, les problèmes rencontrés ou les défis à relever, selon nous, peuvent se classer en trois catégories : l'évolution (1) des réseaux (2) des intérêts des individus et (3) des catégories sémantiques pour décrire l'information qui circule dans les réseaux. La problématique principale abordée dans cette thèse concerne l'évolution des réseaux, mais nous traitons partiellement l'évolution des intérêts et montrons comment notre modèle pourrait résoudre, dans une certaine mesure, l'évolution des catégories sémantiques.

1.2.1 Défi de l'évolution des réseaux en contexte décentralisé

1.2.1.1 Localisation d'individus dans le réseau

Pour pouvoir regrouper ensemble les individus similaires, il faut d'abord les localiser dans le réseau, en faisant des parcours de socialisation : en explorant le réseau d'individu en individu, en suivant les liens. Au cours de ce parcours, chaque fois qu'on visite un individu, on compare, en quelque sorte, nos intérêts et s'ils sont similaires, on vient de rencontrer quelqu'un d'intéressant. Cependant, nos réseaux sociaux, en raison de leur nature dynamique, ne facilitent pas cette tâche de localisation d'individus de profils similaires. Le va-et-vient incessant causé par l'arrivée et le départ d'individus, par la création de nouveaux liens et la cessation d'anciennes relations modifient constamment le patron des connexions entre les individus : un nouvel arrivant crée de nouveaux liens, le départ d'un utilisateur a pour effet de couper tous les liens qui l'unissaient aux autres dans le réseau, etc. Ainsi, les chemins de lien en lien dans le réseau, qui connectent les membres les uns aux autres, sont aussi en perpétuel remaniement. Par conséquent, l'exploration du réseau est d'autant plus difficile. De plus, dans un grand réseau, on ne peut évidemment pas se permettre de visiter tous les individus présents chaque fois qu'on socialise. Le problème, ici, est donc de trouver des chemins dans le réseau, des parcours de socialisation, qui demeurent assez courts par rapport aux nombres d'individus se trouvant effectivement dans le réseau et qui offrent tout de même une bonne chance de rencontrer des individus qui ont des intérêts similaires.

En effet, nous verrons, au chapitre 2, que dans nos réseaux sociaux réels, de tels chemins existent et qu'en général, les individus sont assez bons pour les trouver (Milgram, 1967 ; Dodds et al., 2003). Nous verrons aussi que les propriétés structurales des réseaux sociaux, qui définissent le patron des relations entre les individus présents dans le réseau, jouent un rôle important dans notre capacité à trouver ces chemins de courte distance, et ce, à l'aide d'informations locales uniquement.

Dans cette optique, la solution que nous proposons pour trouver ces chemins de courte distance lors de la recherche d'individus similaires dans un réseau en continuel remaniement est d'imposer aux réseaux générés par notre modèle des propriétés structurales désirables, c'est-à-dire qui favorisent la navigation efficace.

Plus précisément, nous voulons que la structure du réseau fournisse effectivement de courts chemins entre les individus, mais qu'elle ne comporte pas trop de chemins possibles entre les individus. On veut minimiser la redondance des chemins. En effet, plus le nombre de chemins possibles est grand, plus il sera difficile de découvrir celui qui est court. En contexte décentralisé, on navigue dans le réseau, d'individu en individu, et chaque fois qu'on visite un individu, on doit choisir, parmi ses voisins immédiats, le prochain individu à visiter. C'est de cette manière qu'on découvre les chemins, en suivant les liens, un lien à la fois. Si les individus dans le réseau ont tendance à avoir beaucoup de voisins (beaucoup de liens), ils offrent ainsi plusieurs possibilités de chemins différents dans le réseau : chaque fois qu'on visite un individu et qu'on doit choisir le prochain lien à parcourir, on aura ainsi plus de chances de choisir un mauvais chemin. En terme de théorie des graphes, on désire donc une structure qui présente à la fois une petite distance géodésique moyenne entre les individus (présence de courts chemins) et une faible densité (pas trop de liens dans le réseau).

Ainsi, notre objectif est de concevoir des algorithmes d'évolution du réseau, basés sur les mécanismes de socialisation, mais qui de plus contrôlent, dans une certaine mesure, tous les changements structuraux pouvant survenir au cours de l'évolution du réseau : le départ et l'arrivée d'individu dans le réseau, la création de nouveaux liens et la cessation d'anciennes relations. La difficulté, ici, est de maintenir ces propriétés structurales dans le réseau, et ce,

malgré sa constante évolution, de telle sorte que l'on puisse toujours s'y fier et les exploiter lors des parcours de socialisation dans le réseau. Ces algorithmes d'évolution du réseau sont expliqués en détail au chapitre 3 et constituent le cœur de cette thèse.

1.2.1.2 Formation, maintien et renouvellement des communautés d'intérêt

Dans ce contexte évolutif et décentralisé, pour former des communautés d'intérêt, la difficulté est aussi de trouver un mécanisme pour organiser dynamiquement les pairs entre eux, de les regrouper (et de maintenir ces regroupements) de manière à ce que les individus similaires demeurent proches les uns des autres, qu'ils soient accessibles les uns aux autres, et ce, malgré la constante restructuration des réseaux dans lesquels ils évoluent et sans la connaissance de la structure globale du réseau.

La constante évolution des réseaux suppose donc le remaniement continuel des communautés formées au sein de ces réseaux. Le maintien de ces communautés devient donc problématique. Supposons, par exemple, une communauté d'intérêt particulière, au sein d'un réseau, composée des membres A, B, C, D, E, F et G, telle qu'illustrée à la figure 1.1 (a). La figure 1.1 (b) montre que si le membre D s'en va, la communauté initiale se scinde en deux composantes qui ne peuvent plus communiquer l'une avec l'autre. Il faudrait, par exemple, que deux des voisins immédiats de D créent un lien pour maintenir la communauté, comme illustré à la figure 1.1 (c). Nos algorithmes d'évolution du réseau devront donc, en plus de maintenir les propriétés structurales pour une navigation efficace, proposer aussi des mécanismes de relais pour éviter de telles situations, et ce, tout en s'assurant que les propriétés structurales imposées pour favoriser la navigation sont respectées.

Ensuite, non seulement la structure interne de nos communautés déjà formées risque de changer, comme on vient de l'expliquer, mais de plus, de nouveaux individus arrivent constamment dans le réseau et ceux-ci deviennent des membres potentiels pour ces communautés. Cependant, il n'est pas sûr qu'un individu qui arrive dans le réseau trouve tout de suite une communauté d'appartenance (et l'on suppose ici qu'il en existe une). S'il ne trouve pas, il forme donc une nouvelle communauté qui ne contient pour le moment qu'un seul individu. Puis, au fil du temps, d'autres individus de même profil qui arrivent dans le

réseau peuvent finir par se greffer à cette nouvelle communauté. Cependant, la communauté existante qui aurait pu être trouvée, au départ, existe encore. Ce phénomène a donc pour effet de produire différentes communautés éparpillées dans le réseau, qui partagent les mêmes intérêts, mais qui ne se connaissent pas encore. Le problème ici est donc de réunir ces communautés d'intérêts similaires dispersées dans le réseau.

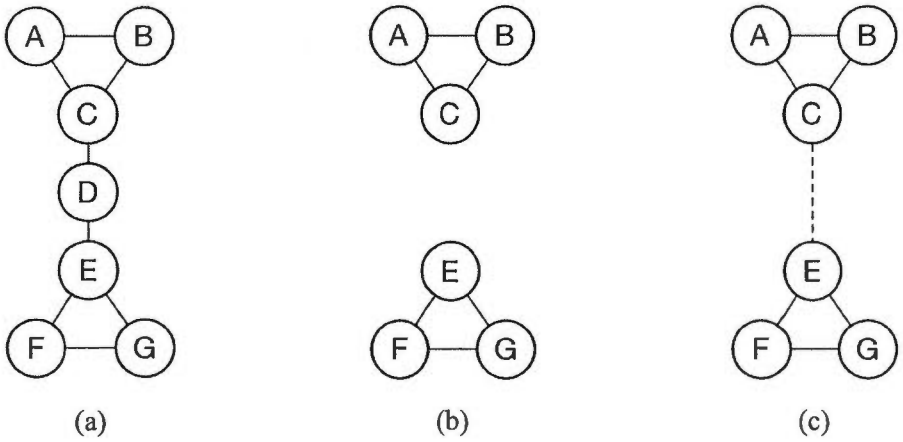


Figure 1.1 Problème de séparation des communautés lorsqu'un individu quitte le réseau.
En considérant le réseau de la figure (a) comme une communauté d'intérêt, la figure (b) montre que la déconnexion du nœud D aura pour effet de scinder la communauté initiale en deux communautés distinctes. La figure (c) montre la nécessité de prévoir un mécanisme de relais pour éviter cette situation.

On a vu que c'est par les parcours de socialisation qu'un individu peut rencontrer d'autres individus de profil similaire. Lorsque deux individus se trouvent, on veut donc les réunir. Cependant, chacun des deux individus appartient peut-être déjà à une communauté d'intérêt. Si tel est le cas, on sait qu'on peut aussi fusionner les deux communautés d'appartenance de chacun d'eux qui sont nécessairement d'intérêts similaires. La socialisation d'un seul individu a donc pour effet de réunir les communautés d'intérêt similaires qui sont dispersées dans le réseau, et profite à tous les membres des deux communautés ainsi fusionnées en une seule. Le procédé est le même lorsqu'au cours d'un parcours de socialisation, on effectue une recommandation. On opère encore une fusion des communautés d'appartenance des deux individus recommandés l'un à l'autre. Ce mécanisme de réunion des communautés s'apparente (sans être identique, cependant) à un phénomène qu'on observe dans nos réseaux sociaux réels qui est la tendance à former des relations avec les amis de nos amis. Au chapitre 2 (voir

sect. 2.3), nous présenterons ce processus de transitivité, aussi connu sous le nom de fermeture des triangles dans la littérature. Notre objectif est donc d'incorporer, dans nos algorithmes d'évolution du réseau, ce mécanisme de fusion des communautés lors des rencontres et des recommandations effectuées dans un parcours de socialisation, et ce, tout en maintenant constamment les propriétés structurales imposées pour la navigabilité du réseau.

Finalement, ce n'est pas tout de rapprocher les individus d'intérêts similaires, le problème ici, est de les organiser, de les connecter de manière à ce qu'ils puissent se trouver, se visiter facilement à l'intérieur d'une même communauté. Il faut que la structure des regroupements induise naturellement un chemin optimal pour qu'un individu puisse parcourir tous les membres de sa communauté ; on ne veut pas refaire de parcours de socialisation chaque fois que l'on désire visiter un membre de notre propre communauté. Évidemment, on crée des communautés pour favoriser l'échange entre les individus qui se ressemblent, ceux-ci doivent donc pouvoir se côtoyer sans difficulté.

Comme on le sait, un individu connaît la liste de ses voisins immédiats. Cependant, certains voisins peuvent ne pas être de même profil que lui : si tous les individus ne possédaient que des voisins de même profil, les communautés seraient toutes séparées les unes des autres en composantes distinctes dans le réseau. Dans nos réseaux sociaux réels, on possède beaucoup de contacts qui ne partagent effectivement pas nos intérêts, ce peut être des membres de notre famille, des collègues de travail, etc. Cependant, cette variété de contacts constitue une richesse en soi et ce sont ceux-là qu'on "partage" lorsqu'on agit en tant qu'intermédiaire entre deux individus qui, on le pense, auraient intérêt à se connaître (on les recommande l'un à l'autre).

Au point de vue structural, on comprend aussi qu'un tel partitionnement des communautés n'est pas souhaitable puisqu'un individu présent dans une composante ne pourrait jamais être découvert par un autre individu qui socialise dans une autre composante. En effet, la connexité du réseau fait partie des propriétés structurales de base qu'on impose lors de l'évolution des réseaux générés par notre modèle.

Alors, étant donné l'information locale disponible dans ces conditions, nos algorithmes d'évolution de réseaux doivent non seulement former, maintenir et renouveler les communautés, mais aussi les former de manière à ce qu'un individu puisse visiter tous les membres de sa communauté, de manière optimale ; en ne parcourant pas plus de liens qu'il n'y a d'individus dans sa communauté.

Par exemple, la figure 1.2 montre un réseau qui contient trois communautés différentes (représentées par trois couleurs différentes). On remarque que l'individu A possède un voisin immédiat blanc, un voisin immédiat gris et un voisin immédiat noir. En considérant seulement la structure (en oubliant les couleurs), on ne peut pas savoir si A appartient à la communauté noire, à la communauté grise ou à la communauté blanche, ou s'il fait même partie d'une de ces communautés. La seule façon de le savoir serait pour A de refaire une comparaison entre son profil d'intérêt et celui de chacun de ses voisins immédiats, pour déterminer lesquels sont de même profil que lui et appartiennent donc à sa communauté. Cependant, l'objectif de réunir les individus en communautés est justement de ne plus avoir à refaire ces comparaisons. On veut qu'une fois rassemblés, les individus puissent se visiter facilement.

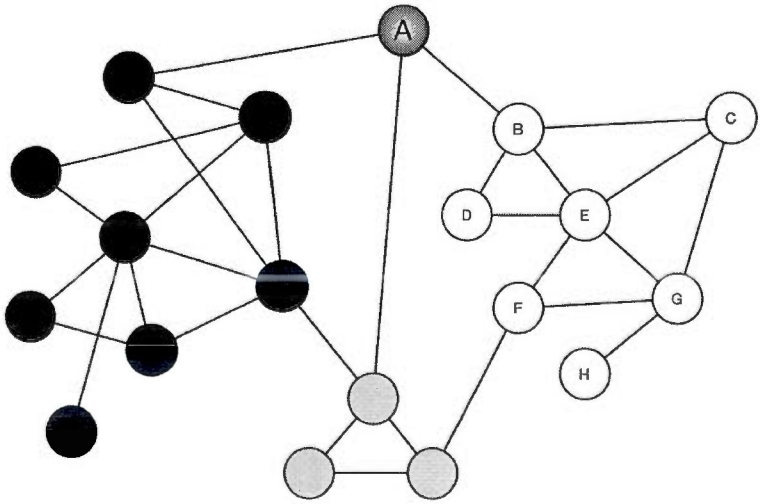


Figure 1.2 Problème d'identification de la communauté d'appartenance d'un individu.
Dans le graphe illustré, on remarque que lorsqu'on ne connaît pas le profil (la couleur) des différentes communautés, on ne peut pas déterminer la communauté d'appartenance du nœud A.

De plus, lors d'une visite communautaire, supposons que A identifie son voisin immédiat similaire (B) dans la communauté blanche et qu'il désire visiter tous les membres de sa communauté. Disons qu'il choisit C comme prochain individu à visiter, puis qu'il choisit G, et puis H. Arrivé sur H, il n'a plus aucun choix possible et doit donc nécessairement revenir sur ses pas pour continuer son parcours. Dans ce cas, il pourrait revenir sur G ou sur C pour reprendre son parcours en choisissant un autre voisin qui n'a pas encore été visité. C'est ce retour arrière qu'on aimerait minimiser. Cette manière de faire suppose aussi que l'individu qui visite sa communauté conserve en mémoire tous les chemins vers tous les individus qu'il a visités, mais dont il n'a pas encore visité tous les voisins.

Encore une fois, nous voulons que les membres d'une communauté soient facilement accessibles les uns aux autres. Notre objectif est donc de trouver une façon de structurer les communautés au sein du réseau, en accord avec les propriétés structurales imposées pour la navigabilité, de telle sorte qu'elles induisent des parcours de visites communautaires simples et efficaces, sans avoir à comparer les profils pour pouvoir identifier les membres de sa propre communauté. Au chapitre 3, nous expliquerons plus en détail le type de structure des communautés que nous proposons pour aborder ce problème.

La difficulté de former, maintenir et renouveler automatiquement les communautés d'intérêts au sein d'un réseau décentralisé et en constante évolution constitue la problématique principale de cette thèse.

D'un point de vue informatique, nous faisons donc l'hypothèse que les mécanismes de socialisation, comme les rencontres et les recommandations qui sont effectuées au cours des parcours de socialisation, favorisent le regroupement des individus ayant des intérêts en commun. De plus, nous pensons qu'en contrôlant l'aspect structural du réseau, ces mécanismes de socialisation peuvent être implémentés de telle sorte qu'ils forment et maintiennent des réseaux qui favorisent la navigabilité, et donc la localisation d'individus similaires. La découverte d'individus similaires favorisant à son tour le rassemblement des communautés d'intérêt similaires qui étaient éparpillées dans le réseau, assure ainsi le renouvellement continuel des communautés dans un réseau en constante évolution.

Nous proposons donc de concevoir un modèle dynamique de réseaux sociaux dont l'évolution est basée principalement sur les mécanismes de socialisation et qui génère des réseaux possédant des caractéristiques structurales favorisant à la fois la navigabilité dans le réseau, et la formation de regroupements d'individus d'intérêts similaires pouvant aisément communiquer entre eux.

En résumé, nos algorithmes d'évolution du réseau devront gérer

- l'arrivée et le départ d'individus,
- la création et la cessation de relations,
- les parcours de socialisation avec rencontres et recommandations (qui sous-entend ici la réunion d'individus et de communautés similaires)

et devront produire des réseaux

- toujours connexes,
- qui permettent une navigation efficace lors des parcours de socialisation,
- qui forment, maintiennent et renouvellent les communautés d'intérêt,
- qui produisent des structures de communautés non équivoques.

Nous verrons aussi au chapitre 2, que les réseaux sociaux réels ne se forment pas de manière complètement aléatoire, mais qu'ils montrent plutôt des propriétés structurales particulières, récurrentes, qui suggèrent ainsi des processus d'évolution intrinsèques des réseaux.

Manifestement, notre choix de contrôler la structure du réseau dans le modèle que nous proposons est motivé par des raisons fonctionnelles (la navigabilité) cependant, la solution proposée est en partie cognitive. Dans cette optique, on peut considérer les mécanismes de socialisation en tant que processus d'évolution des réseaux sociaux.

D'un point de vue sociocognitif, nous faisons donc l'hypothèse que les mécanismes de socialisation peuvent servir à expliquer, du moins en partie, la manière dont se forment les réseaux, comment ceux-ci en viennent à posséder les propriétés structurales particulières

qu'on observe généralement dans les réseaux sociaux réels. Nous analyserons donc la structure des réseaux générés par notre modèle de socialisation pour comparer leurs propriétés structurales à celles qu'on observe dans nos réseaux sociaux habituels.

Il est possible (mais nous ne tentons pas de le prouver) que nos réseaux sociaux se forment et se renouvellent dans le temps de manière à ce qu'ils nous soient utiles pour vivre en société. Nos relations sociales nous procurent, en effet, un certain capital social. L'idée principale derrière la notion de capital social est que l'investissement dans les relations sociales produit des bénéfices (Borgatti et Jones, 1998 ; Lin, 1999). Les individus socialisent, ils interagissent entre eux et se font de nouveaux contacts parce que ça leur est profitable. Par exemple, les liens sociaux permettent aux individus d'avoir accès à l'information et aux opportunités autrement inaccessibles, de se bâtir une bonne réputation qui favorise leur crédibilité sociale, de renforcer leur identité et leur sentiment d'appartenance par la reconnaissance des pairs, etc. Le capital social est alors cette valeur ajoutée que nous procurent nos relations sociales.

En termes de réseaux sociaux, des sociologues comme (Burt, 1992), par exemple, ont suggéré que certaines configurations des relations d'un acteur au sein d'un réseau sont plus profitables que d'autres ; que la position des acteurs dans un réseau influence directement leur source d'opportunités ainsi que leur lot de contraintes. Dans cet esprit, nous faisons aussi l'hypothèse que les mécanismes de socialisation favorisent, dans une certaine mesure, la création de capital social. Nous verrons, au chapitre 2 (voir sect. 2.3), des indicateurs de capital social que nous utiliserons ensuite pour mesurer le capital social dans les réseaux générés par simulation de notre modèle de socialisation.

La section suivante présente une autre problématique connexe, mais que nous n'abordons que partiellement dans le cadre de cette thèse.

1.2.2 Problème de l'évolution des profils d'intérêts

Pour construire des réseaux qui forment et perpétuent automatiquement les communautés d'intérêt, il faut pouvoir déterminer le niveau de similarité entre les individus. Pour ce faire, il

est nécessaire d'obtenir, d'extraire, de déduire, d'une manière ou d'une autre, un profil d'intérêts pour chaque membre de la communauté. Ce problème a déjà été abordé dans la littérature, particulièrement dans le domaine des systèmes sociaux et collaboratifs. Il existe donc déjà diverses méthodes de profilage des individus, de façon explicite ou implicite, dont nous discuterons au chapitre 2 (voir sect. 2.7). Pour simuler notre modèle de socialisation, nous utiliserons une méthode existante d'extraction et de comparaison de profil.

Cependant, il faut noter ici que le modèle proposé est générique au sens où l'on peut fournir au système la procédure d'extraction et de comparaison de profils que l'on désire pourvu que la méthode de comparaison retourne une valeur numérique entre 0 et 1. La valeur 0 doit indiquer une similarité nulle, la valeur 1, une similarité parfaite et toutes les autres valeurs entre 0 et 1, une similarité plus ou moins grande. Ainsi, advenant une implémentation de ce modèle de socialisation dans une application distribuée, on a la possibilité de concevoir des méthodes d'extraction et de comparaison de profils plus adéquates, qui tirent possiblement profit du contexte de l'application. Par exemple, on pourrait implémenter ce modèle de réseau social virtuel décentralisé dans un réseau universitaire, pour former automatiquement des communautés d'intérêts de recherche (et ainsi pouvoir connaître les chercheurs qui travaillent sur des sujets de recherche similaires). On pourrait alors demander à chaque participant de fournir, par exemple, l'URL de sa page Web dans son université d'attache. Le système pourrait ensuite extraire un profil d'intérêts à partir de la page Web de chaque chercheur et utiliser ce profil pour la socialisation.

La généricité du modèle permet aussi de former des communautés qui ne sont pas nécessairement basées sur les intérêts. Par exemple, on pourrait vouloir rassembler les individus selon d'autres critères comme leur âge, leur niveau d'éducation, leur lieu de résidence, leur nationalité, leur langue natale, etc. Il suffit de fournir une méthode de comparaison qui calcule le niveau de similarité, peu importe les critères de comparaison.

L'aspect problématique que nous abordons dans cette thèse concerne surtout l'évolution des intérêts des individus au cours du temps : les intérêts individuels sont susceptibles de changer au cours du temps, ils évoluent. Ainsi, il se peut qu'à un moment donné, un individu se

retrouve associé à des pairs avec qui il n'a plus d'affinités. Il se retrouve donc dans une communauté qui ne répond plus à ses besoins et la communauté, qui abrite maintenant un étranger, perd de sa cohésion.

Pour aborder ce problème, notre objectif est d'incorporer, dans notre modèle de socialisation, un mécanisme supplémentaire de mise à jour des relations des individus qui sera géré par nos algorithmes d'évolution de réseau. Si, à un moment ou à un autre, un individu n'est plus relié à des pairs similaires, la règle de mise à jour s'exécute automatiquement. Nous verrons cet algorithme plus en détail au chapitre 3. De plus, étant donné la généricité de notre modèle, il est donc possible de fournir une méthode de compilation des profils implicite qui s'occupe automatiquement de mettre à jour le profil des acteurs, lorsqu'il y a lieu (nous verrons certaines de ces méthodes au chapitre 2). Dans ce cas, et la mise à jour des profils, et la mise à jour des relations (provoquée par la mise à jour drastique d'un profil) pourront se faire de manière complètement automatique.

Aussi, nous voudrions préciser que dans le cadre de cette thèse, nous avons choisi de considérer des profils centrés sur un seul domaine d'intérêts bien qu'en réalité, les individus en possèdent souvent plusieurs. Nous reviendrons sur ce point à la fin du chapitre 3 (voir sect. 3.5).

Une autre problématique rattachée à notre projet est l'évolution du vocabulaire de description de l'information qui peut circuler dans nos réseaux. Nous n'abordons pas ce problème dans le cadre de cette thèse, mais nous pensons que notre modèle de socialisation pourrait éventuellement résoudre, du moins en partie, le problème exposé dans la section qui suit.

1.2.3 Problème de l'évolution des catégories sémantiques

Dans les communautés ouvertes qui ne cessent d'évoluer, les sujets d'actualité changent, les modes changent, les codes, les pratiques, les discours changent. Au bout du compte, notre façon de décrire l'univers, de classifier les objets du monde évolue aussi. Cela pose problème quant à la description de l'information qui circule dans les réseaux. Dans cette optique, la

spécification explicite et détaillée de la conceptualisation d'un domaine (une ontologie) s'adapte mal à l'évolution des communautés.

Pour aborder ce problème, certains auteurs ont suggéré le concept de *sémantique émergente* comme solution (Mika, 2005). Cette idée suppose que les interactions individuelles d'un grand nombre d'individus rationnels peuvent produire un effet global émergent qui peut être vu comme une sémantique. On parle alors de réseaux sociaux sémantiques. Les ontologies sont vues comme un effet émergeant plutôt qu'un contrat entendu et limité de la majorité.

On observe ce phénomène dans les systèmes de *tagging* collaboratifs comme Delicious (http://delicious.com) et BibSonomy (http://www.bibsonomy.org), par exemple. Sur ces sites, on peut soit naviguer dans le contenu du site, à l'aide des tags, pour trouver des ressources d'intérêt, soit ajouter des ressources en leur assignant nos propres tags. Il est important de mentionner que lorsqu'un utilisateur publie une ressource sur le serveur et que le système lui demande d'entrer des mots-clés (tags) pour décrire le contenu de cette ressource, l'utilisateur peut voir les tags qui ont déjà été assignés à cette ressource par les autres utilisateurs du système.

On décrit donc l'information (des pages Web dans le cas de Delicious, des références dans le cas de BibSonomy) à l'aide de mots-clés (tags) de façon complètement non coordonnée. L'activité collective du *tagging* a pour effet de créer dynamiquement des correspondances entre des ressources et des ensembles de tags partagés par une communauté, qu'on appelle parfois folksonomie (folk + taxonomie). L'ensemble des tags utilisés pour une ressource ainsi que la fréquence d'utilisation de chaque tag à l'intérieur de cet ensemble représente la description combinée de cette ressource par plusieurs utilisateurs. On pourrait s'attendre à ce que les collections individuelles de tags qui varient constamment et les intérêts personnels de chaque utilisateur conduisent à quelque chose de chaotique. Cependant, on a montré que la fréquence relative des tags assignés aux ressources converge assez rapidement vers une valeur assez stable et que l'on assiste plutôt à l'émergence d'une catégorisation clairement définie en termes de tags (Cattuto, 2006 ; Dellschaft et Staab, 2008 ; Golder et Huberman, 2006).

On peut comprendre ce phénomène comme un processus autorenforçant dans lequel les choix des tags précédents renforcent constamment les choix subséquents par imitation. En effet, les utilisateurs de Delicious, par exemple, peuvent facilement imiter la sélection de tags des autres utilisateurs lorsqu'ils assignent des tags aux ressources qu'ils publient, car l'interface de ce système affiche les tags les plus fréquemment utilisés pour une ressource par les utilisateurs précédents. Ceci s'avère utile lorsque les utilisateurs ne savent pas trop comment catégoriser une URL particulière. Cette constante exposition aux tags populaires invite fortement les utilisateurs à les réutiliser, car ils représentent des choix sûrs (acceptés par la communauté) qui ne requièrent pas beaucoup de temps et d'effort. Au cours du temps, certains tags deviendront plus populaires que d'autres. Plus (respectivement moins) les tags seront populaires, plus (respectivement moins) ils auront de chances d'être copiés par des utilisateurs subséquents et ainsi de suite, jusqu'à l'atteinte d'un équilibre. Parce que des patrons stables émergent dans les proportions des tags, les choix minoritaires peuvent coexister avec les choix très populaires sans perturber le consensus établi par la communauté.

Le modèle de socialisation automatique que nous proposons rassemble déjà les individus qui partagent des intérêts similaires pour le partage de connaissances. En supposant que les individus dans le réseau partagent explicitement un certain nombre de documents qu'ils considèrent comme intéressants, il serait donc possible d'incorporer au modèle un mécanisme supplémentaire de partage automatique d'informations (de documents) entre les individus d'une même communauté d'intérêt, au sein du réseau.

Dans cette éventualité, on pourrait aussi ajouter un mécanisme de *tagging* collaboratif, mais décentralisé. Chacun des membres d'une communauté assignerait des mots-clés aux documents qu'il désire partager et lors d'un échange de documents, un membre qui reçoit un document, tout en prenant connaissance des tags déjà assignés par les membres précédents, pourrait aussi, à son tour, assigner d'autres (ou les mêmes) mots-clés au document reçu. Lors de visites communautaires, par exemple, on pourrait alors consolider les tags associés aux documents qu'on a en commun avec les autres membres de notre communauté. De cette manière, le vocabulaire des uns viendrait enrichir le vocabulaire des autres et possiblement

qu'une ontologie légère finirait par s'établir au sein des communautés. Ceci reste à voir et à tester cependant.

Supposons maintenant que le profil de chaque utilisateur soit compilé à l'aide de tous les tags qu'il a associés aux documents qu'il désire partager. Un utilisateur verrait son profil évoluer automatiquement à mesure qu'il ajoute des documents (étiquetés de mots-clés) à sa collection partageable. Mais de plus, son profil serait enrichi significativement des tags ajoutés, par les autres membres de sa communauté, aux documents qu'il a partagés avec ces autres membres. L'échange de vocabulaire, entre les membres d'une même communauté, est un atout significatif, car effectivement, selon les disciplines et selon les époques, on utilise souvent différents termes pour parler des mêmes choses.

Puisque nous n'abordons pas le partage automatique de documents, nous n'abordons pas le problème de l'évolution des catégories sémantiques dans le cadre de cette thèse. Cependant, nous réutiliserons cette idée de description des ressources à l'aide de mots-clés et d'extraction de profils d'intérêts à partir de l'agrégation de tous les mots-clés associés aux ressources partagées, pour tester notre modèle.

En résumé, ce travail de recherche propose un modèle dynamique de réseau social décentralisé (comme nos réseaux sociaux usuels, et non comme les sites de réseautage en ligne) dont l'évolution se base sur les mécanismes de socialisation observés dans nos réseaux sociaux réels.

Pour résoudre le problème de la localisation des individus de profils similaires, de la formation, du maintien et du renouvellement des communautés d'intérêts dans un réseau évolutif et décentralisé, notre objectif principal, au niveau informatique/fonctionnel, est de concevoir un modèle dont les algorithmes d'évolution du réseau doivent :

- produire des structures de réseaux efficacement navigables pour la localisation d'individus de profils similaires, lors des parcours de socialisation (formation et renouvellement de communautés).

- fournir des mécanismes de relais qui assurent la persistance des communautés lors du départ d'individus (maintien des communautés).
- fusionner les communautés d'intérêt similaires dispersées dans le réseau (formation et renouvellement des communautés).
- structurer efficacement les communautés pour que les visites communautaires soient simples et efficaces (faciliter les échanges entre les membres d'une même communauté).

Pour résoudre partiellement le problème de l'évolution des intérêts, notre objectif est de concevoir notre modèle de manière à ce que son architecture générique permette l'incorporation de diverses méthodes de comparaisons de profils pouvant s'adapter au contexte d'utilisation (pour une éventuelle implémentation du modèle). Nous voulons aussi que notre modèle puisse offrir un mécanisme de mise à jour des relations pour prendre en charge le repositionnement dans le réseau d'un individu dont les intérêts ont changé et dont la communauté d'appartenance ne répond plus à ses besoins.

Nous attendons de ce modèle qu'il fournisse des pistes utiles quant à l'implémentation éventuelle d'applications distribuées basées sur la formation et le maintien automatique de communautés d'intérêt au sein d'un réseau évolutif et décentralisé.

Au niveau sociocognitif, notre objectif est de montrer que les mécanismes de socialisation proposés dans notre modèle peuvent servir à expliquer, dans une certaine mesure, la présence de propriétés structurales typiques et récurrentes, observées dans nos réseaux sociaux réels, et que de surcroît, ceux-ci contribuent à créer un certain capital social.

Étant donné les différents points de vue abordés, on remarquera, dans nos travaux, une tension constante entre des considérations de modélisation sociocognitive (vraisemblance sociale du modèle) et des considérations fonctionnelles (efficacité informatique du modèle). Nous accorderons cependant la primauté au niveau fonctionnel.

1.3 Méthodologie

1.3.1 Approche expérimentale par simulation

Pour valider notre modèle, nous allons étudier son comportement par simulation. Nous composerons divers scénarios de simulations pour observer le comportement des réseaux générés par notre modèle au cours de leur évolution, et ce, dans divers contextes.

Plus précisément, nous voulons d'une part, concevoir des simulations avec socialisation et des simulations sans socialisation afin de les comparer pour déterminer si nos mécanismes de socialisation favorisent effectivement la formation de communautés d'intérêt au sein des réseaux.

D'autre part, nous composerons des scénarios avec données 1) fortement structurées, 2) moyennement structurées et 3) faiblement structurées. Le niveau de structuration signifie le niveau de similarité des acteurs présents dans le réseau. Effectivement, le comportement du modèle risque de varier selon qu'il y a beaucoup, moyennement ou peu d'utilisateurs de profils identiques dans le réseau, et donc plus ou moins de chances de former diverses communautés plus ou moins fournies.

Nous exécuterons donc 6 scénarios de simulation :

Scénario 1 : avec données fortement structurées
a) sans socialisation
b) avec socialisation

Scénario 2 : avec données moyennement structurées
a) sans socialisation
b) avec socialisation

Scénario 3 : avec données faiblement structurées
a) sans socialisation
b) avec socialisation

1.3.2 Collecte de données pour peupler nos simulations

Pour tester notre modèle par simulation, nous avons besoin d'acteurs dans nos réseaux et ces acteurs doivent posséder un profil d'intérêts. Nous avons choisi, à cette fin, de représenter les intérêts de chaque acteur par une liste de mots-clés.

Pour obtenir ces diverses listes de mots-clés qui seront attribuées aux divers acteurs de notre simulation, lors de leur création, nous voulons collecter cette information sur le site Web Delicious (http://delicious.com). Comme nous l'avons déjà expliqué brièvement, les utilisateurs de ce site publient des URL qui pointent vers des pages Web qu'ils apprécient pour les partager avec les autres. Chaque URL ainsi publiée est aussi associée de mots-clés qui décrivent le contenu de la page Web adressée par cette URL. Donc, à chaque utilisateur est associée une collection d'URL qui sont elles-mêmes associées à une collection de mots-clés. La réunion de tous les mots-clés de toutes les URL publiées d'un utilisateur peut donc servir à représenter son profil d'intérêt. Nous allons donc utiliser la base de données du site Delicious pour recueillir les collections d'URL de divers utilisateurs desquelles nous extrairons les différentes listes de mots-clés qui représenteront les profils de nos acteurs. La collecte et le traitement de nos données de simulations seront abordés plus en détail au chapitre 4 (voir sect. 4.2).

De plus, étant donné que les utilisateurs de Delicious peuvent utiliser le même mot-clé pour décrire plusieurs des URL qu'ils publient, les mots-clés ainsi compilés, pour la collection d'un utilisateur de Delicious, sont associés à une fréquence d'utilisation : certains mots-clés sont utilisés plus souvent que d'autres. Avec cette information, nous pourrons donc représenter nos profils d'intérêts (liste de mots-clés) sous forme de vecteurs de fréquences et nous pourrons comparer ces profils dans un espace vectoriel afin de déterminer la similarité entre deux acteurs. Cette méthode de comparaison sera expliquée plus en détail au chapitre 3 (voir art. 3.4.2).

1.3.3 Mesures et critères d'évaluation

Pour analyser les résultats de nos simulations et vérifier nos hypothèses, nous allons mesurer divers attributs des réseaux à plusieurs moments au cours de leur évolution.

D'un point de vue fonctionnel, nous avons fait l'hypothèse que nos mécanismes de socialisation favorisent la formation de communautés d'intérêts et que le maintien de certaines propriétés structurales du réseau avantage la navigabilité.

Les mesures que nous utiliserons pour vérifier que les propriétés structurales imposées favorisent effectivement l'efficacité des parcours dans nos réseaux générés sont 1) la densité, 2) la distance moyenne, 3) le diamètre et 4) le nombre de nœuds pivots. Nous nous attendons à obtenir une densité faible, une distance moyenne courte et un petit diamètre. Ces trois caractéristiques structurales impliquent déjà qu'il n'existe pas trop de chemins différents dans le réseau (densité faible) et que de courts chemins existent (faible distance moyenne et petit diamètre). De plus, nous nous attendons à ce que ces caractéristiques soient encore plus marquées dans les scénarios avec socialisation. En ce qui concerne le nombre de nœuds pivots, nous verrons, au chapitre 3, que nous pourrons limiter les parcours de connexion/socialisation, grâce aux propriétés structurales imposées, à un sous-ensemble de nœuds (qu'on appelle les pivots) dans nos réseaux. Donc moins il y a de nœuds pivots dans les réseaux, moins les parcours risquent d'être longs. Nous nous attendons à ce que le nombre de noeuds pivots soit beaucoup plus petit dans les scénarios avec socialisation.

Ensuite, pour évaluer la capacité de nos mécanismes de socialisation à former des communautés d'intérêt, nous utiliserons une mesure 1) d'homophilie et 2) de la proportion des rencontres effectuées. La première mesure évalue à quel point sont similaires les individus qui sont reliés ensemble dans le réseau et la seconde calcule la proportion moyenne du nombre d'individus qui se sont rencontrés sur le nombre d'individus qui auraient dû se rencontrer. Encore une fois, nous nous attendons à ce que la valeur de ces deux mesures soit beaucoup plus élevée dans les scénarios avec socialisation.

D'un point de vue sociocognitif, nous avons fait l'hypothèse que les mécanismes de socialisation peuvent servir à expliquer, du moins en partie, la présence de propriétés structurales particulières qu'on observe dans les réseaux sociaux réels : l'effet des petits mondes, une transitivité élevée, l'émergence de communautés structurales et une distribution des degrés suivant une loi de puissance (nous verrons ces propriétés typiques au chapitre 2).

Nous utiliserons 1) la mesure de la distance moyenne pour mesurer l'effet des petits mondes, 2) le coefficient de *clustering* pour mesurer le niveau de transitivité, 3) une mesure de modularité pour vérifier la présence de communautés structurales bien définies dans nos réseaux et finalement, nous vérifierons 4) la distribution des degrés de nos réseaux pour voir si elle suit bien une loi de puissance. Nous nous attendons à ce que les réseaux générés par simulation présentent toutes ces caractéristiques typiquement observées dans les réseaux sociaux usuels, à l'exception d'une seule, la transitivité élevée. En effet, nous avons délibérément éliminé, le plus possible, le taux de *clustering* dans nos réseaux (parce qu'il augmente la densité) au profit d'une plus grande efficacité au point de vue fonctionnel.

De plus, nous avons supposé que les mécanismes de socialisation favorisent, dans une certaine mesure, la création de capital social et en ce sens, nous nous attendons à une mesure de centralité de proximité plus élevée dans les scénarios avec socialisation.

Les mesures discutées dans cette section (à l'exception du nombre de nœuds pivots) seront expliquées en détail au chapitre 2 (voir art. 2.3.1).

1.4 Contributions de la thèse

La contribution principale de cette thèse est un modèle de socialisation automatique cohérent qui gère la dynamique d'un réseau social complètement décentralisé en maintenant, entretenant et renouvelant les communautés d'intérêts au sein de ce réseau en constante évolution.

Dans une optique de modélisation sociocognitive, notre modèle s'inspire de la réalité sociale et plus précisément de la manière dont les individus tendent à se regrouper entre eux, selon les intérêts qu'ils partagent, par le biais de la socialisation. Nous avons formalisé deux types de mécanismes de socialisation, les rencontres aléatoires et les recommandations. Dans le cadre de cette thèse, nous proposons ces mécanismes comme processus d'évolution pouvant expliquer, du moins en partie, la formation de nos réseaux sociaux réels, au niveau structural. À notre connaissance, il n'existe pas de modèle d'évolution de réseaux sociaux qui propose de tels mécanismes. De plus, notre modèle sert à expliquer, dans une certaine mesure, les motivations sous-jacentes à l'acte de socialisation : les individus socialisent parce que ça leur est utile, parce que ça leur apporte un certain capital social.

D'un point de vue informatique/fonctionnel, notre apport principal se situe au niveau de l'efficacité du modèle à 1) localiser les individus de profils similaires et 2) les regrouper ensemble au sein de communautés d'intérêts en n'utilisant que l'information locale disponible et malgré le remaniement continuel de la structure du réseau.

Pour ce faire, nous proposons des mécanismes de socialisation originaux qui permettent le regroupement des individus similaires par la restructuration du patron des relations entre les individus à mesure que ceux-ci se rencontrent, lors de parcours effectués dans le réseau. De plus, l'originalité de notre approche réside dans le fait que nous implémentons nos algorithmes d'évolution du réseau (les règles d'évolution du modèle) de manière à ce qu'ils contrôlent et maintiennent certaines propriétés structurales qui favorisent la navigabilité. Plus précisément, ces propriétés structurales nous permettent de :

- *former de communautés non équivoques au niveau structural* : les propriétés structurales que nous imposons permettent à chaque membre d'une même communauté de se trouver, se visiter les uns les autres, de manière optimale. Ainsi, une fois que les individus sont regroupés, ceux-ci ont facilement accès à tous les membres de leur communauté d'appartenance.

- *faire des parcours efficaces dans le réseau* : nous verrons, au chapitre 3 (art. 3.3.1), que la recherche dans les graphes est généralement un problème computationnel difficile, de complexité exponentielle. En perpétuant ces propriétés structurales désirables pour la navigabilité, nous sommes capables de rendre plus efficaces de simples parcours heuristiques dans le graphe. Ceux-ci sont efficaces au sens où ils sont courts, qu'ils permettent de trouver les individus similaires pour les regrouper ensemble et qu'ils fonctionnent en temps réel, dans un réseau qui évolue constamment.

Pour la formation et le maintien efficaces des communautés d'intérêts au sein du réseau, nous proposons donc une dynamique de renforcement entre la navigabilité et la socialisation : la navigabilité favorise la localisation d'individus similaires lors des parcours de socialisation, les mécanismes de socialisation avantagent à leur tour le rapprochement des individus, ce rapprochement favorisant encore davantage la navigation, et ainsi de suite.

Comme contribution secondaire, notre modèle prévoit aussi un mécanisme de mise à jour des relations pour prendre en charge automatiquement le repositionnement d'un individu au sein du réseau lorsque ses intérêts ont changé radicalement et qu'il ne se trouve plus dans une communauté qui lui ressemble. De plus, l'architecture générique de notre modèle, qui nécessite de fournir une méthode spécifique de comparaison des profils, permet de combiner ce mécanisme de mise à jour des relations avec un mécanisme implicite (automatique) de mise à jour des profils d'intérêts, rendant ainsi possible l'automatisation complète du processus de mise à jour.

Lorsque des choix s'imposaient, nous avons privilégié l'utilisabilité informatique (sur la vraisemblance sociale), et en ce sens, nous proposons ce modèle dans l'optique qu'il puisse être utilisé pour la conception et l'implémentation d'applications distribuées, basées sur la formation et le maintien automatiques de communautés d'intérêt pour l'échange et le partage de connaissances entre les individus.

1.5 Organisation de la thèse

Au chapitre 2, nous présentons divers concepts et différents travaux tirés de la littérature qui sont en lien avec les aspects de notre recherche. En particulier, nous voyons d'abord que sur les sites de réseautage en ligne, les mécanismes de socialisation sont très similaires à ceux qui ont lieu dans nos réseaux sociaux réels. Ensuite, nous abordons les études portant sur la structure des réseaux sociaux. Nous verrons différentes mesures qui servent à caractériser la structure des réseaux, les propriétés structurales typiques qu'on observe dans nos réseaux sociaux usuels (réels et virtuels) ainsi que divers modèles qui tentent d'expliquer la présence de ces propriétés structurales récurrentes. Nous verrons aussi que dans les réseaux sociaux, le patron des relations entre les individus influence leur navigabilité ainsi que la manière dont l'information y circule. Finalement, nous présentons diverses études sur les systèmes sociaux et collaboratifs qui, comme la socialisation, tentent de tirer profit de la collaboration entre les individus pour filtrer les connaissances et ainsi mieux personnaliser l'accès à l'information en tenant compte des besoins spécifiques (intérêts) des individus.

Au chapitre 3, nous décrivons notre modèle de socialisation. Tout d'abord, nous formalisons les règles d'évolution de notre modèle, dont les mécanismes de socialisation (rencontres aléatoires et recommandations) que nous proposons. Puis, nous discutons de la structure choisie pour former les communautés d'intérêt au sein des réseaux générés par notre modèle et pour quelles raisons celle-ci favorise leur navigabilité. Nous expliquons ensuite nos algorithmes qui implémentent les règles d'évolution de notre modèle, selon deux variantes du regroupement choisi pour la formation des communautés : les regroupements 1) par nœuds pivots simples et 2) par nœuds pivots chaînés. Enfin, nous voyons les méthodes que nous avons choisies pour représenter et comparer les profils d'intérêts des acteurs du réseau géré par notre modèle pour en déterminer la similitude.

Au chapitre 4, nous présentons les résultats de la validation de notre modèle de socialisation, par simulation, au point de vue informatique et sociocognitif. Nous expliquons d'abord la manière dont nous avons obtenu des profils d'intérêts réalistes pour la création des acteurs de nos réseaux. Nous détaillons ensuite nos scénarios de simulations ainsi que les mesures

utilisées pour analyser les réseaux générés par notre modèle. Nous discutons finalement les résultats de nos analyses quant à l'efficacité de notre modèle à regrouper les acteurs similaires en communautés d'intérêt (au niveau fonctionnel) et quant à sa capacité à représenter les propriétés structurales qu'on observe dans les réseaux sociaux usuels ainsi qu'à créer un certain capital social (au niveau sociocognitif).

Finalement, au chapitre 5, nous présentons un bilan des travaux effectués dans le cadre de cette thèse, puis nous proposons quelques pistes de recherche pour le futur, en continuité avec les travaux accomplis.

CHAPITRE II

CONTEXTE THÉORIQUE

2.1 Introduction

La socialisation est un moyen de découvrir nos réseaux, de découvrir les individus intéressants qui s'y cachent et par le fait même, de l'information potentiellement utile et pertinente. En plus des lieux de socialisation traditionnels, la montée en popularité des communautés en ligne, et plus particulièrement des réseaux sociaux virtuels, exacerbe continuellement nos possibilités de socialisation. À la section 2.2, nous expliquons brièvement ce que sont les sites de réseautage en ligne ainsi que leur fonctionnement en général pour montrer que la manière de socialiser au sein de ces réseaux virtuels est très similaire à notre façon de procéder dans le monde réel. Que ce soit dans le monde réel ou virtuel, la recherche active de nouveaux contacts requiert toujours un certain investissement. C'est dans cette optique que nous pensons que l'automatisation de la socialisation serait un avantage notable.

À la section 2.3, nous abordons l'étude structurale des réseaux sociaux, réels ou virtuels. Nous expliquons tout d'abord certaines mesures utiles pour l'analyse de la structure des réseaux, tirées de la littérature. Au chapitre 4, plusieurs d'entre elles nous serviront alors à valider notre modèle et à caractériser la structure des réseaux générés par celui-ci afin de les comparer aux structures qu'on observe dans le monde (réel ou virtuel).

Nous présentons ensuite plusieurs travaux portant sur l'analyse des propriétés structurales des réseaux sociaux, en insistant sur les propriétés générales et récurrentes qu'on observe dans la réalité. Au chapitre 4, d'un point de vue sociocognitif, nous analyserons alors dans quelle mesure les mécanismes de socialisation proposés dans notre modèle peuvent servir à expliquer les propriétés structurales effectivement constatées dans les réseaux sociaux. Au

chapitre 3, les études structurales présentées dans cette section nous serviront aussi à motiver nos choix quant à la structure des communautés d'intérêts dans nos réseaux.

À la section 2.4, nous faisons un survol des modèles importants proposés dans la littérature, suggérant divers processus d'évolution des réseaux qui expliqueraient la formation de ces structures particulières observées dans la réalité. À notre connaissance, aucun modèle ne propose les mécanismes de socialisation tels que décrits dans cette thèse comme processus d'évolution des réseaux sociaux.

À la section 2.5, nous verrons que de connaître les propriétés structurales des réseaux sociaux aide à mieux guider la navigation à l'aveugle dans le patron des connexions qui les forme et permet ainsi une localisation d'informations/individus plus efficace. Nous verrons aussi, à la section 2.6, comment la topologie des réseaux peut influencer la diffusion de l'information au sein de ces réseaux. Nous voulons montrer ici que les caractéristiques structurales des réseaux déterminent, dans une certaine mesure, l'efficacité de la navigation/recherche et de la dissémination de l'information. C'est en ce sens que nous proposons de contrôler, en partie, la structure des réseaux générés par notre modèle de telle sorte qu'en plus de rassembler les individus d'intérêts communs, elle favorise aussi la navigation à l'aveugle lors de la recherche de contacts similaires dans les réseaux (lors des parcours de socialisation que nous verrons plus en détail au chapitre 3).

Finalement, à la section 2.7, nous faisons un bref survol des méthodes utilisées par les systèmes dits "sociaux" ou "collaboratifs". En effet, notre projet s'inscrit dans la lignée de ce type de systèmes, car la socialisation est, à notre sens, un processus fondamentalement collaboratif au sens de l'échange : échange d'informations, échange de contacts personnels.

De plus, un bon nombre de travaux dans le domaine de l'informatique sociale s'est penché sur le problème de la création de profils d'intérêts des utilisateurs, à partir d'informations diverses disponibles selon le contexte. Nous aborderons quelques-unes de ces méthodes parmi lesquelles nous choisirons celle que nous utiliserons pour caractériser le profil des utilisateurs qui peupleront nos simulations au chapitre 4. Il est à noter que notre modèle de socialisation

est générique au sens où il permet l'utilisation de toutes sortes de profils (d'intérêts ou autre), pourvu que l'information pour les concevoir soit disponible et qu'on puisse fournir une méthode de comparaison qui retourne une valeur numérique indiquant le niveau de similarité entre deux profils comparés.

2.2 La socialisation dans les réseaux virtuels

2.2.1 Qu'est-ce qu'un site de réseautage en ligne (médias sociaux)

Il existe aujourd'hui plusieurs sites de réseautage en ligne (ou réseaux sociaux virtuels), qui continuent à se multiplier et dont la popularité ne cesse de croître. Bien qu'ayant souvent des motivations sous-jacentes diverses, ces sites offrent généralement la possibilité aux participants de créer et maintenir des contacts sociaux, de publier diverses formes de contenu et de consulter le profil ou le contenu publié par les autres membres du réseau.

Ces réseaux sociaux en ligne constituent une nouvelle forme de réseaux d'informations, différente des réseaux plus traditionnels comme le Web. Pour n'en nommer que quelques-uns, parmi les plus populaires, on retrouve Facebook (http://www.facebook.com), Twitter (https://twitter.com) et LinkedIn (http://www.linkedin.com). Pour le partage de contenu plus spécifique, on pense à Flickr (http://www.flickr.com) pour la publication de photos, YouTube (http://www.youtube.com) pour le partage de vidéos et d'autres encore comme Blogster (http://www.blogster.com) pour le partage de blogues et Delicious (http://delicious.com) pour le partage d'URL. Pour un aperçu plus exhaustif des sites bien connus de réseautage en ligne, voir Wikipédia (http://en.wikipedia.org/wiki/List_of_social_networking_websites).

2.2.2 Mécanismes généraux d'utilisation et de fonctionnement

Quelle que soit leur orientation, on observe des mécanismes de fonctionnement et d'utilisation récurrents dans la majorité des sites de réseautage en ligne. Tout d'abord, les utilisateurs doivent créer un compte, une identité, pour s'inscrire sur le réseau et obtenir l'accès au contenu du site à quelques exceptions près qui permettent l'accès du contenu public

sans connexion explicite au système à l'aide d'un compte utilisateur. Les participants peuvent alors fournir des informations qui les concernent comme leur lieu de naissance, leur sexe, leur niveau d'étude, leurs intérêts, etc. L'ensemble de ces informations constitue leur profil d'utilisateur.

Pour la formation du réseau en tant que tel, chaque participant doit construire son réseau personnel de contacts en faisant la demande explicite de connexion aux autres utilisateurs qu'il aimerait avoir comme relation. Certains sites ne permettent la création d'un lien que si l'utilisateur sollicité accepte, d'autres n'imposent pas cette condition. Chaque membre du réseau peut aussi publier du contenu sur son compte pour le partager avec les autres utilisateurs. En général, les sites offrent la possibilité de spécifier si le contenu est privé et donc réservé à la seule vue des contacts immédiats de l'utilisateur ou bien s'il est public et accessible à tous.

Le réseau de contacts ainsi que le contenu publié par un utilisateur sont accessibles aux autres participants. On peut donc naviguer dans le réseau, en suivant les liens, de profil en profil et consulter le contenu publié de chacun des utilisateurs visités. Si au passage, on rencontre des profils qui nous semblent intéressants, on peut décider d'en faire des contacts personnels. Aussi, plusieurs sites, comme Facebook par exemple, permettent de faire des recommandations de contacts : suggérer à l'un de ses contacts personnels la création d'une relation avec un autre de ses contacts personnels, parce qu'on pense qu'ils auraient intérêt à se connaître.

En ce sens, la manière de socialiser, sur un site de réseautage en ligne, est très similaire à celle qu'on utilise dans nos réseaux sociaux du monde réel, telle que décrite au chapitre 1 (sect. 1.2). On navigue dans le réseau des utilisateurs dans l'espoir de faire des rencontres intéressantes et l'on se recommande les uns aux autres. Certains sites, comme LinkedIn, fixent une limite sur l'étendue du voisinage d'un utilisateur qu'on peut visiter tandis que d'autres, qui n'imposent pas cette limitation, permettent de consulter de cette manière tous les utilisateurs (et leur contenu) présents dans le réseau.

Les réseaux sociaux en ligne, comme les réseaux hors ligne, fournissent non seulement des manières de diffuser du contenu, mais aussi de localiser des individus intéressants par navigation dans le réseau. De mieux comprendre la structure des réseaux, c'est-à-dire les caractéristiques structurales particulières des patrons de connexions (des chemins de navigation) qui forment le réseau, peut ainsi conduire au développement de meilleurs systèmes de recherche et de diffusion de l'information.

Dans cette optique, à la section suivante, nous expliquons tout d'abord quelques mesures de réseau utiles pour caractériser quantitativement la structure des réseaux et nous présentons ensuite diverses études qui se sont penchées sur l'analyse des propriétés structurales des réseaux sociaux, du monde réel et virtuel.

2.3 Propriétés structurales des réseaux

2.3.1 Formalisation et mesures des réseaux

Un réseau social peut être défini comme étant un ensemble de relations entre un ensemble fini d'individus, ceux-ci faisant figure d'acteurs à l'intérieur du réseau. On peut considérer plusieurs types de relations (collaboration, soutien, contrôle, conseil, influence, parenté, amitié, etc.) qu'entretiennent les acteurs les uns avec les autres au sein d'un ensemble social. À titre d'exemple, prenons les employés d'une entreprise. On peut alors se demander qui collabore avec qui au sein de cette société. On obtient alors un réseau social formé de l'ensemble des relations de collaboration entre les employés de cette entreprise. Dans cet exemple, le réseau est constitué de relations entre individus, mais nous pourrions, de la même façon, considérer les relations qui existent entre différents groupes sociaux. Dans ce cas, l'acteur n'est plus un individu, mais un ensemble d'individus. Par exemple, nous pourrions étudier les relations de collaboration entre plusieurs entreprises, chaque entreprise faisant figure d'acteur à l'intérieur du réseau. Dans cet esprit, lorsqu'on parle du réseau global, c'est toujours une question de point de vue.

Les réseaux sociaux, en raison de leur nature, se représentent bien sous forme de graphe : chaque acteur d'un réseau social est représenté par un sommet (ou noeud) du graphe et chaque relation entre deux individus est représentée par un lien. Par analogie, il est donc possible d'appliquer les concepts de la théorie des graphes aux réseaux sociaux.

2.3.1.1 Quelques concepts de base en théorie des graphes

On parle d'un *graphe orienté* lorsque les liens entre les sommets sont orientés (les relations sont unidirectionnelles). Un lien orienté s'appelle un *arc*. Par opposition, un *graphe non orienté* est composé de liens non dirigés (les relations sont bidirectionnelles) qu'on appelle *arêtes*.

Dans le cas des graphes non orientés, le *degré* d'un sommet correspond à son nombre d'arêtes. Dans le cas d'un graphe orienté, le *demi-degré intérieur* d'un sommet est le nombre d'arcs entrant sur ce sommet et le *demi-degré extérieur* est le nombre d'arcs sortant de ce sommet. Le *degré* d'un sommet dans un graphe orienté est alors la somme du demi-degré intérieur et du demi-degré extérieur.

En théorie des graphes, un *chemin* est une succession de liens que l'on parcourt pour se rendre d'un sommet i à un sommet j. Le nombre de liens qui constitue un chemin est la *longueur du chemin.* S'il n'existe pas de chemin allant de i à j, nous dirons que la longueur du chemin est égale à 0. On dit que deux sommets sont *adjacents* s'il existe un lien entre ces deux sommets (un chemin de longueur égale à 1). On parle aussi de *relation directe* dans ce cas. On dira alors qu'il existe une *relation indirecte* entre i et j s'il existe un chemin entre i et j, mais dont la longueur est plus grande que 1.

On dit qu'un graphe est *connexe* lorsqu'il existe au moins un chemin pour relier chacun des couples de sommets dans le réseau.

Les mesures de réseaux sont nombreuses et de nouvelles apparaissent constamment. Nous en présentons ici quelques-unes qui nous seront utiles pour discuter de nos résultats d'analyse

présentés au chapitre 4. Pour une liste plus substantielle, se référer à (Wassermann et Faust, 1994 ; Scott, 2000).

2.3.1.2 Densité

Sachant que dans un graphe de n sommets, le nombre maximum possible de liens est égal à $n(n-1)$ pour un graphe orienté et à $\frac{n(n-1)}{2}$ pour un graphe non orienté, on peut alors calculer la densité d'un graphe comme suit :

$$\text{densité} = \frac{\text{nombre de liens}}{\text{nombre maximum possible de liens}} .$$

La densité indique à quel point un graphe est fortement connecté. Un graphe complètement connecté (dans lequel le nombre de liens est égal au nombre maximum possible de liens) aura donc une densité égale à 1, et plus il est creux (ou éparse), plus la densité tendra vers 0.

2.3.1.3 Distance

Le *plus court chemin* entre deux sommets i et j est le chemin de longueur minimale entre i et j.

La *distance* (ou distance géodésique) entre deux sommets i et j est la longueur du plus court chemin entre i et j. La distance entre i et j égale 0 s'il n'y a aucun chemin qui relie i à j.

Le *diamètre* d'un graphe correspond à la plus grande distance séparant n'importe quelle paire de sommets dans le graphe. C'est la distance géodésique maximale entre deux membres.

On peut calculer la *distance moyenne* d'un graphe en calculant la moyenne des distances entre toutes les paires de sommets i et j, où $i \neq j$.

La distance moyenne et le diamètre nous informent dans quelle mesure les sommets du graphe sont éloignés les uns des autres.

2.3.1.4 Modularité

Dans un réseau, on définit les communautés structurales comme étant des ensembles de sommets très fortement connectés les uns aux autres, à l'intérieur desquels, donc, la densité des liens est beaucoup plus élevée que la densité des liens qui relient ces communautés entre elles. Lorsqu'on examine les communautés au sein des réseaux, on a souvent besoin d'une mesure objective qui évalue à quel point une division particulière d'un réseau en communautés est significative. Une de ces mesures est la mesure de modularité proposée par (Newman, 2004b).

Supposons un réseau partitionné en k communautés. Soit E, la matrice carrée symétrique dont les éléments e_{ij} représentent la fraction des liens du réseau qui connectent les sommets de la communauté i à la communauté j (par rapport à tous les liens dans le réseau). On remarque ici que les éléments e_{ii} sur la diagonale de E correspondent à la fraction des liens du réseau qui se trouvent à l'intérieur d'une même communauté i. On définit ensuite $a_i = \sum_j e_{ij}$ qui représente la fraction de tous liens qui touchent les sommets dans la communauté i. Dans un réseau où les liens seraient distribués entre les sommets sans tenir compte des communautés auxquels ils appartiennent, nous aurions $e_{ij} = a_i a_j$. Donc, on peut définir une mesure de modularité Q comme suit :

$$Q = \sum_i (e_{ii} - a_i^2) = Tr\,(E) - \left\| E^2 \right\|,$$

où $Tr\,(E)$ est la trace de la matrice E (la somme des éléments sur la diagonale de E) et $\left\| E^2 \right\|$ représente la somme de tous les éléments de la matrice E^2.

Ainsi, la modularité est une mesure de la proportion des liens intra-communautés moins la valeur de la même quantité qu'on obtiendrait dans un réseau avec le même partitionnement en communautés, mais dont les liens seraient distribués aléatoirement, sans tenir compte du partitionnement. Une valeur de modularité égale à 0 indique qu'on ne trouve pas plus de structure de communautés qu'on n'en retrouverait dans un graphe aléatoire tandis qu'une valeur positive significative représente la présence de communautés dans la structure. En pratique, Newman fait remarquer que des valeurs au-dessus de 0.3 semblent indiquer des structures de communautés significatives et qu'en général, on observe dans les réseaux réels, des valeurs entre 0.3 et 0.7 (1 étant le maximum pour cette mesure). Au chapitre 4, nous utiliserons cette mesure pour vérifier que notre partitionnement en communautés d'intérêt se reflète aussi au niveau structural.

2.3.1.5 Distribution des degrés

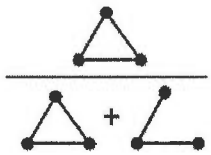
La distribution des degrés est une fonction *P(d)* qui décrit la proportion des sommets de degré *d* dans le réseau.

La plupart des réseaux sociaux ne se forment pas de manière complètement aléatoire. Par exemple, nous verrons que la distribution des degrés de certains réseaux suit une loi de puissance selon laquelle seulement une petite proportion des noeuds du réseau possèdent un très grand nombre de liens tandis qu'une grande proportion des noeuds n'ont qu'un faible degré.

2.3.1.6 Transitivité (ou clustering)

La présence significative de transitivité dans un réseau est un autre indicateur de structuration qui signifie que la structure du réseau ne s'est pas formée de manière complètement aléatoire.

La transitivité, qu'on appelle aussi *clustering*, indique à quel point les voisins immédiats des sommets sont aussi reliés entre eux. C'est une mesure de la proportion du nombre de triangles (ou triades fermées) par rapport aux triades connexes dans le réseau :



Le coefficient de *clustering* C_i, pour le sommet i est donc :

$$C_i = \frac{\text{nombre de triangles connectés à } i}{\text{nombre de triades connexes centrées sur } i} \quad \text{(Newman, 2003),}$$

où $C_i = 0$ pour les sommets de degré 0 ou 1 (pour lesquels le dénominateur est égal à 0).

On peut ensuite calculer le coefficient de *clustering* C pour le réseau complet en prenant la moyenne des C_i :

$$C = \frac{1}{n} \sum_i C_i \quad \text{, où } n \text{ est le nombre de sommets dans le réseau.}$$

Le coefficient de *clustering* varie entre 0 et 1 et des valeurs plus élevées signifient une plus grande tendance à former des cliques au sein du réseau. On remarque, en effet, une transitivité élevée dans beaucoup de réseaux sociaux, car souvent, les amis de nos amis sont aussi nos amis.

Les mesures de centralité suivantes déterminent l'importance des acteurs au sein du réseau. Elles servent à mesurer à quel point la position des acteurs dans le réseau est susceptible de leur procurer du pouvoir ou un certain contrôle sur la circulation des ressources. Ce sont donc des mesures du capital social au niveau des acteurs. Nous présentons ici trois mesures de centralité bien connues dans la littérature (Lazega, 1998) que nous utiliserons pour analyser nos réseaux générés par simulation, au chapitre 4.

2.3.1.7 La centralité

Centralité de degré (degree centrality)

On dit qu'un acteur possède une grande centralité de degré lorsqu'il possède plusieurs contacts, c.-à-d. que le sommet qui le représente dans le graphe a un degré élevé. C'est donc la mesure de la taille du réseau d'un acteur. Plus un acteur est central, plus il est actif dans le réseau.

La centralité de degré pour un acteur i, Cd_i, est alors la somme de ses liens directs :

$$Cd_i = \sum_{j=1}^{n} x_{ij} \quad \text{où } x_{ij} \text{ est la valeur du lien de } i \text{ à } j, \text{ pour tout } j \neq i .$$

La valeur maximum de cette mesure est n-1 dans le cas d'un acteur qui est relié à tous les autres acteurs du réseau. Pour obtenir une centralité de degré $C'd_i$ dont les valeurs sont entre 0 et 1, on peut donc standardiser comme suit :

$$C'd_i = \frac{Cd_i}{n-1} .$$

Centralité de proximité (closeness centrality)

La centralité de proximité mesure à quel point un acteur est proche de tous les autres. Le score de centralité pour un acteur i est obtenu en prenant l'inverse de la somme des distances reliant cet acteur à tous les autres :

$$Cc_i = \frac{1}{\sum_{j=1}^{n} d_{ij}} \quad \text{où } d_{ij} \text{ est la distance géodésique entre les acteurs } i \text{ et } j, \text{ pour tout } j \neq i .$$

Le score maximum possible pour la centralité de proximité est alors de $\frac{1}{n-1}$ où n est le nombre d'acteurs dans le réseau. On peut donc standardiser les valeurs entre 0 et 1 en calculant $C'c_i$ de cette manière :

$$C'c_i = \frac{n-1}{\sum_{j=1}^{n} d_{ij}} = (n-1)\, Cc_i \ .$$

Ainsi, $C'c_i$ vaut 1 lorsque l'acteur i est adjacent à tous les autres.

Centralité d'intermédiarité (betweenness centrality)

La centralité d'intermédiarité mesure à quel point un acteur se trouve sur des passages obligés lors de communications entre d'autres acteurs. En effet, lorsque deux acteurs ne sont pas adjacents, ils dépendent des autres pour communiquer ensemble. Ainsi, plus un acteur est central de ce point de vue, plus il exerce un contrôle sur la circulation de l'information dans le réseau. Cet indice, pour un acteur i, représente le rapport entre tous les plus courts chemins (les géodésiques) entre j et k qui passent par i et l'ensemble total de tous les plus courts chemins entre j et k :

$$Cb_i = \frac{\sum_{i \neq j \neq k} g_{jk}(i)}{g_{jk}} \ ,$$

où g_{jk} représente l'ensemble des géodésiques entre j et k et où $g_{jk}(i)$ est un plus court chemin entre j et k passant par i.

Le score minimum de centralité d'intermédiarité est 0 lorsque i ne se trouve sur aucun géodésique et le maximum est $\frac{(n-1)(n-2)}{2}$ si i tombe sur toutes les géodésiques. La valeur standardisée entre 0 et 1, $C'b_i$, est donc :

$$C'b_i = \frac{Cb_i}{\frac{(n-1)(n-2)}{2}} \quad .$$

Pour mieux comprendre ce que ces mesures représentent en terme de capital social, des bénéfices reliés à ses relations sociales, regardons la figure 2.1 qui illustre le « Kite Network » développé par David Krackhardt. Cette figure illustre très bien ce que signifient les différentes mesures de centralité.

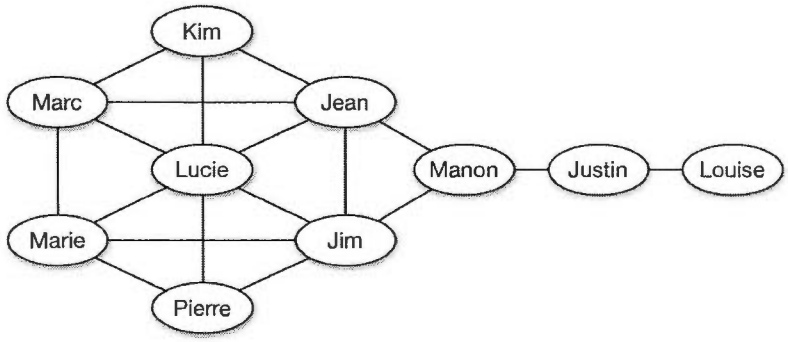


Figure 2.1 Illustration du « Kite Network » développé par David Krackhardt.
(Tirée de http://www.orgnet.com/sna.html.)

On observe que dans ce réseau, Lucie obtient le meilleur résultat pour la centralité de degré, car c'est elle qui possède le plus de liens directs. Cependant, le fait d'avoir plusieurs relations n'est pas toujours aussi important que les personnes avec qui l'on entretient ces relations. En effet, on remarque que les contacts de Lucie font tous partie de sa propre clique, ils sont tous reliés entre eux. Il aurait été plus avantageux pour Lucie d'avoir des contacts hors clique afin de devenir un intermédiaire obligé entre les membres de sa clique et ces contacts extérieurs.

De son côté, Manon possède peu de liens directs, moins que la moyenne dans le réseau. Son score de centralité de degré est donc relativement petit par rapport aux autres membres du réseau. Par contre, du point de vue de la centralité d'intermédiarité, elle occupe une des meilleures positions dans le réseau. Elle joue le rôle de l'unique intermédiaire entre les membres de la clique de gauche et les deux membres à sa droite. Elle peut ainsi contrôler les communications entre ces deux parties. Elle est en position de pouvoir, car sans elle, Justin et

Louise ne pourraient pas avoir accès à l'information et aux connaissances provenant de la clique de Lucie. Un acteur ayant un fort degré de centralité d'intermédiarité a beaucoup d'influence sur la circulation des connaissances dans le réseau.

Jean et Jim possèdent un peu moins de connexions que Lucie cependant, la structure de leurs connexions directes et indirectes leur permet d'avoir accès à tous les autres acteurs du réseau plus rapidement que n'importe quel autre membre. Ils ont accès aux plus courts chemins vers tous les autres acteurs. Ils sont proches de tout le monde et possèdent donc le score de centralité de proximité le plus élevé. Ils ont accès à l'information plus rapidement que tous les autres. Ils sont alors dans une excellente position pour exercer un contrôle sur le flux d'information dans le réseau, car ils ont la meilleure visibilité de ce qui y circule.

2.3.1.8 Densité du réseau personnel d'un acteur (ego-network)

La densité, calculée sur le réseau des contacts personnels d'un acteur *i*, est aussi un indicateur du capital social. C'est une mesure de la proportion des contacts de l'acteur *i* qui sont reliés ensemble. Une densité élevée est négative, car elle signifie que les contacts personnels de l'acteur sont très fortement reliés et qu'ils sont dès lors redondants. Il est plus profitable d'avoir des pairs qui ont eux-mêmes des contacts ailleurs dans le réseau, qui forment ainsi des chemins possibles vers d'autres opportunités.

2.3.1.9 Mesures de cohésion du réseau

La distance moyenne (ou le diamètre), comme discutée précédemment, peut aussi informer sur la capacité du réseau à propager l'information entre les individus. En terme de capital social, plus la distance moyenne est petite, plus les communications entre les membres du réseau seront rapides, ce qui est un atout.

Il existe, dans la littérature, beaucoup d'autres mesures du capital social et nous n'avons exposé ici que celles qui seront utiles pour notre analyse. Pour une revue plus complète, nous référons le lecteur à (Borgatti et al., 1998 ; Lazega, 1998).

Les mesures suivantes sont des mesures qui serviront à évaluer l'efficacité de nos mécanismes de socialisation par rapport à la formation de communautés d'intérêt.

2.3.1.10 Homophilie

L'homophilie informe à quel point les acteurs qui entretiennent des relations sont semblables entre eux. En terme de capital social, cette mesure est considérée comme négative. En effet, peu d'homophilie devrait procurer une plus grande exposition à la diversité et aux nouvelles idées, mais en réalité, l'homophilie est une caractéristique observée dans la plupart de nos réseaux sociaux réels ou virtuels (McPherson et al., 2001 ; Crandall et al., 2008 ; Lewis et al., 2008).

Dans le cadre de cette thèse, l'homophilie est cruciale puisque nous voulons explicitement former des communautés d'individus qui se ressemblent. Notons cependant que, dans le cas qui nous occupe, on parle de similarité des intérêts et non de similarité globale. On veut construire des communautés autour des intérêts spécifiquement. Ainsi, rien n'empêche qu'au sein d'une communauté, les individus diffèrent sur d'autres plans comme leur lieu de naissance, leur langue natale, leur milieu et niveau d'éducation, leur âge, leur domaine professionnel, leur milieu culturel, etc. Dans le cadre de ce travail, lorsque nous parlons d'homophilie, c'est au sens de la similarité au niveau des intérêts uniquement.

Pour mesurer l'homophilie moyenne globale (désirable) au niveau du réseau, nous calculons la moyenne du score de similarité entre tous les individus qui sont en relation dans le réseau comme ceci :

$$H = \frac{\sum_{i,j=1}^{n} \text{sim}(i,j)}{\text{nombre de liens dans le réseau}} ,$$

où $j \neq i$, i et j sont reliés ensemble, n est le nombre de nœuds dans le réseau et sim(i,j) est le score de similarité entre i et j. Il est à noter que la mesure de similarité sim(i,j) que nous utiliserons pour simuler notre modèle sera expliquée en détail au chapitre 3 (art. 3.4.2).

Une homophilie élevée signifie que beaucoup d'acteurs de profils d'intérêts similaires sont reliés ensemble dans le réseau (ils sont regroupés). Cette mesure servira à montrer, au chapitre 4, que les mécanismes de socialisation favorisent effectivement le rapprochement d'individus dont les intérêts sont proches.

Nous pouvons faire un parallèle entre notre mesure d'homophilie et la mesure de précision qui est souvent utilisée en recherche d'informations. La précision est une mesure de pertinence des résultats obtenus lors d'une recherche avec requête, qui calcule la proportion du nombre de documents pertinents rapportés sur le nombre total de documents rapportés. Dans notre contexte, la précision consisterait à mesurer la pertinence des regroupements formés au sein du réseau, à quel point les individus appartenant à une même communauté sont effectivement de profils similaires. En ce sens, notre mesure d'homophilie incorpore, en quelque sorte, la notion de précision.

2.3.1.11 Proportion des rencontres effectuées

Considérant qu'une rencontre a été effectuée entre tous les membres regroupés au sein d'une même communauté d'intérêt – les membres d'un groupe se connaissent – on peut calculer, pour chaque acteur *i*, la proportion du nombre de rencontres effectuées (parmi les individus de même profil que *i* dans le réseau) sur le nombre de rencontres qu'il aurait pu effectuer (nombre d'individus de même profil que *i* dans le réseau) :

$$R_i = \frac{\text{nombre de rencontres effectuées par } i}{\text{nombre de rencontres possibles pour } i} .$$

On peut ensuite calculer une moyenne globale des R_i, qui représente le taux moyen de rencontres effectué pour l'ensemble des acteurs dans le réseau. Nous utiliserons aussi cet indice pour évaluer l'efficacité des mécanismes de socialisation à provoquer des rencontres.

Cette mesure des rencontres effectuées s'apparente à la mesure de rappel qu'on utilise souvent en recherche d'informations pour évaluer l'efficacité d'un système à trouver les documents selon une requête effectuée. C'est le calcul de la proportion du nombre de documents

pertinents trouvés sur le nombre de documents pertinents existants (qui auraient pu être rapportés). Dans notre cas, la mesure des rencontres effectuées évalue, dans le même esprit, la proportion du nombre de rencontres pertinentes sur le nombre possible de rencontres pertinentes (qui auraient pu être effectuées).

Dans la section qui suit, nous voyons les propriétés structurales qu'on observe dans les réseaux sociaux existants. Au chapitre 4, nous comparerons ces propriétés à celles de nos réseaux générés pour discuter de la pertinence de notre modèle en tant que modèle d'évolution des réseaux sociaux.

2.3.2 Étude de la structure des réseaux complexes

L'objectif ultime de l'étude des propriétés structurales des réseaux complexes est de mieux comprendre et expliquer le fonctionnement des systèmes qui sont construits sur ces réseaux. On a observé, en effet, que ces réseaux ne se forment généralement pas de manière aléatoire, mais qu'ils présentent plutôt, selon les cas étudiés, des caractéristiques structurales particulières. Dans cette optique, de nombreux travaux se sont penchés sur la découverte de propriétés structurales typiques de divers réseaux : les réseaux biologiques comme les réseaux métaboliques (Fell et Wagner, 2000) et les réseaux neuronaux (White et al., 1986), les réseaux technologiques comme les réseaux électriques (Amaral et al., 2000) et Internet (Chen et al., 2002), les réseaux d'informations comme les réseaux de citations (Redner, 1998), le Web (Huberman, 2001) ou les réseaux sémantiques de mots (Motter et al., 2002) et, bien entendu, les réseaux sociaux comme les réseaux de collaboration (Barabási et al., 2002) et les réseaux de communication (Aiello et al., 2000 ; Ebel et al., 2002).

De même que l'on a amélioré les algorithmes de recherche d'informations dans le Web en étudiant sa topologie (Brin et Page, 1998 ; Page et al., 1998), on veut comprendre et ainsi pouvoir exploiter les propriétés structurales qui favorisent ou contraignent la circulation de l'information et des connaissances, au sein des réseaux sociaux. Nous parlerons cependant plus explicitement de la diffusion et la recherche d'informations aux sections 2.5 et 2.6.

Il existe plusieurs travaux qui se sont intéressés à la structure des réseaux. Nous en présentons ici quelques-uns qui se sont penchés plus particulièrement sur les réseaux sociaux et qui montrent la présence de caractéristiques structurales spécifiques dans plusieurs réseaux comme une distribution des degrés suivant une loi de puissance, un coefficient de *clustering* élevé et un petit diamètre. Au chapitre 4, nous reprendrons ces propriétés typiques des réseaux sociaux pour discuter des réseaux générés par simulation de notre modèle de socialisation.

2.3.2.1 L'effet des petits mondes

Une des propriétés notables qu'on observe dans les réseaux sociaux est que les individus sont relativement proches les uns des autres, que la distance moyenne dans les réseaux sociaux est petite. On appelle cette propriété l'effet des petits mondes (*small world effect*).

L'effet des petits mondes a été largement étudié. La première étude expérimentale conduite en ce sens fut celle de (Milgram, 1967) qui constata effectivement que le monde était petit, que la distance moyenne qui sépare les individus sur la planète (en terme de nombre de contacts interposés), se situe autour de six. Pour son expérience, Milgram demanda à plusieurs participants (au Nebraska) de faire parvenir une lettre à celui de leurs contacts personnels qu'il croyait être le plus proche d'un individu cible donné au Massachusetts (non connu des participants). Le contact ainsi choisi, qui recevait la lettre, devait faire de même et ainsi de suite jusqu'à ce que l'individu cible reçoive le message. Plusieurs des lettres se sont perdues au cours de l'expérience, mais environ le quart des messages atteignirent leur cible en passant par six contacts, en moyenne. Plus tard, cette idée fut reprise par (Guare, 1990) et donna lieu à l'expression bien connue des six degrés de séparation.

De nombreuses études expérimentales ultérieures prouvèrent aussi que si le nombre six n'est pas nécessairement le nombre juste, il n'en demeure pas moins que la chaîne de contacts qui relie entre eux deux individus, choisis aléatoirement sur la planète, est très petite par rapport à la population entière. Plus récemment, par exemple, (Dodds et al., 2003) ont refait cette expérience en utilisant des chaînes de courriels. Ils ont demandé à plus de 60 000 participants d'essayer de rejoindre un des 18 individus cibles, répartis dans 13 pays différents, et ils ont

trouvé que la médiane du nombre de contacts sur les chaînes des lettres qui sont parvenues à destination se situait entre 5 et 7.

On a montré la particularité des petits mondes dans plusieurs réseaux existants. Par exemple, dans l'une des premières études sur la structure du graphe du Web (Albert, Jeong, Barabasi, 1999), ont a calculé, sur un échantillon de 325 729 pages Web, que la distance moyenne entre les pages, lorsqu'on suit les hyperliens, était égale à 11.2. Par la suite (Broder et al., 2000) ont trouvé, sur un échantillon de 50 millions de pages, que la distance moyenne n'était que de 16.

Dans le cas particulier des réseaux sociaux, on observe aussi l'effet des petits mondes. Par exemple, (Watts et Strogatz, 1998) ont construit un réseau de collaboration entre acteurs à partir de la base de données IMDb - *Internet Movie Database* (http://us.imdb.com) - qui contenait 225 226 acteurs au moment de l'étude. Dans ce réseau, les acteurs sont reliés ensemble lorsqu'ils ont participé à un même film. On a montré que la distance moyenne entre les acteurs n'était que de 3.65.

Dans (Albert et Barabási, 2002 ; Newman, 2003), on retrouve plusieurs autres exemples, qui corroborent l'effet des petits mondes au sein des réseaux sociaux par l'étude de réseaux réels, construits à partir de données diverses comme la collaboration scientifique dans la publication d'articles, l'échange de courriels, les communications téléphoniques, etc.

D'autres études ont étudié plus particulièrement la relation entre la distance moyenne et la taille du réseau. On a trouvé que la distance moyenne entre les individus augmente effectivement très peu avec la taille des réseaux. Par exemple, l'étude conduite par (Newman , 2001b) sur différents réseaux de collaboration scientifique a montré que la distance moyenne augmente de manière logarithmique avec la taille des réseaux étudiés. D'autres études, comme celle de (Leskovec et al., 2007), ont même révélé un facteur de croissance négatif du diamètre en fonction de la taille pour divers réseaux étudiés, dont un réseau de communication par courriel, et un réseau de collaboration d'acteurs sur différents films.

Les communautés en ligne ne sont pas en reste quant à leur étude structurale. Au contraire, la montée en popularité de ces communautés constitue une mine d'or quant à l'obtention de données de réseaux à étudier et l'on voit de plus en plus de travaux portant sur l'analyse de la structure de ces grands réseaux en ligne.

On a montré que les communautés virtuelles en réseaux exhibent aussi l'effet des petits mondes. Par exemple, (Adamic et al., 2003) ont étudié le Club Nexus qui est un système de communication et de réseautage en ligne pour les étudiants à l'université de Stanford. Ceux-ci peuvent envoyer des courriels et des invitations, clavarder, publier des événements, vendre et acheter des biens usagés, rechercher des étudiants partageant leurs intérêts, et gérer leur liste de contacts. Le réseau obtenu en cartographiant le Club Nexus comprend 2469 utilisateurs reliés entre eux par 10119 liens. Les auteurs ont déterminé une distance moyenne d'environ 4 pour ce réseau en ligne. (Kumar et al., 2006) ont montré qu'en janvier 2006, le diamètre de la plus grosse composante connexe des réseaux sociaux virtuels Flickr (http://www.flickr.com), composés d'environ 1 million d'individus et de 8 millions de liens orientés, et Yahoo!360 (http://360.yahoo.com), comportant autour de 5 millions d'utilisateurs et 7 millions de liens dirigés, était de seulement 6.01 et 8.26, respectivement. Plusieurs autres études abondent en ce sens (Mislove, 2009 ; Ferrara et Fiumara, 2011 ; Nazir et al., 2008).

On comprend facilement que l'effet des petits mondes a une implication directe dans la diffusion de l'information au sein des réseaux sociaux : moins le nombre de pas nécessaires pour transmettre une information est grand, plus cette information se dispersera rapidement parmi les individus du réseau.

(Watts et Strogatz, 1998) ont défini qu'un réseau est petit (qu'il présente l'effet des petits mondes) si la distance moyenne ou le diamètre est comparable à celui d'un graphe aléatoire de taille similaire (en termes de nombre de sommets et de liens). Les mesures de distances des graphes aléatoires deviennent ainsi des points de références pour l'analyse de ce phénomène dans les réseaux sociaux.

2.3.2.2 Les amis de mes amis sont aussi mes amis

Une autre caractéristique structurale largement observée dans la majorité des réseaux sociaux est la transitivité (appelée aussi *clustering*). En effet, dans plusieurs réseaux sociaux, on remarque que lorsqu'un sommet *i* est connecté à un sommet *j* et que le sommet *j* est connecté à un sommet *k*, il y a de fortes chances que *i* et *k* soient aussi connectés ensemble. C'est la tendance de la formation de liens entre les individus qui ont des contacts en commun. Comme indicateur de cette propriété, on peut calculer un coefficient de *clustering* en mesurant la densité des triangles (triades complètement connectées) dans le réseau.

On observe que souvent, les réseaux complexes possèdent un coefficient de *clustering* plus élevé que celui qu'on obtiendrait pour un **graphe aléatoire** possédant le même nombre de sommets et de liens. Ce dernier est considéré comme le point de comparaison pour affirmer un fort ou faible taux de *clustering* dans les réseaux étudiés.

On peut calculer le coefficient de *clustering* $C_{aléa}$ d'un graphe aléatoire comme suit (Albert et Barabási, 2002) :

$$C_{aléa} = \frac{d_{moy}}{n}$$, où d_{moy} est le degré moyen et n, le nombre de sommets.

Lors d'une étude structurale, on peut donc comparer le coefficient de *clustering* obtenu lors d'analyses de données de réseaux réelles avec le coefficient de *clustering* théorique calculé pour un graphe aléatoire de même taille que le réseau observé. Par exemple, (Albert et Barabási, 2002) expliquent qu'on a trouvé que la topologie du réseau Internet possède un coefficient de *clustering* qui varie entre 0.18 et 0.3, ce qu'on peut considérer comme élevé en comparaison au coefficient de *clustering* d'un graphe aléatoire de paramètres similaires qui est d'environ 0.001. Pour le réseau du Web, (Adamic, 1999) a aussi rapporté un fort coefficient de *clustering,* égal à 0.1078, en comparaison avec son pendant aléatoire dont le coefficient de *clustering* serait égal à 3.2×10^{-4}.

Typiquement, les réseaux sociaux présentent aussi une transitivité de beaucoup supérieure à celle qu'on retrouve dans les graphes aléatoires. Le tableau 2.1 présente le coefficient de *clustering* réel C versus le coefficient de *clustering* dans le graphe aléatoire correspondant $C_{aléa}$, pour divers réseaux sociaux analysés. Cette information est tirée de (Albert et Barabási, 2002). Dans le tableau 2.1, le réseau d'acteurs de films est celui dont nous avons parlé à la section précédente. Les réseaux de co-auteurs sont des réseaux construits à l'aide de base de données électronique d'articles scientifiques où deux auteurs sont connectés ensemble s'ils sont co-auteurs d'un même article. MEDLINE est une base de données en recherche biomédicale, SPIRES est une base de données en physique et NCSTRL est une base de données en informatique.

Tableau 2.1
Coefficients de *clustering* observés dans divers réseaux sociaux réels

Réseau	C	$C_{aléa}$	Référence
Réseaux d'acteurs de films IMDb	0.79	$2.7\text{x}10^{-4}$	Watts and Strogatz, 1998
Réseaux de co-auteurs sur MEDLINE	0.066	$1.1\text{x}10^{-5}$	Newman, 2001b
Réseaux de co-auteurs sur SPIRES	0.726	0.003	Newman, 2001b
Réseaux de co-auteurs sur NCSTRL	0.496	$3\text{x}10^{-4}$	Newman, 2001b
Réseau de co-auteurs en mathématique	0.59	$5.4\text{x}10^{-5}$	Barabási *et al.*, 2002
Réseau de co-auteurs en neurosciences	0.76	$5.5\text{x}10^{-5}$	Barabási *et al.*, 2002

Du côté des sites de réseautage en ligne, on retrouve le même phénomène. Par exemple, on a rapporté que le coefficient de *clustering* du club Nexus (Adamic et al., 2003) est égal à 0.17, soit 40 fois plus grand qu'il le serait pour un graphe aléatoire de même taille en termes de nombre de sommets et de liens. Dans une étude sur la structure de Facebook, (Nazir et al., 2008) ont étudié le graphe des interactions entre les utilisateurs, via différentes applications : deux utilisateurs A et B interagissent lorsque A exécute une activité sur B (ou vice versa) ou bien lorsque A et B exécute une activité sur un ami commun C. Les trois graphes étudiés,

construits selon les interactions via trois applications différentes, présentent un coefficient de *clustering* très supérieur à celui qu'on obtiendrait pour des graphes aléatoires similaires : 0.81 versus 0.0062 pour l'application *Fighters' Club*, 0.31 versus 0.000016 pour l'application *Got Love* et 0.41 versus 0.000085 pour l'application *Hugged.* (Mislove, 2009) note aussi ce phénomène dans les sites de réseautage en ligne Flickr, LiveJournal, Orkut.

2.3.2.3 Distribution de degrés en loi de puissance

Les réseaux en loi de puissance sont des réseaux dans lesquels la probabilité *P(d)* qu'un nœud soit de degré d est $P(d) = cd^{-\alpha}$ où c est une constante et α, qu'on appelle souvent le coefficient de loi de puissance, est plus grand que 1. La figure 2.2 montre à quoi ressemble une telle distribution.

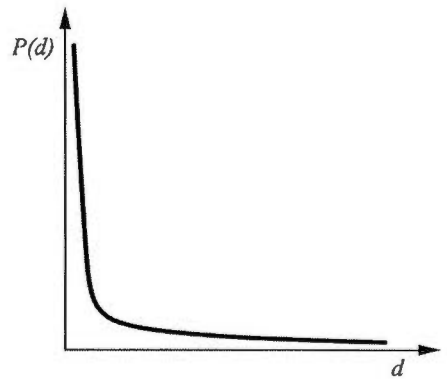


Figure 2.2 Distribution des degrés suivant une loi de puissance.

Dans la figure 2.2, on remarque une très forte probabilité d'avoir des sommets de degré faible qui décroît (très rapidement au début et doucement par la suite) à mesure que le degré augmente. En d'autres termes, une grande proportion des sommets sont de degré très faible tandis qu'une toute petite proportion de ceux-ci sont de degré très élevé. Une distribution en loi de puissance (avec sa longue queue / *long tail*) décroît plus graduellement qu'une distribution exponentielle, ce qui permet l'existence d'un petit nombre de sommets de degré très élevé.

Dans un graphe orienté, une telle distribution des demi-degrés intérieurs implique la présence d'une petite fraction de sommets qui reçoivent beaucoup de liens, tandis qu'une telle distribution sur les demi-degrés extérieurs indique que seulement une petite fraction des sommets propagent une grande quantité de liens.

La figure 2.3 illustre une représentation sous forme de graphe du réseau de contacts sexuels étudié par (Potterat et al., 2002) et dont la distribution des degrés suit une loi de puissance. On remarque clairement la présence d'un petit ratio de sommets qui possèdent beaucoup de liens par rapport au reste des sommets, qui n'en ont que très peu. Nous verrons que les réseaux construits par notre modèle de socialisation ont une forme graphique très similaire et nous allons montrer, au chapitre 4, que nos réseaux présentent aussi une distribution de degrés qui s'apparente fortement à une distribution en loi de puissance.

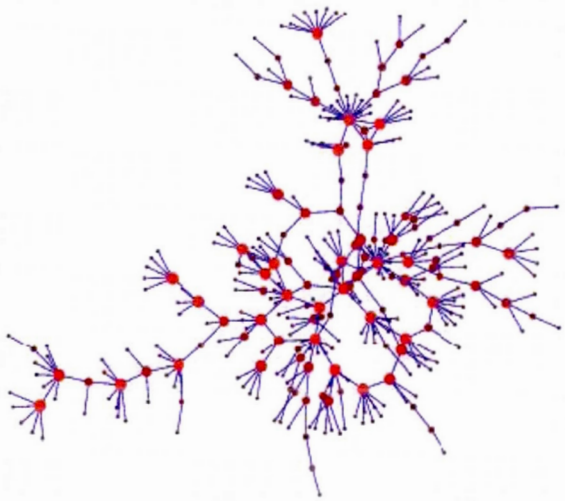


Figure 2.3 Graphe du réseau de contacts sexuels étudié par (Potterat et al., 2002).
(Tirée de Newman, 2003.)

Dans la littérature, on a montré qu'une grande diversité de réseaux complexes évolue de manière à produire une distribution des degrés en loi de puissance : la topologie d'Internet (Faloutsos et al., 1999), les réseaux neuronaux (Braitenberg et Schz, 1991), les réseaux linguistiques (Ferrer-i-Cancho et Solé, 2001), les réseaux écologiques (Solé et Montoya,

2001), etc. Le Web, aussi, est reconnu pour avoir une telle distribution des degrés, à la fois pour son demi-degré intérieur (nombre de liens entrants) et son demi-degré extérieur (nombre de liens sortants) (Barabási et Albert., 1999 ; Kumar et al., 1999).

Divers travaux ont montré que la structure de plusieurs réseaux sociaux, comme les réseaux de collaboration scientifique étudiés par (Newman 2001b ; Barabási et al., 2002) et les réseaux de partenaires sexuels investigués par (Liljeros et al., 2001 ; Potterat et al., 2002) ont une distribution des degrés particulière qui suit une loi de puissance. On a rapporté le même phénomène dans plusieurs communautés de réseaux en ligne. Par exemple, (Nazir et al., 2008), qui ont étudié le réseau des interactions (via des applications) entre les utilisateurs de Facebook, ont montré que la distribution des degrés du graphe obtenu, pour les trois applications étudiées, suit une loi de puissance. (Adamic et al., 2003) ont aussi observé cette distribution dans leur étude du site de réseautage en ligne Club Nexus de l'université de Stanford et d'autres encore, comme (Mislove, 2009), ont fait le même constat pour les sites de réseautage en ligne Flickr, LiveJournal et YouTube.

Cette distribution particulière sur les demi-degrés intérieurs, par exemple, a aussi des implications en recherche d'informations. Les sommets qui attirent (et donc, reçoivent) beaucoup de liens peuvent être vus comme des autorités (*authorities*), des sommets de réputation et de popularité élevées. Pour la recherche sur le Web, l'algorithme PageRank de Google (Brin et Page, 1998 ; Page et al., 1998) tire profit de la présence de ces autorités dans le graphe du Web pour mesurer l'importance relative des pages Web. Nous aborderons plus en détail l'influence de la structure des réseaux dans la recherche d'informations ou d'individus à la section 2.5. Il est à noter que ce ne sont pas tous les réseaux sociaux qui présentent une distribution des degrés en loi de puissance. Cette caractéristique est plutôt typique aux réseaux sociaux qui, d'abord, croissent continuellement et ensuite, dans lesquels les liens ne représentent pas des relations qui demandent beaucoup d'effort et d'investissement pour les entretenir, contrairement à nos relations avec les amis et la famille proches, par exemple. Dans ce dernier cas, l'effort demandé impose nécessairement une limite sur le nombre de relations sérieuses qu'un individu peut avoir et bien que cette limite

varie d'un individu à l'autre, elle proscrit la présence de super-connecteurs, possédant énormément de liens, typiques des réseaux en loi de puissance.

Les sites de réseautage en ligne, par contre, sont des lieux où il est possible de se créer un ensemble considérable (quasi illimité) d'amis ou contacts, moyennant un effort minimal et l'on retrouve ainsi très souvent cette propriété de distribution des degrés en loi de puissance lorsqu'on analyse ces communautés.

Pour les mêmes raisons, on observe aussi cette propriété lorsque le type des relations dans les réseaux étudiés est de nature sporadique, temporaire ou utilitaire, comme les réseaux de collaboration scientifique et les réseaux de contacts sexuels, mentionnés précédemment, qui ne demandent pas l'entretien continu de la relation et qui peuvent s'accroître continuellement.

2.3.2.4 Émergence de communautés structurales

La plupart des réseaux sociaux sont formés de sous-groupes plus cohésifs qu'on appelle communautés (Wasserman et Faust, 1994 ; Scott, 2000 ; Mislove et al., 2007). Ce sont des ensembles de sommets très fortement liés dans lesquels on observe une densité de liens beaucoup plus forte que la densité des liens qui relient ces communautés entre elles.

En effet, il est assez intuitif de comprendre que les individus se divisent en groupes selon leurs intérêts, leur culture, leur domaine professionnel, leur âge, etc. Ce phénomène d'homophilie est d'ailleurs souvent observé dans les réseaux sociaux (McPherson et al., 2001 ; Crandall et al., 2008 ; Lewis et al., 2008 ; Traud et al., 2012) et il n'est pas déraisonnable de penser que ce phénomène se reflète aussi au niveau structural.

Dans une étude récente (Ferrara et Fiumara, 2010), les auteurs ont mesuré la modularité Q de divers réseaux sociaux 1) de collaboration scientifique, construits à partir des articles archivés sur Arxiv (http://arxiv.org), 2) de communication, construit à partir de l'échange de courriels entre les membres de *Federal Energy Regulatory Commission*, 3) d'amis/contacts en ligne sur Facebook et YouTube 2007 et 4) de suffrages (qui vote pour qui) obtenu du système de vote de Wikipédia lors des élections des administrateurs de 2008. Le tableau 2.2 montre quelques-

uns des résultats qu'ils ont obtenus. On remarque que les valeurs de Q trouvées, bien que variant d'un réseau à l'autre, montrent des partitions significatives des réseaux en différentes communautés. On considère, en effet, qu'en pratique, une valeur supérieure à 0.3 semble indiquer de manière significative la formation de communautés dans un réseau (Newman, 2004 ; Newman 2004b).

Tableau 2.2
Valeurs de la modularité Q pour différents réseaux sociaux

Réseau	Q
Collaboration (Arxiv - astro physique)	0.628
Collaboration (Arxiv - physique, matière condensée)	0.731
Collaboration (Arxiv - relativité générale et cosmologie quantique	0.861
Communication (Échange de courriels)	0.615
Amis / Contacts (Facebook)	0.634
Amis / Contacts (YouTube 2007)	0.447
Suffrages (Élection Wikipédia 2008)	0.418

Au chapitre 4, nous utiliserons la mesure de modularité pour vérifier si les communautés formées automatiquement autour d'intérêts communs se reflètent bien au niveau structural, que les réseaux générés sont bien partitionnés en regard de la densité des liens à l'intérieur des groupes par rapport à la densité des liens entre les groupes et nous comparerons nos résultats à ceux du tableau 2.2.

Dans cette section, nous avons vu les propriétés structurales typiques qu'on remarque, de façon récurrente, dans les réseaux complexes et particulièrement dans les réseaux sociaux du monde réel. Ces observations indiquent clairement que nos réseaux sociaux ne se forment pas de manière aléatoire, mais semblent plutôt évoluer selon des mécanismes de structuration bien particuliers. Cette connaissance a donné lieu à la conception d'un bon nombre de modèles de réseaux, qui tentent d'expliquer la présence de ces propriétés structurales particulières au sein des réseaux.

2.4 Modélisation des réseaux

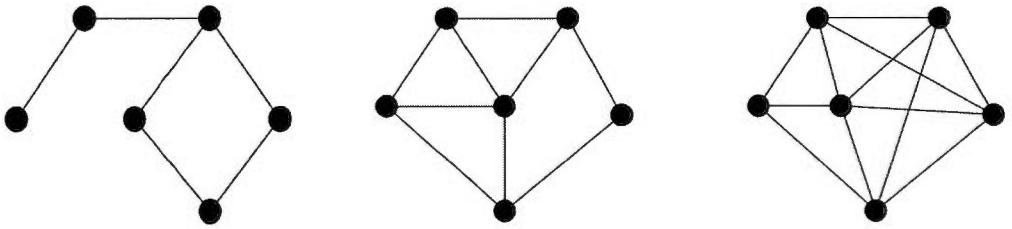
Notre modèle de socialisation propose les mécanismes de socialisation comme mécanismes d'évolution d'un réseau social. À notre connaissance, il n'existe aucun modèle basé sur ces mécanismes tels que nous les décrivons dans cette thèse.

Nous faisons ici la revue de modèles bien connus qui tentent d'expliquer les propriétés structurales comme l'effet des petits mondes, la présence d'une transitivité élevée et la distribution des degrés suivant une loi de puissance, observées dans plusieurs réseaux complexes. Les modèles discutés sont ou bien statiques, et proposent des mécanismes de construction de réseaux ayant des propriétés structurales reflétant celles observées dans la réalité ou bien dynamiques, expliquant la structure particulière des réseaux à l'aide de règles d'évolution.

2.4.1 Les graphes aléatoires

Un des premiers modèles proposés pour modéliser l'effet des petits mondes dans les réseaux, qu'on croyait de nature aléatoire, à l'époque, est le modèle des graphes aléatoires. Ce modèle a d'abord été introduit par (Solomonoff et Rapoport, 1951) et a été ensuite redécouvert et largement étudié par (Erdös et Rényi, 1960).

Le modèle des graphes aléatoires est extrêmement simple. C'est un réseau de taille n ayant un degré moyen z, appelé nombre de coordination du réseau. Ce réseau contient ainsi $\frac{1}{2}nz$ liens (bidirectionnels). On peut construire un tel graphe en traçant tout simplement n sommets et en reliant $\frac{1}{2}nz$ paires de sommets qu'on choisit au hasard. La figure 2.4 illustre différents graphes aléatoires de 6 sommets ayant des connectivités moyennes (z) différentes.



$n = 6, z = 2$ $n = 6, z = 3$ $n = 6, z = 4$

Figure 2.4 Divers graphes aléatoires de taille n et de degré moyen z.

On peut montrer facilement que ce type de réseau présente l'effet des petits mondes. Supposons i, un sommet d'un graphe aléatoire. i possède donc environ z voisins qui possèdent à leur tour z voisins et ainsi de suite. Il s'ensuit que i possède z voisins directs, z^2 voisins de distance 2, z^3 voisins de distance 3, etc.

Le diamètre d d'un graphe aléatoire est donc donné par $z^d = n$ qui implique que $d = \log n / \log z$. On voit que D augmente très lentement avec l'augmentation de la population n. Si l'on considère qu'un individu a en moyenne entre 100 et 1000 connaissances (disons 500) et que la population mondiale tourne autour de 7 milliards, si le monde était aléatoire, on serait séparé les uns des autres d'une distance d'environ $\log_{500} 7\ 000\ 000\ 000 = 3.65$. Clairement, les graphes aléatoires montrent l'effet des petits mondes !

On peut montrer que ces graphes aléatoires présentent une distribution suivant une loi de Poisson (Newman, 2003) où la plupart des sommets ont approximativement le même degré, dont la valeur est proche du degré moyen. Cette distribution n'est pas très réaliste comme modèle de réseaux réels. On a donc étendu les graphes aléatoires en introduisant la possibilité de spécifier d'avance la distribution des degrés voulue. On appelle cette généralisation du modèle des graphes aléatoires, modèle de configuration (*configuration model*). On construit ce type de réseau en spécifiant d'abord une distribution des degrés p_d où p_d est la proportion des sommets qui sont de degré égal à d. On détermine donc à l'avance le degré k_i de tous les sommets $[1..n]$ du réseau, puis on ajoute des connexions, de manière aléatoire, entre des paires de sommets, dont le degré (prédéterminé) n'est pas encore atteint. Le modèle de configuration a été largement étudié par plusieurs chercheurs dont (Bender et Canfield, 1978

; Molloy et Reed, 1995 ; Newman et al., 2001 ; Chung et Lu, 2002) et des extensions plus sophistiquées ont été proposées pour les graphes orientés et les graphes bipartis, par exemple. Nous référons le lecteur à (Newman, 2003) pour une description plus détaillée des divers modèles de graphes aléatoires.

Bien que le modèle des graphes aléatoires (et ses nombreuses extensions) montre parfaitement la propriété des petits mondes et qu'il puisse être adapté pour générer une distribution des degrés plus réaliste, celui-ci ne représente pas la propriété de transitivité (*clustering*) que l'on peut observer dans la plupart des réseaux sociaux : si i connaît j et k, la probabilité que j et k se connaissent est beaucoup plus élevée que la probabilité que deux individus, simplement choisis au hasard, se connaissent puisqu'ils ont un contact commun, i, par l'intermédiaire duquel ils risquent fortement de se rencontrer. Le modèle bien connu des petits mondes, présenté ci-dessous, tente de pallier ce manque.

2.4.2 Les petits mondes

Les petits mondes sont une classe de graphes qui montrent à la fois l'effet des petits mondes, qu'on retrouve dans les graphes aléatoires, et un degré de *clustering* élevé, typique des graphes fortement structurés comme les treillis réguliers.

Supposons un graphe représenté par un treillis régulier à une dimension, pouvant être circulaire, dans lequel chaque sommet est relié à ses z voisins les plus proches. La figure 2.5 montre un tel graphe avec $z = 4$. On note que plusieurs des voisins immédiats de chaque sommet sont aussi des voisins l'un de l'autre. Ce type de graphe présente donc la propriété de transitivité (*clustering*). Le coefficient de *clustering* C est la fraction moyenne des paires de voisins d'un nœud qui sont aussi des voisins l'un de l'autre. Dans un graphe complet, où tous les sommets sont reliés à tous les autres sommets, la transitivité est à son maximum avec $C = 1$. Dans un graphe aléatoire, $C = z / n$, ce qui est très petit pour un grand réseau (Albert et Barabási, 2002). Dans le cas d'un graphe similaire à la figure 2.5 (un treillis régulier), le

coefficient de transitivité est $C = \frac{3(z-2)}{4(z-1)}$ et tend vers $\frac{3}{4}$ pour des z très grands (Newman, 1999).

Un treillis régulier possède donc un coefficient de *clustering* très élevé, mais ne représente pas du tout l'effet des petits mondes, car la distance moyenne séparant deux sommets augmente de façon linéaire avec la taille du réseau.

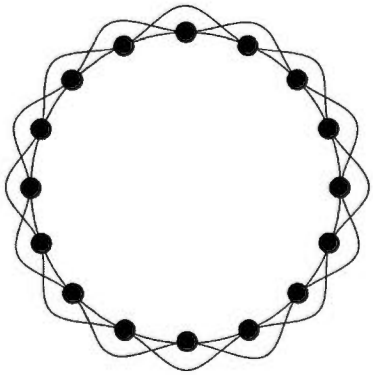


Figure 2.5 Treillis régulier avec un degré moyen $z = 4$.

(Watts et Strogatz, 1998 ; Watts, 1999) ont donc proposé un modèle se situant entre ces deux extrêmes, pour construire un graphe montrant à la fois l'effet des petits mondes (caractéristique des graphes aléatoires) et la propriété de transitivité (typique aux treillis réguliers). Le premier modèle, le modèle α, se construit comme le modèle d'Erdös et Rényi, mais lors de l'ajout de liens au graphe, plutôt que de choisir deux sommets de manière aléatoire, on les choisit avec une probabilité qui augmente proportionnellement au nombre de contacts communs que possèdent les deux sommets qu'on veut relier. Cela conduit rapidement à la formation de sous-groupes cohésifs (clusters) dans la structure tout en conservant un petit diamètre. Le deuxième modèle, le modèle β, part d'une structure en forme de treillis régulier circulaire, mais y ajoute un certain degré de hasard. Pour ce faire, parmi tous les liens du graphe de la figure 2.5, on choisit quelques liens selon une probabilité p et, pour chaque lien choisi, on déplace une des extrémités de ce lien vers un autre sommet

déterminé aléatoirement. La figure 2.6 illustre l'exemple d'un graphe obtenu par cet algorithme.

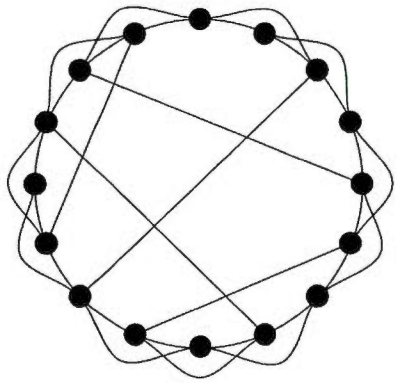


Figure 2.6 Modèle des petits mondes β de Watts et Strogatz.

La figure 2.6 montre que, pour une petite probabilité p, le treillis obtenu demeure quand même assez régulier et conserve un fort taux de *clustering*, mais possède désormais quelques raccourcis qui le traversent en produisant ainsi l'effet des petits mondes. En effet, Watts et Strogatz ont montré, par simulation numérique, que la distance moyenne entre deux sommets, lorsque la probabilité $p = ¼$, est égale à 3.6. Ce résultat est à peine supérieur à la distance moyenne égale à 3.2, calculée pour un graphe aléatoire de même taille. De plus, le nombre de voisins d'un sommet particulier peut être supérieur ou inférieur à z, mais en moyenne, il est égal à z. Clairement, la valeur du coefficient de transitivité C, pour de petites valeurs de p et un grand z, sera proche de celui calculé pour un treillis parfaitement régulier, comme celui de la figure 2.5. Ce modèle montre donc l'effet des petits mondes ainsi que la propriété de transitivité observée dans plusieurs réseaux sociaux réels.

L'effet des petits mondes, selon la vision de Watts et Strogatz, se produit grâce à quelques raccourcis aléatoires, passant à travers le treillis. (Dorogovtsev et Mendes, 2000a) ont trouvé une autre façon de provoquer ce phénomène. Il suffit d'ajouter au graphe quelques sommets qui ont un nombre de coordination (z) inhabituellement plus élevé que les autres sommets c'est-à-dire, qui sont reliés aléatoirement à un plus grand nombre de voisins que les autres sommets du graphe. Ces quelques sommets particuliers deviennent alors des points de

passage entre deux nœuds quelconques éloignés l'un de l'autre. Cet arrangement produit le phénomène des petits mondes tout en conservant la propriété de transitivité qu'offre la structure en treillis. La figure 2.7 montre à quoi ressemble un tel graphe.

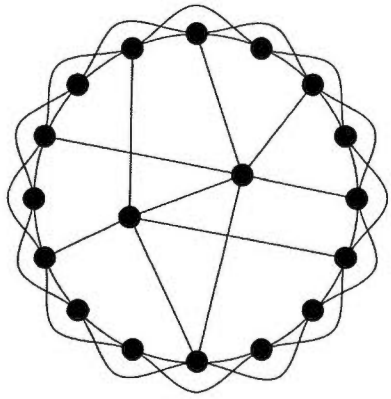


Figure 2.7 Modèle de petit monde de Dorogovtsev et Mendes.

Un autre modèle de petits mondes a été proposé par (Kleinberg, 2000b) dont l'argument en défaveur du modèle de Watts et Strogatz porte sur le fait qu'un tel modèle, basé sur des connexions arbitraires entre des paires de sommets et selon une probabilité uniforme, est une mauvaise représentation des situations qu'on retrouve dans le monde réel. Il propose alors un modèle basé sur un treillis carré à deux dimensions dans lequel on a ajouté aléatoirement quelques connexions entre des paires de sommets i,j, mais selon une probabilité non uniforme égale à d_{ij}^{-r} suivant une loi de puissance où r est le coefficient caractéristique qui détermine l'allure de la courbe. Cette probabilité est fonction de la distance d_{ij} entre i et j où $d_{ij} = |x_i - x_j| + |y_i - y_j|$ et où (x_i, y_i) et (x_j, y_j) sont les coordonnées des sommets i et j dans le treillis. Il semble possible, en effet, qu'un modèle qui considère le facteur de distance représente mieux les interactions sociales où la probabilité que deux individus se connaissent diminue avec la distance qui les sépare. D'autres auteurs ont aussi examiné cette possibilité où les liens relient préférentiellement des sommets qui sont proches les uns des autres dans la structure sous-jacente du treillis (Jespersen et Blumen, 2000 ; Moukarzel et de Menezes, 2002).

Les modèles de petits mondes, bien que représentant l'effet des petits mondes et la transitivité élevée observés dans les réseaux sociaux ne présentent cependant pas une distribution des degrés plausible dans la réalité.

Les modèles décrits jusqu'à présent tentent de créer des réseaux ayant des propriétés structurales similaires à celles observées dans les réseaux réels. Nous présentons maintenant des modèles de croissance des réseaux qui tentent d'expliquer comment les réseaux en viennent à posséder ces caractéristiques particulières, au cours de leur évolution. Typiquement, ces modèles proposent des mécanismes de croissance des réseaux (ajout de sommet et de liens au cours du temps) qui semblent bien refléter les mécanismes d'évolution des réseaux réels et qui produisent ainsi des structures semblables à celles observées dans la réalité.

2.4.3 Modèles basés sur l'attachement préférentiel

2.4.3.1 Le modèle à invariance d'échelle de Barabási et Albert et ses descendants

Pour expliquer le phénomène de la distribution des degrés suivant une loi de puissance, observée dans plusieurs réseaux complexes (dont certains réseaux sociaux et le Web), (Barabási et Albert, 1999) proposent le modèle à invariance d'échelle (*scale-free model*) que nous appellerons le modèle BA, pour références ultérieures.

Les auteurs argumentent que les modèles de réseaux existants (les graphes aléatoires et petits mondes) ne tenaient pas compte de deux caractéristiques importantes : 1) certains réseaux grandissent continuellement par l'addition de nouveaux sommets et de nouveaux liens et 2) les nouveaux sommets se connectent préférablement à des sommets qui ont une connectivité élevée. Ce second mécanisme est connu sous le nom d'attachement préférentiel. Les auteurs ont démontré que la combinaison de la croissance continuelle du réseau et de l'attachement préférentiel est directement responsable de la distribution de la connectivité en loi de puissance. Ce modèle dynamique est donc défini par deux règles :

Croissance : en débutant avec un petit nombre de sommets n_0 (et aucun lien), à chaque pas de temps, on ajoute un nouveau sommet qui possède n nouveaux liens (où $n \leq n_0$) qui seront connectés à n sommets déjà présents dans le graphe.

Attachement préférentiel : Chaque lien du nouveau sommet k est alors attaché à un sommet i existant dans le graphe avec une probabilité proportionnelle à son degré. Pour chaque lien de k, la probabilité $p(d_i)$ que le nouveau sommet k connecte un de ses liens au sommet i dépend du degré d_i de ce sommet tel que :

$$p(d_i) = \frac{d_i}{\sum_j d_j} .$$

Au bout de t pas de temps, on obtient un réseau contenant $t + n_0$ sommets et nt liens dont la distribution des degrés suit une loi de puissance $P(d) \sim d^{-2.9\pm0.1}$.

Ce modèle a été largement étudié et a donné lieu à plusieurs variantes dans l'optique de le rendre encore plus réaliste. Par exemple, (Albert et Barabási, 2000) proposent eux-mêmes une dynamique qui incorpore des reconnexions et des suppressions de liens dans le temps.

Aussi, on remarque que dans le modèle BA, les sommets plus anciens accumulent plus de liens par renforcement : plus un sommet a de liens, plus il en attire, plus il en attire, plus il a de liens. Les premiers sommets (les plus vieux) qui arrivent dans le réseau sont les premiers à recevoir de nouveaux liens avec les sommets subséquents et renforcent ainsi leurs chances d'en recevoir encore plus, et ainsi de suite. (Adamic et Huberman, 2000) ont montré, cependant, que le Web ne montrait pas cette corrélation entre l'âge et le degré comme dans le cas du modèle BA. Ils suggèrent que le grand degré de certaines pages Web s'explique plutôt parce que celles-ci ont une valeur intrinsèque plus grande, qu'elles sont plus utiles ou plus attrayantes que d'autres et attirent ainsi plus de liens. Pour modéliser ce phénomène, (Bianconi et Barabási, 2001) ont proposé une extension du modèle BA, le *fitness model*. Dans ce modèle, on assigne à chaque sommet un attribut dont la valeur représente leur niveau d'attrait et donc, leur propension à attirer de nouveaux liens. La probabilité d'attirer de nouveaux liens ne dépend plus seulement du degré, mais aussi de cette valeur d'attrait (*fitness value*). Une autre variante de (Dorogovtsev

et Mendes, 2000b) suggère l'idée que la tendance d'un site (dans le cas du Web) à attirer de nouveaux liens diminue selon son ancienneté (contrairement au modèle BA) : plus un site est âgé, moins il attire de nouveau liens.

Encore dans le cas du Web, on a observé que le degré moyen d'un sommet a tendance à augmenter avec le temps, que le paramètre n du modèle BA varie dans le temps. (Dorogovtsev et Mendes, 2001) ont alors proposé une autre variation du modèle BA qui incorpore ce processus. (Amaral et al., 2000), pour leur part, ont suggéré un coût ou une contrainte de capacité qui ralentit le processus de la distribution des degrés. Pour une revue plus exhaustive des descendants du modèle BA, nous référons le lecteur à (Albert et Barabási, 2002).

Il existe plusieurs autres modèles, se basant aussi sur l'attachement préférentiel, mais qui sont souvent plus spécifiques des réseaux étudiés, et donc plus difficilement comparables. Le modèle suivant en est un exemple.

2.4.3.2 Modèle d'évolution de réseaux sociaux en ligne de Kumar et al.

Les auteurs de ce modèle (Kumar et al., 2006) ont d'abord procédé à l'analyse structurale de deux grands réseaux sociaux en ligne : Flickr et Yahoo! 360. Ils ont remarqué que les deux réseaux étudiés pouvaient se partitionner en 3 types de composantes distinctes :

Singletons : ce sont les utilisateurs qui ont rejoint la communauté, mais qui n'ont jamais créé de relations avec d'autres utilisateurs. Ce sont des sommets de degré 0.

Composante géante : cette composante représente la communauté principale du réseau. Elle contient des utilisateurs qui sont typiquement les plus actifs dans le système, qui sont très fortement connectés entre eux.

Région intermédiaire : Cette région couvre le reste. Elle est composée de diverses communautés isolées qui sont des petits groupes d'utilisateurs connectés ensemble, mais non connectés à la composante principale du réseau. Cette région contient environ 33 % des

utilisateurs dans le cas de Flickr et autour de 10 % des utilisateurs dans le cas de Yahoo! 360. Les auteurs observent aussi que la majeure partie de ces petites communautés sont structurées en forme d'étoile : un seul utilisateur très populaire connecté à plusieurs autres utilisateurs qui ont eux-mêmes très peu de connexions. De plus, la distribution des degrés, pour les deux réseaux, suit aussi une loi de puissance.

Dans les réseaux étudiés, les utilisateurs peuvent rejoindre le réseau de deux manières : ils peuvent décider de s'inscrire eux-mêmes ou peuvent être invités par une de leurs connaissances. Les auteurs expliquent que les structures étoilées représentent typiquement les inscriptions par invitation. En effet, les individus qui en invitent d'autres sont plutôt motivés par l'envie de faire migrer leurs amis hors ligne vers la communauté en ligne que de construire de nouvelles relations parmi les utilisateurs qui sont déjà présents dans le réseau. Inversement, les membres de la composante principale font surtout du réseautage actif.

Modélisation des propriétés et mécanismes d'évolution observés

Pour modéliser les caractéristiques principales des réseaux étudiés, les auteurs introduisent une version biaisée du mécanisme d'attachement préférentiel qui, en plus de préférer les connexions à des sommets de degré élevé, celle-ci tient également compte de la plus grande facilité de trouver un contact potentiel (de créer une connexion) dans la composante géante que dans les communautés isolées de la région médiane.

Premièrement, ils définissent trois types d'utilisateurs : les passifs (*passive*), les connecteurs (*linkers*) et les inviteurs (*inviters*). Les passifs sont ceux qui rejoignent le réseau, pour une raison ou une autre, mais qui ne s'investissent dans aucune activité. Les inviteurs sont ceux qui souhaitent faire migrer en ligne, leur réseau social hors ligne. Les connecteurs sont ceux qui cherchent activement à parfaire leur réseau de contacts en créant de nouvelles connexions dans le réseau. Il est à noter, aussi, que ce modèle génère un graphe où les liens sont orientés. De manière informelle, le modèle se décrit comme suit :

À chaque pas de temps :

1. on ajoute un nouveau sommet au graphe. On détermine aléatoirement si ce nouveau sommet est un (P)assif, un (I)nviteur ou un (C)onnecteur selon une distribution de probabilité *p*. Cette distribution sur les types d'utilisateurs est prédéterminée (c'est un paramètre du modèle).

2. on ajoute ensuite *e* nouveaux liens au réseau et pour chaque lien :

    a. La source du lien est choisie au hasard parmi les I et les C présents dans le réseau en utilisant l'attachement préférentiel (normal).

    b. **Si la source est un I,** alors elle "invite" un non-membre à joindre le réseau et ainsi, la destination est un nouveau sommet qui devient un P.

    c. **Si la source est un C**, alors le sommet de destination est choisi parmi les I et C existants, en utilisant, cette fois, l'attachement préférentiel **biaisé**. – ceci reflète le fait que la région intermédiaire est plus difficile à découvrir que la composante principale lorsqu'on navigue dans le réseau.

Selon les auteurs, ce modèle reproduit fidèlement les composantes observées dans les deux réseaux étudiés, et ce, bien que les deux réseaux diffèrent d'un point de vue quantitatif.

Les modèles présentés précédemment nécessitent de connaître le degré de tous les sommets dans le réseau pour calculer la probabilité d'attachement à un sommet donné. En effet, cette probabilité est obtenue en calculant le rapport entre le degré de ce sommet et la somme de tous les degrés dans le réseau. Dans cette optique, on a donc proposé des modèles décentralisés qui ne demandent pas de connaissance sur la structure globale du réseau. Par exemple, le modèle suivant, basé sur des marches aléatoires dans le graphe, n'utilise que des règles locales pour former des réseaux dont la distribution des degrés suit une loi de puissance.

2.4.3.3 Le modèle basé sur les marches aléatoires de Vázquez

Le modèle de (Vázquez, 2003) s'inspire de la manière dont les utilisateurs naviguent sur le Web : un utilisateur peut commencer par sélectionner une page Web au hasard, parmi les

choix d'une collection de liens obtenue par l'intermédiaire d'un moteur de recherche, par exemple. Ensuite, il peut soit décider de revenir à la collection de pages initiales (ou faire une autre recherche) et choisir une autre page à visiter ou alors, il peut décider de suivre un des liens de la page sur laquelle il se trouve.

Pour modéliser ce phénomène de navigation sur le Web, l'auteur propose deux règles. Initialement, le réseau ne contient qu'un seul sommet, puis on effectue itérativement les deux règles suivantes :

1. **Ajouter** : on ajoute un nouveau sommet avec un lien (orienté) pointant vers un sommet déjà dans le réseau et choisi au hasard, puis on exécute la règle "Marcher".

2. **Marcher** : tant qu'un nouveau lien a été créé dans le réseau, ajouter, selon une probabilité q, un nouveau lien qui part du sommet courant (celui qui vient de recevoir un lien entrant) et arrive sur l'un de ces voisins, choisi au hasard. Lorsqu'aucun lien n'est créé, retourner exécuter la règle "Ajouter".

La règle "Ajouter" modélise le fait de choisir une nouvelle page Web à consulter et la règle "Marcher" représente l'exploration du réseau, de page en page, en suivant les hyperliens.

L'auteur du modèle a montré que le graphe résultant suit une loi de puissance (sur les demi-degrés intérieurs) dont le coefficient est égal à 2 pour une probabilité q proche de 1.

Le mécanisme des marches aléatoires produit celui de l'attachement préférentiel dans une dynamique d'auto-renforcement (l'attachement préférentiel est ici un effet émergent de la dynamique du réseau). En effet, les sommets qui ont un demi-degré intérieur supérieur ont plus de chances de se trouver sur les chemins parcourus lors des marches aléatoires dans le graphe et ainsi, reçoivent constamment plus de liens, ce qui favorise encore plus leurs chances de se trouver sur des chemins subséquents et ainsi de suite. Pour les graphes non orientés, (Saramäki et Kaski, 2004) ont proposé un modèle similaire, basé aussi sur les marches aléatoires.

2.4.4 Modèles basés sur la fermeture des triangles

Les modèles dynamiques basés sur l'attachement préférentiel représentent bien la distribution des degrés en loi de puissance et l'effet des petits mondes (petit diamètre) observés dans plusieurs réseaux complexes, cependant, plusieurs d'entre eux ne traitent pas le phénomène de transitivité élevé qu'on observe aussi dans la majorité des réseaux sociaux.

Les modèles présentés ici exploitent spécifiquement le processus de fermeture des triangles pour créer un certain degré de *clustering* dans les réseaux qu'ils génèrent : lorsqu'on ajoute des liens au réseau, on le fait de manière préférentielle entre deux sommets qui ont des voisins en commun (on ferme le triangle).

2.4.4.1 Modèle de réseaux de contacts personnels de Jin, Girvan et Newman

(Jin, Girvan et Newman, 2001) proposent un modèle représentant l'évolution d'un réseau social de contacts personnels (ou d'amis). Les propriétés de ce modèle diffèrent des règles du modèle BA (*scale-free* model) pour plusieurs raisons.

Premièrement, les auteurs considèrent qu'étant donné le va-et-vient continuel dans les réseaux sociaux (les individus voyagent et changent de réseau, ils font de nouvelles connaissances, certains meurent, d'autres naissent, etc.), il est raisonnable de penser qu'au bout du compte, le nombre de sommets dans le graphe du réseau demeure assez constant. Seuls le nombre et la structure des liens varient.

Deuxièmement, selon les auteurs, la distribution des degrés, dans un réseau social, ne suit pas nécessairement une loi de puissance, et semble souvent tourner autour d'un degré moyen. Ils expliquent ce phénomène en soulignant que l'entretien d'une amitié demande des ressources, en temps et en effort, et qu'il y a forcément une limite au nombre d'amis qu'un individu peut avoir. L'absence d'une distribution des degrés suivant une loi de puissance suggère que le mécanisme de l'attachement préférentiel n'est pas très important dans les réseaux d'amis.

Enfin, comme on l'a vu précédemment, les réseaux sociaux doivent tenir compte de la propriété de transitivité qu'on ne retrouve pas dans le modèle BA et ses descendants.

Pour modéliser les propriétés ci-dessus énumérées, les auteurs proposent les règles et propriétés suivantes :

Un nombre fixe de sommets : Considération d'une population fermée d'une grandeur fixe.

Un degré limité : La probabilité qu'un individu développe une nouvelle amitié doit diminuer fortement lorsque le nombre de ses amis atteint un certain niveau. On pose donc le paramètre z^* qui représente la limite sur le nombre z d'amis qu'un individu peut avoir.

Lorsque le degré d'un sommet du graphe atteint la valeur z^*, la probabilité que ce sommet reçoive d'autres liens (qu'un individu se fasse de nouveaux amis) diminue rapidement.

Transitivité : Pour modéliser la transitivité par le processus de fermeture des triangles, la probabilité que deux individus deviennent des amis doit être significativement plus élevée s'ils ont un ou plusieurs amis communs : des paires d'individus se rencontrent avec une probabilité par unité de temps qui dépend du nombre d'amis qu'ils ont en commun.

La probabilité par unité de temps p_{ij} de former un lien entre les sommets i et j dépend de deux facteurs :

1. Le nombre d'amis z_i et z_j que i et j possèdent déjà.
2. Le nombre m_{ij} d'amis communs de i et j.

$$p_{ij} = f(z_i)f(z_j)g(m_{ij}) \ .$$

La fonction f doit être assez grande et assez constante pour de petits z, mais doit diminuer considérablement lorsque z approche de la valeur de transition z^*. La fonction de Fermi présente ces caractéristiques et c'est elle que les auteurs ont utilisée pour leur modèle.

$f(z_i) = \frac{1}{e^{\beta(z_i - z^*)} + 1}$, où le paramètre β contrôle la précision de la diminution à z^*.

La fonction *g(m)* représente l'augmentation attendue de la chance que deux individus se rencontrent s'ils ont un ou plusieurs amis communs.

$$g(m) = 1 - (1 - p_0)e^{-\alpha m} ,$$

où p_0 représente la probabilité d'une rencontre entre deux individus sans amis communs et où le paramètre α contrôle la vitesse à laquelle $g(m)$ augmente.

Cessation d'une amitié : Pour éviter la stagnation du réseau, il est nécessaire que des liens se brisent (et que d'autres se forment), étant donné le nombre fixe de sommets degré limité pour chaque sommet. Deux amis doivent se rencontrer régulièrement pour maintenir leur amitié. Dans le cas contraire, la relation s'atténue pour finalement disparaître.

On représente donc la force d'une relation avec un attribut *s*, pour chaque lien. La valeur de *s* sera plus ou moins élevée selon que les rencontres seront plus ou moins fréquentes, au cours de l'évolution du réseau. Chaque fois que deux amis se rencontrent, la force s_{ij} est ajustée à la valeur 1. Puis, à mesure que le temps passe et qu'ils ne se rencontrent pas de nouveau, s diminue exponentiellement selon :

$$s_{ij} = e^{-k\Delta t} ,$$

où Δt représente l'intervalle de temps depuis la dernière rencontre de *i* et *j* et où *k* est un paramètre ajustable du modèle. Si *i* et *j* se rencontre une autre fois, s_{ij} est remis à 1. On peut ainsi poser une limite minimale à s_{ij} en dessous de laquelle on considère que la relation se termine (le lien entre *i* et *j* est supprimé).

La figure 2.8 illustre un graphe généré par ce modèle. On note de façon très évidente la formation de triangles dans le réseau. Pour des valeurs de paramètres spécifiques, les auteurs du modèle obtiennent un coefficient de *clustering* égal à 0.45, très élevé par rapport à 0.02 pour un graphe aléatoire de même taille en termes de nombre de sommets et de liens. On y voit aussi la formation de communautés distinctes, comme observées dans nos réseaux sociaux usuels.

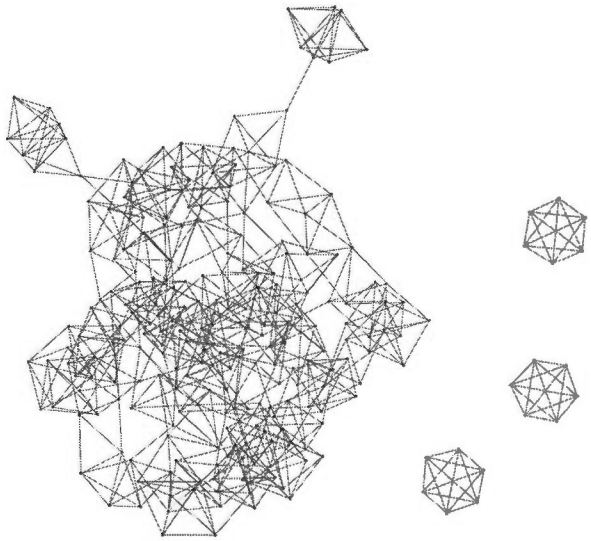


Figure 2.8 Illustration du graphe généré par le modèle de Jin, Girvan et Newman.
(Tirée de Lord, 2007.)

Le modèle de Jin, Girvan et Newman est plausible pour représenter l'évolution de communautés plutôt fermées, c.-à-d. qui ne croissent pas rapidement avec le temps comme c'est le cas des réseaux sociaux virtuels ou du Web, par exemple.

Le prochain modèle est une autre généralisation du modèle BA (voir art. 2.4.3) qui produit des réseaux dont la distribution des degrés suit une loi de puissance, mais auquel on a intégré le mécanisme de fermeture des triangles pour ajouter l'effet de *clustering* observé aussi dans la plupart des grands réseaux qui croissent continuellement.

2.4.4.2 Le modèle transitif à invariance d'échelle de Holme et Kim (clustered scale-free model)

Pour incorporer la transitivité dans les réseaux générés, qui fait défaut au modèle BA, (Holme et Kim, 2002) propose une version du modèle à invariance d'échelle avec *clustering* en ajoutant une étape supplémentaire de fermeture des triangles au modèle BA original.

Le modèle de Holme et Kim comprend donc les mêmes étapes de croissance et d'attachement préférentiel du modèle BA, responsables de la distribution en loi de puissance, plus une étape supplémentaire, qu'ils nomment **formation de triades**. Les trois règles du modèle sont alors :

Croissance : en débutant avec un petit nombre de sommets n_0 (et aucun lien), à chaque pas de temps, on ajoute un nouveau sommet qui possède n nouveaux liens (où $n \leq n_0$) qui seront connectés à n sommets déjà présents dans le graphe.

Attachement préférentiel : Chaque lien du nouveau sommet j est alors attaché à un sommet i existant dans le graphe avec une probabilité proportionnelle à son degré. Pour chaque lien de j, la probabilité $p(d_i)$ que le nouveau sommet j connecte un de ses liens au sommet i dépend du degré d_i de ce sommet tel que :

$$p(d_i) = \frac{d_i}{\sum_j d_j} .$$

Formation de triades : Si un lien a été ajouté entre j et i à l'étape d'attachement préférentiel précédente, on ajoute alors un lien de plus entre j et un voisin immédiat de i choisi aléatoirement (fermeture des triangles), selon une probabilité P_t. S'il ne reste plus de paires à connecter (si tous les voisins immédiats de i sont déjà connectés à j), refaire une autre étape d'attachement préférentiel selon une probabilité (P_t - 1).

Le paramètre du modèle m_t , qui représente le nombre moyen d'exécutions de l'étape de formation des triades (pour chaque ajout de sommet) est donné par :

$$m_t = (m-1)P_t \ .$$

C'est ce paramètre qui permet de contrôler le niveau de *clustering*. Le modèle se conduit donc exactement comme le modèle BA lorsque $m_t = 0$.

Peu importe la valeur de m_t, la distribution en loi de puissance résultante est toujours la même que celle du modèle BA soit $P(d) \sim d^{-3}$. Le coefficient de *clustering* peut être paramétré (avec le paramètre m_t) selon le niveau désiré. Par exemple, pour un nombre de sommets qui tend vers l'infini, $m_t = 0.6$ génère un réseau ayant un coefficient de *clustering* autour de 0.15, pour $m_t = 1.8$, le coefficient de *clustering* est de 0.5. On remarque aussi que la distance moyenne augmente de manière logarithmique avec le nombre de sommets, ce qui est cohérent avec le comportement du modèle de petits mondes de Watz et Strogatz (ainsi que le modèle BA). Le réseau se qualifie donc en tant que petit monde.

Pour expliquer la formation d'une distribution des degrés en loi de puissance ainsi qu'un degré de *clustering* élevé, les auteurs suggèrent que non seulement, en contexte social, un nouvel arrivant aura tendance à se lier avec un acteur populaire, mais de plus, une fois cette relation établie, il aura de bonnes chances de rencontrer aussi les amis de celui-ci.

Dans la littérature, on retrouve plusieurs autres auteurs ayant proposé des modèles basés sur le mécanisme de fermeture des triangles dont (Watts, 1999 ; Davidsen et al., 2002 ; Klemm et al., 2002 ; Vázquez et al., 2003).

Dans la section suivante, nous montrons comment les propriétés structurales des réseaux influencent leur navigabilité lors de recherches d'informations ou d'individus.

2.5 Navigation et recherche dans les réseaux d'information

Comme l'ont montré (Milgram, 1967 ; Dodds et al., 2003), il existe en effet des chemins de très courte distance, dans les réseaux sociaux, qui relient des individus à première vue très éloignés les uns des autres. En général, les individus connaissent leurs amis et possiblement

quelques amis de leurs amis, mais pas beaucoup plus loin. Pourtant, les lettres de Milgram se sont rendues à destination en quelques pas seulement. Cela laisse supposer que les structures relationnelles ont quelque chose de particulier qui fait que les individus, en ne possédant que de l'information locale (les amis et les amis des amis), arrivent tout de même à trouver ces courts chemins assez aisément.

Dans cette optique, des chercheurs se sont penchés sur le problème de navigation dans les réseaux décentralisés soit en tentant de tirer parti des propriétés structurales connues ou bien en essayant de déterminer les structures qui fonctionnent le mieux, dans le but d'améliorer la recherche dans les réseaux d'informations.

Supposons qu'une ressource d'intérêt soit stockée dans un sommet quelconque d'un réseau décentralisé : une information sur une page Web, un fichier stocké dans une base de données distribuée ou sur un ordinateur distant dans un réseau pair-à-pair, etc. Le problème est de pouvoir localiser rapidement cette information, et ce, à partir d'information locale seulement.

Un moyen direct d'accomplir cette tâche est de parcourir tous les liens du réseau de façon exhaustive (*crawling*) pour cataloguer le réseau et construire un index local des données trouvées. C'est la méthode utilisée par les moteurs de recherche classiques.

Avec le moteur de recherche Google, (Brin et Page, 1998 ; Page et al., 1998) ont été les premiers à implémenter un algorithme de recherche efficace (*PageRank*) qui utilise la structure du réseau pour trier, selon leur pertinence, les données obtenues à l'étape du *crawling*. En effet, en sachant que le Web montre une distribution des degrés en loi de puissance qui suppose la présence d'une petite portion de sommets de degré très élevé, on peut profiter de cette propriété structurale. Leur stratégie consiste donc à utiliser l'information non seulement présente dans le contenu des pages Web, mais aussi dans les liens (hyperliens) entre les pages, pour ordonner les résultats selon leur pertinence. Cela suppose évidemment une indexation préalable non seulement des pages, mais aussi des liens. L'idée générale de cet algorithme est de tenir compte du degré des noeuds dans le réseau. Pour évaluer la pertinence des pages, on fait l'hypothèse 1) qu'une page est importante ou

populaire si beaucoup d'autres pages pointent vers elle (nombre de liens reçus = demi-degré intérieur) et 2) encore plus importante si les pages qui pointent vers elle sont aussi des pages populaires.

Des raffinements successifs de cet algorithme ont été proposés, par exemple, par (Kleinberg, 1999) qui, en plus de tenir compte des pages populaires (*authorities*) étant réputées contenir de l'information pertinente sur un sujet donné, il considère aussi les pages dont le demi-degré extérieur est élevé (*hubs*). Ces dernières qui, bien que ne contenant pas nécessairement d'informations pertinentes, sont cependant une source importante de références (liens projetés) vers d'autres pages possiblement pertinentes, sur un sujet donné. Ces types de méthodes de classification par pertinence peuvent aussi être étendues à d'autres types de réseaux d'informations. Par exemple, le moteur de recherche de la librairie numérique *CiteSeer* utilise une méthode de classification d'articles semblable pour l'ordonnancement des publications pertinentes dans le réseau des citations (Giles et al., 1998).

Plutôt que de faire un catalogage exhaustif (*crawling*) du réseau (ou d'une portion du réseau), d'autres méthodes de recherche exploitent la structure du réseau pour choisir plus efficacement la portion à explorer ou les liens à suivre lors du catalogage. Par exemple, la stratégie du *Web crawler* de (Menczer et Belew, 2000) est fondée sur l'observation que les pages (ou sommets) qui contiennent de l'information sur un sujet particulier tendent à être regroupées ensemble (*clusters*) dans des régions locales du graphe. Ils basent ainsi leur algorithme sur la connaissance de cette propriété structurale. Ils utilisent donc un type d'algorithme génétique en déployant un bataillon d'agents pour parcourir les liens du Web au hasard, recherchant les pages qui contiennent, par exemple, les mots ou des sous-ensembles de mots donnés dans la requête de l'utilisateur. Plus un agent trouve de pages qui correspondent à la requête, plus son score de performance augmente. Les agents les plus performants sont alors dupliqués de telle sorte que la densité des agents devient graduellement plus élevée dans les régions prometteuses et les agents les moins performants sont tués.

Dans le cas d'un réseau pair-à-pair, disons un système de partage de fichiers décentralisé, par exemple, une méthode de recherche naïve serait de propager la requête à tous les utilisateurs dans le réseau : l'utilisateur qui recherche un document envoie d'abord une requête à tous ses voisins. Lorsqu'un voisin n'a pas le document demandé, il propage la requête vers ses propres voisins et ainsi de suite jusqu'à ce que la ressource recherchée soit trouvée ou bien que tous les noeuds du réseau aient été interrogés sans succès. Dans le pire cas, la requête se propage à travers tout le réseau. (Adamic et al., 2001) ont proposé une amélioration de cet algorithme en raffinant le processus de propagation de la requête. Dans leur version, la requête n'est plus transmise à tous les voisins, mais plutôt au voisin le plus populaire (qui a lui-même le plus de voisins). Cette stratégie fait en sorte que la requête se propage le long d'une chaîne de pairs de degré de plus en plus élevé jusqu'à ce qu'elle atteigne le pair le plus populaire. Elle suppose aussi des techniques de marche arrière pour éviter les culs-de-sac. On peut comprendre que ce protocole de recherche est particulièrement efficace dans les réseaux à invariance d'échelle (scale free) puisque ceux-ci ne contiennent qu'une petite proportion de noeuds de degré très élevé (hubs).

Supposons maintenant qu'on veuille trouver un noeud cible (on connaît la cible) à partir d'un noeud choisi de façon aléatoire dans le réseau. (Kleinberg, 2000) a montré qu'avec un modèle spécifique de réseau petit monde et sous des conditions très particulières, ce modèle permet de trouver le noeud cible de façon efficace en n'utilisant qu'un algorithme glouton. L'algorithme glouton, ici, est une stratégie locale simple qui consiste à choisir le prochain noeud (parmi les voisins du nœud courant) qui nous rapproche le plus de la cible. Pour concevoir son modèle petit monde, Kleinberg a utilisé un treillis géométrique à deux dimensions (n x n) pour disposer les noeuds. Pour tenir compte de la forte structuration locale typique au modèle petit monde, chaque noeud u est relié à ses quatre voisins immédiats (nord, sud, est et ouest). La particularité du modèle vient du fait que, pour tenir compte de l'effet des petits mondes, l'ajout de connexions de longue distance entre un noeud u et un noeud v choisi au hasard, se fait selon une probabilité qui suit une loi de puissance $P(d) = d^{-\alpha}$ où d est la distance de treillis (distance de Manhattan) entre u et v et où $\alpha \geq 0$ est l'exposant de *clustering* qui est un paramètre fixe du modèle. En utilisant un algorithme glouton (décentralisé) pour parcourir le réseau, si l'on considère le temps de parcours en

terme de nombre de noeuds à traverser pour atteindre la cible, Kleinberg montre que, pour $\alpha = 2$, le temps de parcours est borné par n'importe quel polynôme en log N (où N est le nombre de sommets dans le graphe). Cependant, pour toute autre valeur de α, le temps de parcours devient asymptotiquement beaucoup plus grand. De façon générale, pour un treillis à m dimensions, Kleinberg montre que la valeur critique est $\alpha = m$. Ces résultats montrent bien qu'étant donné les propriétés structurales d'un réseau, la navigation ou la recherche dans le réseau s'effectue plus ou moins efficacement et qu'une navigation efficace n'est pas une propriété de tous les réseaux petits mondes.

(Watts et al., 2002) et (Kleinberg, 2001b) ont abordé le même problème de recherche d'un individu cible à partir d'un individu choisi au hasard dans un réseau petit monde, mais en proposant un modèle plus plausible de la société et en utilisant plutôt une mesure de distance sociale qui tient compte de l'identité des individus (ex. : leur emploi, leur lieu de résidence, leur langue, etc.). L'algorithme glouton pour choisir le prochain individu à visiter consiste à sélectionner l'individu qui est le plus proche (en terme de distance sociale) de la cible. On a montré que cet algorithme est performant pour un large éventail de paramètres du modèle et qu'avec des paramètres appropriés, la recherche pouvait s'effectuer en temps $O(\log N)$, encore une fois. Les deux modèles présentés par Watts et al. et Kleinberg ont montré que la probabilité de formation des relations entre les individus présents dans le réseau doit être corrélée à la proximité (en termes d'attributs similaires) entre les individus pour qu'une simple stratégie gloutonne puisse fonctionner.

Ces quelques exemples sur les stratégies de navigation dans les réseaux montrent bien comment leurs propriétés structurales influencent leur navigabilité. Nous verrons, au chapitre 3, comment les caractéristiques topologiques que nous imposerons aux réseaux générés par notre modèle de socialisation nous permettront d'effectuer des parcours plus efficaces dans le réseau, lors de la recherche d'individus de profils similaires (afin de les regrouper en communautés d'intérêts).

La section suivante présente des travaux effectués sur la diffusion de l'information au sein de divers types de réseaux et montre comment les caractéristiques structurales contraignent ou favorisent la dissémination de l'information dans ces réseaux.

2.6 Diffusion de l'information dans les réseaux sociaux

Une des raisons principales d'étudier les réseaux est de comprendre les mécanismes de diffusion qui influencent la propagation des maladies, des virus informatiques, des rumeurs, de l'information, etc.

La théorie des réseaux sociaux a donc donné lieu à plusieurs études qui démontrent que les différentes propriétés structurales des réseaux contraignent ou favorisent la manière et la vitesse avec lesquelles l'information peut se diffuser à travers eux. Pour n'en nommer que quelques-unes, mentionnons, par exemple, les travaux en sociologie de (Granovetter, 1973) qui ont montré que les liens de longue distance (les contacts qui forment des ponts entre les petits groupes) facilitent la diffusion globale de l'information. (Friedkin, 1990), quant à lui, a investigué l'effet de la structure des réseaux de communication interpersonnelle sur la transmission des influences. Parmi les études sur la diffusion des innovations, (Mason et al., 2008) ont montré expérimentalement, entre autres, que les réseaux complets (complètement connectés) favorisent l'imitation à cause de leur forte structuration spatiale tandis que les réseaux petits mondes modèrent l'imitation au profit d'une certaine exploration d'autres solutions (grâce à leurs connexions de longue distance). Ces conclusions sont venues corroborer les résultats d'une expérience antérieure similaire effectuée par (Lazer et Friedman, 2005) à l'aide d'un modèle basé agents.

Dans une étude par modélisation sur la diffusion des rumeurs, (Nekovee et al., 2007) ont trouvé, par exemple, que le taux de diffusion initial d'une rumeur est beaucoup plus élevé dans les réseaux à invariance d'échelle que dans les réseaux aléatoires à cause des super connecteurs (*hubs*). Aussi, les travaux de (Zanette, 2002) ont montré que, dans les réseaux petits mondes, en variant le degré de hasard (et donc de structure spatiale), à un certain moment donné, le modèle opère une transition critique en passant d'un régime dans lequel la

rumeur s'éteint dans le voisinage immédiat de son origine vers un régime dans lequel elle se diffuse dans une partie de la population entière.

Encore une fois, on voit que les propriétés structurales des réseaux influencent grandement la manière dont l'information y circule. La diffusion de l'information n'est pas directement abordée dans cette thèse, mais on pourrait très bien imaginer une composante à notre modèle de socialisation qui automatise aussi la diffusion de l'information au sein des participants. Nous reparlerons brièvement de cette possibilité au chapitre 5 (art. 5.2.2).

2.7 Systèmes sociaux et collaboratifs

Les nouvelles technologies ont un effet notable sur la manière dont nous collaborons aujourd'hui. Nous sommes plus connectés que jamais, les uns avec les autres puis avec le reste, par l'intermédiaire d'ordinateurs, de téléphones cellulaires et autres appareils mobiles. On assiste au foisonnement de communautés virtuelles et d'espaces de collaboration et d'interactions "dans les nuages" (enseignement à distance, laboratoire à distance, wikis, blogues, etc.) qui rendent possible l'interaction en temps réels (synchrone ou asynchrone) de plusieurs individus décentralisés. La collaboration en temps réel n'est plus soumise aux frontières du temps et de l'espace.

On constate dès lors l'émergence récente de plusieurs outils et systèmes dits "sociaux" ou "collaboratifs" qui exploitent ces communautés virtuelles pour améliorer la recherche et la diffusion de l'information. En effet, plusieurs travaux soulignent l'importance de considérer la nature sociale de la gestion de la connaissance pour la conception de meilleurs outils d'accès à l'information basés sur la collaboration et les relations sociales. Nous faisons ici un bref parcours des idées principales sous-jacentes aux principes de l'informatique sociale et collaborative. En effet, le travail proposé dans cette thèse, basé principalement sur l'idée de socialisation, s'inscrit dans cette lignée de paradigmes collaboratifs qui tentent de tirer profit de l'expérience et des connaissances des uns pour enrichir et personnaliser l'expérience et les connaissances des autres.

En effet, la socialisation peut être perçue comme paradigme collaboratif puisqu'elle est basée sur l'échange. En échangeant avec des individus cognitifs, on ne récupère pas simplement de l'information à l'état brut, mais de l'information augmentée de l'expérience des individus, de l'information déjà assimilée, élaguée, contextualisée, analysée, filtrée, etc. On profite donc de ce traitement cognitif fait par les autres lorsqu'on socialise. De plus, par l'intermédiaire de mécanismes de socialisation comme les rencontres et les recommandations, les contacts des uns viennent enrichir le réseau de contacts des autres.

Nous en profiterons aussi pour aborder quelques techniques bien connues d'extraction de profils d'intérêt des utilisateurs, qui sont utilisées dans les systèmes sociaux et collaboratifs. Dans cette thèse, nous n'abordons pas le problème de création de profils, mais nous aurons à choisir une méthode pour simuler notre modèle de socialisation, que nous verrons plus en détail au chapitre 3 (art. 3.4.2).

2.7.1 La navigation sociale et la recherche collaborative

L'idée générale de la navigation sociale, introduite par (Dourish et Chalmers, 1994), consiste à se diriger vers des regroupements sociaux ou des objets particuliers parce qu'ils ont été examinés par d'autres avant nous. Dans le contexte d'un environnement d'hyperliens comme le Web, elle consiste à récupérer les traces électroniques (les comportements ou l'historique de navigation) laissées par les utilisateurs précédents afin de guider les utilisateurs subséquents. Un peu comme lorsqu'on demande conseil à l'un de nos amis sur l'achat d'un ordinateur parce qu'on sait qu'il vient justement de s'en procurer un et qu'il en a magasiné plusieurs (a visité plusieurs magasins). Dans cette optique, le système *Footprints* de (Wexelblat et Maes, 1999) et des travaux comme ceux de (Farzan et Brusilovsky, 2005), par exemple, exploitent les comportements de navigation des utilisateurs passés (les pages Web accédées, le temps de lecture de chaque page, le patron de navigation entre les pages, les annotations et commentaires laissés par les utilisateurs, etc.) en les interprétant comme des mesures d'appréciation pour ensuite guider les utilisateurs subséquents (qui ont des comportements similaires) vers l'information qui semble la plus intéressante et pertinente.

La recherche collaborative, de façon similaire, exploite le comportement de recherche des uns pour guider la recherche des autres. En effet, les moteurs de recherche traditionnels fonctionnent assez bien lorsque l'utilisateur a une bonne connaissance du sujet et qu'il peut dès lors formuler une requête précise quant à l'information recherchée. Cependant, bien souvent, l'utilisateur n'est pas un expert dans le domaine et se retrouve submergé par la quantité d'informations, plus ou moins pertinentes, qu'il reçoit au terme de sa requête.

Dans le but de personnaliser les résultats de recherche, des études en recherche d'informations collaborative comme celle de (Teevan et al. 2005; Smyth et al., 2005), ont montré comment les comportements de recherche des utilisateurs passés (les requêtes effectuées, les documents consultés ou sauvegardés...) peuvent être utilisés comme indicateur de pertinence et d'intérêt pour guider les recherches futures d'autres utilisateurs. Aussi, d'autres outils construisent et utilisent des réseaux d'affinités entre les utilisateurs pour aider à la recherche d'informations (Memmi et Nérot, 2003). Ceux-ci augmentent les requêtes imprécises en utilisant la connaissance d'experts sur le sujet, qui se trouvent dans le réseau de l'utilisateur. Dans le même esprit, Google Personalized Search (http://www.google.com/psearch) personnalise les résultats de recherche des utilisateurs en considérant (entre autres) les comportements de recherche passés d'utilisateurs similaires.

Étant donné que la navigation sociale et la recherche d'informations collaborative se font souvent de façon conjointe à partir d'un même fureteur, il semble en effet assez naturel d'intégrer ces deux techniques. C'est ce qu'ont fait (Freyne et al., 2007) en soulignant les bénéfices d'une telle intégration. Dans le même ordre d'idées, (Papagelis et al., 2008) proposent un système qui étend les fonctionnalités des fureteurs classiques en offrant une série d'outils qui favorisent la collaboration entre les internautes, en temps réel. Leur système permet entre autres de visualiser, sur un radar virtuel, les autres utilisateurs qui sont actuellement connectés au système et qui présentent des comportements de navigation ou de recherche plus ou moins similaires au navigateur courant. Celui-ci peut à tout moment profiter des outils de communication instantanée disponibles pour communiquer directement avec les autres utilisateurs.

Une autre approche plus récente et très intéressante pour améliorer la navigation et la recherche d'informations sur le Web est le *tagging* social et les fameux "*tag clouds*" qui en découlent et deviennent si populaires. Cette pratique de classification collaborative consiste en l'agrégation des tags ajoutés par les utilisateurs pour qualifier une ressource. De cette manière, à chaque ressource est associée une liste de mots-clés, pondérés par leur fréquence (ou leur popularité pour décrire cette ressource). On a montré que les tags associés aux ressources, lorsqu'utilisés comme métadonnées, améliorent effectivement la navigation (Millen et Feinberg, 2006) et la recherche d'informations (Heymann et al., 2008). Dans la même optique, (Zubiaga, 2009) propose une façon de rendre la navigation et la recherche d'informations plus efficace dans un système classique de classification taxonomique comme Wikipédia en le combinant avec un mécanisme de *tagging* social. Il expose différentes méthodes de navigation qui seraient possibles avec les tags. Par exemple, on pourrait naviguer parmi les articles populaires selon un certain tag c.-à-d. parmi les articles qui sont associés très fréquemment avec ce tag particulier. Dans le cas de la recherche, les tags fournissent de nouveaux termes qui ne sont pas nécessairement présents dans le contenu de l'article et qui pourraient être utilisés par le moteur de recherche pour répondre à une requête.

2.7.2 Les systèmes de recommandation par filtrage collaboratif

Parmi les approches prometteuses pour pallier la surabondance d'informations sur Internet et personnaliser la diffusion de l'information, on retrouve aussi les systèmes de recommandation qui se basent sur les évaluations (notes d'appréciation) de produits ou services (livres, musique, films, restaurants...) faites par les utilisateurs. Les profils d'intérêts (les préférences) des utilisateurs sont donc déduits des items qu'ils ont consommés et des évaluations qu'ils en ont faites. Aussi, parfois le profil est augmenté d'autres informations implicites déduites du comportement des utilisateurs. Ainsi, le problème de la recommandation consiste essentiellement à prédire les évaluations pour les items qui ne sont pas encore connus d'un utilisateur donné, afin de déterminer si ces items devraient lui être recommandés ou non : une fois qu'on a estimé les scores d'évaluation des items encore jusque-là non évalués, on peut alors proposer à l'utilisateur les items ayant les meilleurs scores d'évaluation estimés. Les systèmes de recommandation peuvent se classer en trois

catégories selon qu'ils utilisent une méthode de recommandation 1) basée sur le contenu, 2) basée sur filtrage collaboratif ou 3) hybride, basée sur la combinaison des deux premières approches.

La première approche, basée sur le contenu, consiste à estimer le score d'appréciation d'un item donné, pour un utilisateur particulier, en fonction des évaluations qu'il a lui-même faites dans le passé. Pour recommander de nouveaux livres à un utilisateur, par exemple, un système de recommandation basé sur le contenu tentera de comprendre ce qu'il y a de commun (auteurs, genres, sujets, etc.) parmi tous les livres que l'utilisateur a déjà évalués et auxquels il a attribué un score d'appréciation élevé. Puis, le système recommandera seulement les livres qui présentent une similarité forte avec les livres préférés de l'utilisateur.

Cependant, un des problèmes de cette approche consiste en la surspécialisation des recommandations effectuées. En effet, puisque le système peut seulement recommander des items qui ressemblent beaucoup au profil de l'utilisateur, les recommandations sont limitées à des items très similaires à ceux déjà évalués par celui-ci. L'approche par filtrage collaboratif, décrite ci-dessous, résout ce problème.

Les systèmes de recommandation collaboratifs (ou basés sur le filtrage collaboratif), pour leur part, essaient de prédire les items intéressants pour un utilisateur particulier, en se basant sur les évaluations faites par d'autres utilisateurs qui lui ressemblent. Par exemple, pour recommander des livres à un utilisateur donné, un système de recommandation collaboratif tente de trouver des utilisateurs qui ont des préférences similaires c.-à-d. qui ont évalué les mêmes livres de manière semblable. Puis, seulement les livres les plus appréciés des homologues de cet utilisateur lui seront alors recommandés.

Ainsi, étant donné que les systèmes de recommandation collaboratifs utilisent les recommandations des autres utilisateurs, ils peuvent recommander des items beaucoup plus variés que dans le cas précédent, où les recommandations ne sont effectuées que sur la base du contenu des items évalués dans le passé, par l'utilisateur lui-même.

Comme exemple de systèmes de recommandation collaboratifs, on peut citer le projet de recherche *GroupLens* (Resnick et al., 1994 ; Konstan et al., 1997) qui filtre les nouvelles de Usenet en construisant un réseau d'affinités et en utilisant un algorithme de filtrage basé sur les k plus proches voisins. Dans cet algorithme, les recommandations pour un utilisateur donné sont effectuées en choisissant d'abord un sous-ensemble des k utilisateurs les plus similaires (dans leurs préférences) à l'utilisateur donné. Puis, en effectuant une moyenne pondérée des évaluations de ces k utilisateurs, le système prédit les articles susceptibles d'intéresser cet utilisateur.

Le système de recommandation collaboratif (item à item), du site d'achat en ligne Amazon (http://amazon.com), utilise une méthode un peu différente pour effectuer les recommandations (Linden, 2003). Au lieu d'associer les utilisateurs à d'autres consommateurs similaires, ce système associe tous les achats et les items évalués d'un utilisateur à d'autres articles similaires et combine ensuite les items qui se ressemblent le plus sous forme d'une liste de recommandations. Plus précisément, pour déterminer les items les plus semblables à un item donné, leur algorithme construit une matrice de similarités entre les items en considérant que les produits achetés par un même consommateur sont possiblement similaires. Donc, pour un item i, on cherche d'abord tous les consommateurs c qui ont acheté l'item i. Ensuite, on prend chaque item j acheté par le consommateur c et l'on enregistre qu'un consommateur a acheté à la fois l'item i et l'item j. Finalement, pour tous les items j trouvés, on calcule la similarité entre l'item i et l'item j. Avec cette matrice de similarités entre les items, l'algorithme peut rapidement trouver, pour un utilisateur donné, les items qui sont semblables à tous ceux qui ont été achetés ou évalués par cet utilisateur. On recommande ensuite ceux qu'on a trouvé qui sont les plus similaires.

Comme exemple de système hybride, qui utilise à la fois les recommandations basées sur le contenu et celles basées sur le filtrage collaboratif, on peut mentionner Netflix (http://netflix.com), qui est un site sur lequel il est possible de visionner divers films et séries télévisées. Les utilisateurs sont amenés à donner leur appréciation des contenus vidéo qu'ils visionnent et Netflix leur propose ensuite des suggestions de films ou séries susceptibles de

leur plaire parce qu'ils ressemblent à ceux qu'ils ont aimés dans le passé et parce qu'ils ont été appréciés par d'autres utilisateurs ayant des préférences similaires.

Pour un survol plus approfondi des systèmes de recommandation et des diverses techniques servant à calculer et comparer les préférences des utilisateurs, nous référons le lecteur à (Adomavicius et Tuzhilin, 2005)

Le filtrage collaboratif utilise donc l'expérience des uns (les évaluations) pour enrichir l'expérience des autres. Comme dans la socialisation, les individus agissent ici comme des filtres de l'information.

2.7.3 Compilation de profils d'intérêts

Comme on vient de le voir, la plupart des systèmes sociaux et collaboratifs se basent sur l'extraction et la comparaison de profils pour identifier les utilisateurs de profils similaires et créer des liens implicites entre eux. Pour construire des profils, outre l'utilisation des évaluations explicites des utilisateurs, plusieurs méthodes sont possibles.

Tout d'abord, un utilisateur peut définir explicitement son profil en spécifiant, par exemple, ses intérêts musicaux, ses passe-temps préférés, son domaine d'étude, etc. Dans le cas de systèmes automatiques, il peut être préférable, cependant, de pouvoir extraire implicitement les profils à partir de l'information disponible selon le contexte.

Dans le cas de la navigation sociale dans un fureteur Web, on peut extraire les préférences d'un utilisateur, par exemple, selon les pages qu'il visite lorsqu'il navigue sur le Web, combien de temps il s'attarde sur cette page et s'il la parcourt au complet ou non (*scrolling*). Dans le cas de la recherche collaborative, on peut examiner les mots qu'un individu utilise dans ses requêtes sur un moteur de recherche et puis les pages qu'il consulte dans la liste des pages retournées par la requête, s'il enregistre une de ces pages dans ses favoris (marque-pages ou signets). Cette manière de procéder ne se pratique cependant qu'en contexte Web.

Dans les sites de réseautage en ligne, on peut extraire un profil d'intérêts implicite d'un membre en observant le contenu qu'il publie, les autres profils qu'ils visitent, les intérêts de ses contacts personnels, etc. Dans le cas des sites d'achats en ligne, on examine les produits consommés par un utilisateur pour en déduire ses préférences. Pour un résumé des objets classique qu'on peut récupérer, selon les contextes, afin d'en déduire implicitement les préférences des utilisateurs, voir (Kelly et Teevan, 2003).

Plus récemment, l'utilisation de tags devient de plus en plus populaire pour extraire les préférences des utilisateurs. En effet, dû à la popularité des systèmes de *tagging* collaboratif (ou folksonomies), on remarque l'apparition de quelques travaux qui visent à concevoir des méthodes pour extraire des profils d'intérêts à partir des *tags* associés aux ressources des utilisateurs (Michlmayr et Cayzer, 2007 ; Diederich et Iofciu, 2006).

Les systèmes de *tagging* collaboratif sont des sites Web comme Delicious (http://delicious.com) ou BibSonomy (http://bibsonomy.org) qui permettent aux utilisateurs de transférer des ressources sur un serveur et de les partager avec les autres utilisateurs. Dans le cas de Delicious, les ressources sont des URL de pages Web et dans le cas de BibSonomy, ce sont des références au format BibTeX. Les utilisateurs peuvent ensuite classer leurs ressources par mots-clés en leur assignant les tags de leur choix. En faisant l'hypothèse raisonnable que les tags associés aux ressources reflètent le contenu de ces ressources et que les ressources partagées par les utilisateurs reflètent leurs intérêts, l'ensemble des tags d'un utilisateur est alors une source d'information utile pour extraire son profil d'intérêts. En fait, la plupart des réseaux sociaux en ligne qui permettent la publication de contenu permettent aussi de lui assigner des mots-clés.

Bref, les traces électroniques récupérables pour construire des profils d'intérêts de manière implicite sont nombreuses et l'information disponible dépend essentiellement du contexte. Comme nous l'avons déjà mentionné, notre modèle de socialisation n'impose pas de méthode de profilage spécifique, mais nous devrons tout de même en choisir une pour composer les profils d'intérêts des acteurs qui peupleront nos simulations. Nous expliquerons la méthode choisie plus en détail au chapitre 3 (art. 3.4.2) et au chapitre 4 (sect. 4.2).

2.8 Conclusion

Nous avons vu, dans ce chapitre, plusieurs notions qui touchent les réseaux sociaux et la collaboration en général. Tout d'abord, nous avons présenté ce que sont les médias sociaux et comment ceux-ci offrent de nouveaux lieux de socialisation très accessibles, et utiles pour parfaire son réseau personnel de contacts et échanger des connaissances. Cependant, on a montré que la socialisation active, tant dans nos réseaux réels que virtuels, demande tout de même un certain investissement. D'un point de vue informatique, c'est dans cette optique que nous proposons un modèle de socialisation qui permet la formation automatique de communautés d'intérêt, dans un réseau ouvert, en constante évolution.

Ensuite, nous avons abordé diverses mesures qui caractérisent l'aspect structural des réseaux représentés sous forme de graphe, et nous avons montré que les réseaux sociaux (réels et virtuels) ne semblent pas évoluer de manière aléatoire, mais produisent plutôt des structures particulières. Pour valider notre modèle d'évolution de réseaux, d'un point de vue sociocognitif, nous utiliserons certaines des mesures présentées pour analyser différents aspects structuraux des réseaux générés par notre modèle de socialisation. Nous comparerons ensuite nos résultats aux propriétés structurales typiques qu'on observe dans la réalité.

De plus, nous avons décrit divers modèles qui tentent d'expliquer les propriétés structurales récurrentes qu'on retrouve dans plusieurs réseaux sociaux : l'effet des petits mondes, la transitivité (*clustering*), la distribution des degrés suivant une loi de puissance et l'émergence de communautés structurales. Nous avons aussi présenté des modèles dynamiques qui proposent deux mécanismes plausibles d'évolution des réseaux sociaux : l'attachement préférentiel (*the rich get richer*), qui tente d'expliquer la distribution des degrés suivant une loi de puissance observée dans plusieurs grands réseaux ouverts et la fermeture des triangles (les amis de nos amis sont nos amis), qui explique la présence d'une transitivité élevée dans nos réseaux sociaux en général. Dans le cadre de cette thèse, au niveau de la modélisation sociocognitive, nous proposons les mécanismes de socialisation (parcours, rencontres, recommandations) comme mécanismes d'évolution des réseaux sociaux qui pourraient expliquer certaines propriétés structurales observées dans la réalité. À notre connaissance, il

n'existe pas de modèles proposant ces mécanismes particuliers pour expliquer l'évolution des réseaux.

Nous avons ensuite présenté plusieurs études portant sur la recherche et la diffusion de l'information au sein des réseaux sociaux. Nous avons vu que le patron des connexions, que la structure des liens qui unissent les individus les uns aux autres, dans un réseau social, influence grandement sa capacité à diffuser l'information ainsi que sa navigabilité. En nous inspirant de ces travaux, d'un point de vue informatique, nous proposons donc de contrôler l'évolution des réseaux générés par notre modèle afin qu'ils maintiennent certaines propriétés structurales qui favorise leur navigabilité lors des parcours de socialisation. Nous verrons, au chapitre suivant, que d'un point de vue fonctionnel, nous aurons à négocier certaines propriétés structurales désirables pour la navigabilité, mais un peu moins plausibles du point de vue de la modélisation sociocognitive.

Finalement, nous avons abordé les systèmes sociaux et collaboratifs. En effet, notre projet de recherche s'inscrit bien dans cette philosophie qui consiste à réutiliser l'expérience des uns au profit des autres. Comme nous l'avons mentionné à plusieurs reprises, la socialisation est un processus fondamentalement collaboratif. De plus, que ce soit pour la navigation sociale, pour la recherche collaborative ou dans les systèmes de recommandation collaboratifs, on retrouve cette idée d'association des individus similaires par comparaison des profils. Dans notre modèle de socialisation, nous reprenons cette idée d'associations déduites par calcul, de création de liens implicites entre les individus de profils similaires, à la différence que nous transformons ces liens implicites en liens explicites dans le réseau. La similarité entre deux individus, une fois découverte, s'enregistre directement dans la structure du réseau, sous forme de communautés d'intérêts. Ainsi, l'information est conservée sans recours à un lieu de stockage centralisé.

Au chapitre qui suit, nous présentons notre modèle de socialisation.

CHAPITRE III

DESCRIPTION DU MODÈLE DE SOCIALISATION

3.1 Introduction

Dans un premier temps, nous formalisons notre modèle de socialisation par analogie aux mécanismes sociaux que l'on observe dans nos réseaux sociaux usuels. Nous proposons quatre règles d'évolution du modèle :

1. la règle de connexion, qui modélise l'arrivée d'un nouvel individu dans le réseau
2. la règle de déconnexion, qui représente le départ d'un individu
3. la règle de socialisation, qui formalise les mécanismes des rencontres aléatoires et des recommandations, lors de parcours de socialisation
4. la règle de mise à jour des relations, qui modélise le fait qu'un individu peut devoir changer de communauté lorsque ses intérêts, qui ont évolué au cours du temps, ne correspondent plus à ceux des membres de sa communauté d'appartenance.

Nous discutons ensuite du type de regroupement structural choisi pour former les communautés d'intérêt au sein du réseau, en tentant de conserver une certaine vraisemblance sociale, mais en privilégiant d'abord, d'un point de vue fonctionnel, leur navigabilité.

Puis, nous expliquons en détail les algorithmes qui implémentent les règles d'évolution du modèle, selon deux variantes du type de regroupement choisi : les regroupements autour de nœuds pivots simples et les regroupements autour de nœuds pivots chaînés. Nous utiliserons ici un système numérique simple pour représenter et comparer les profils d'intérêts des acteurs dans le réseau.

Finalement, nous expliquons une approche plus réaliste pour représenter les profils d'intérêts des acteurs du réseau ainsi que la méthode que nous utiliserons pour comparer ces profils

entre eux. Ce sont ces méthodes d'extraction et de comparaison des profils que nous utiliserons pour simuler notre modèle au chapitre suivant.

3.2 Formalisation du modèle

3.2.1 Composants du modèle

3.2.1.1 Le réseau, les individus et les liens

Le modèle que nous proposons est un modèle de socialisation dans un réseau ouvert, c.-à-d. qui évolue constamment avec l'arrivée et le départ d'individus, la formation de nouvelles associations ainsi que la cessation d'anciennes relations. Certains liens perdureront longtemps, d'autres ne sont que d'une durée temporaire et c'est ainsi que le réseau évolue constamment dans le temps. Cette évolution implique donc la restructuration constante du patron des relations entre les membres du réseau et contrairement aux petits réseaux fermés, personne ne connaît tout le monde dans le réseau ni la structure globale du patron des relations qui existent entre tous ses membres. Cependant, chaque personne connaît une portion plus ou moins grande du réseau, constituée de ses contacts personnels. La réunion de ces petits réseaux locaux de contacts personnels, se chevauchant parfois, parfois non, forme ainsi le réseau global.

Soit le graphe non orienté G représentant un tel réseau social dans lequel chaque nœud (ou sommet) représente un individu. Un lien (une arête) entre deux nœuds, n et m, signifie que l'individu représenté par le nœud n connaît (est en relation avec) l'individu représenté par le nœud m. Le fait que le graphe soit non orienté signifie que si n connaît m alors m connaît aussi n. En d'autres termes, les échanges sociaux impliquent des relations mutuelles.

Nous utiliserons le terme *réseau global* pour référer à G par opposition au terme *réseau local* que nous utiliserons pour parler du réseau personnel d'un individu, c.-à-d. le réseau formé par l'ensemble de ses contacts personnels. On peut comprendre le réseau global comme étant composé de plusieurs réseaux locaux qui se chevauchent.

Nous utiliserons le terme *voisin de n* pour désigner un nœud du graphe *G* qui est connu de *n*, soit directement ou indirectement. Nous dirons alors qu'un *voisin immédiat de n* est un nœud *v* de *G* tel qu'il existe un lien entre *n* et *v* dans *G* et qu'un *voisin indirect de n* est un nœud *v* de *G* tel qu'il n'existe pas de lien entre *n* et *v*, mais qu'il existe un chemin allant de *n* à *v* et que *ce chemin est connu* (ou peut être connu) de *n* et de *v*. Nous verrons, à la section 3.3 (par. 3.3.2.2), que tous les membres de la communauté de *n*, même s'ils ne sont pas des voisins immédiats de *n*, peuvent en effet être facilement localisés et donc visités par *n*. Dans cette optique, nous considérons que le réseau local (ou personnel) d'un nœud *n* est composé de l'ensemble de ses voisins (immédiats et indirects). Plus précisément, tous les voisins immédiats et indirects de même profil que *n* constituent sa communauté d'intérêt d'appartenance.

3.2.1.2 Représentation numérique des profils d'intérêts et mesure de similarité

Pour le moment, par souci de simplicité et de clarté, nous représenterons le profil d'intérêts de chaque individu à l'aide d'un attribut numérique (> 0) assigné à chaque nœud du graphe. C'est la valeur de cet attribut que nous allons comparer pour calculer la similarité des profils entre deux nœuds. Tout simplement, plus la valeur du profil du nœud *n* est proche de celle du profil du noeud *m*, plus *n* et *m* seront dit similaires. Nous utiliserons la fonction ci-dessous pour calculer la similarité *sim* entre deux profils numériques, $p1$ et $p2$. Cette fonction retourne un nombre positif plus petit ou égal à 1. Une valeur retournée égale à 1 indique que les deux profils comparés sont identiques sinon, plus la valeur de *sim* se rapproche de 1, plus les deux profils sont similaires.

$$sim(p1, p2) = \frac{1}{|p1 - p2| + 1}.$$

À la section 3.4 (art. 3.4.2), nous proposerons une solution plus réaliste quant à la représentation et la comparaison des profils d'intérêts des participants du réseau.

3.2.2 Comportements et règles d'évolution du modèle

Nous décrivons ici l'idée générale des règles de fonctionnement de notre modèle par analogie aux mécanismes de socialisation observés dans nos réseaux sociaux habituels. Nous reformulerons cependant ces mécanismes de manière plus concrète à la section 3.3 qui traite des algorithmes implémentant ces règles d'évolution et qui introduira des considérations supplémentaires plutôt axées sur l'aspect fonctionnel du modèle que sur son aspect sociocognitif.

Les règles de fonctionnement du modèle sont des règles strictement locales, c.-à-d. qu'elles sont exécutées au niveau des sommets du graphe qui représente le réseau social et non au niveau du réseau global. Dans cette optique, on peut considérer chaque nœud du graphe comme un acteur capable d'effectuer certaines tâches (sociales) simples : des *connexions*, des *déconnexions*, des *parcours de socialisation* et des *mises à jour de relations*.

3.2.2.1 Règle de connexion

Cette règle a pour but de modéliser l'arrivée d'un nouvel individu dans le réseau qui commence à socialiser de manière aléatoire en se déplaçant d'un individu à un autre, à la recherche d'individus partageant ses intérêts.

Formellement, dans notre modèle, on ajoute un nouveau nœud dans le graphe G, disons le nœud *i*. Ensuite, *i* effectue un parcours de connexion, tel que décrit ci-dessous.

Parcours de connexion

Le nouveau nœud *i* choisit d'abord, de manière aléatoire, un nœud déjà présent dans le graphe à partir duquel il entreprend son parcours de connexion. Disons qu'il choisit le nœud *j*. À partir de *j*, donc, *i* choisit le prochain nœud à parcourir en prenant le voisin immédiat de *j* dont le profil se rapproche le plus son profil (en utilisant notre fonction de similarité). Supposons que *i* ait choisi le voisin *v*, il choisit ensuite, parmi les voisins immédiats de *v*, celui qui lui ressemble le plus et ainsi de suite, de nœud en nœud, sans jamais revenir en

arrière sur le parcours. Le trajet s'arrête lorsque *i* trouve un nœud de profil identique ou bien qu'il ne reste aucune possibilité locale pour le choix d'un nœud suivant (un nœud qui n'a pas déjà été visité). Si un nœud de profil identique a été trouvé, *i* crée une nouvelle relation (un lien) avec ce noeud, il se connecte au nœud trouvé (ou bien à l'un des nœuds de même profil appartenant à la communauté d'intérêt du nœud trouvé). Si aucun nœud de profil identique n'a été trouvé, *i* se connecte au nœud le plus similaire rencontré lors du parcours (ou à l'un des nœuds appartenant à la communauté d'intérêt du nœud le plus similaire rencontré).

Nous verrons ce qui détermine la longueur d'un parcours lorsque nous aborderons plus en détail l'algorithme de connexion à la section 3.3. En effet, on ne veut pas parcourir tous les nœuds du graphe de la même manière qu'un individu, dans un grand réseau ouvert, ne peut pas socialiser avec tous les individus présents.

3.2.2.2 Règle de déconnexion

Ce mécanisme modélise le fait que, dans un réseau dynamique, on observe un va-et-vient constant et que tandis que le réseau accueille de nouveaux arrivants (connexion), d'autres le quittent, que ce soit pour des raisons de changements de lieu, de changements d'intérêts, de disponibilité, etc.

Dans notre modèle, cela correspond à la suppression d'un noeud *i* de *G* et à la suppression de tous les liens de *i*. Nous verrons, à la section 3.3, que nous implémenterons aussi des mécanismes supplémentaires, plutôt liés à la performance du modèle, pour éviter, entre autres, qu'à la déconnexion d'un noeud, le réseau ne se fragmente en plusieurs composantes non connexes, ce qui aurait pour effet de créer plusieurs réseaux indépendants et de limiter l'accès des membres d'un sous-réseau aux individus des autres sous-réseaux.

3.2.2.3 Règle de socialisation avec rencontres et recommandations

La règle de socialisation décrit les comportements des individus lorsqu'ils décident de socialiser activement. Ceux-ci explorent le réseau en naviguant d'individu en individu, à la recherche de personnes intéressantes et intéressées dans le but d'échanger. Échanger des

connaissances, échanger des contacts. Comme nous l'avons déjà mentionné, notre modèle se concentre sur l'échange de contacts personnels, par le biais des rencontres aléatoires et des recommandations qui sont responsables de la restructuration du réseau pour le rapprochement d'individus ayant des intérêts communs.

Parcours de socialisation

Nous appellerons l'individu qui effectue un parcours de socialisation un *socialisateur*. Formellement, dans notre modèle, le parcours de socialisation se fait de la même manière que le parcours de connexion. Cependant, contrairement à l'individu qui se connecte au réseau, le socialisateur se trouve déjà dans le réseau global et possède donc un réseau de contacts personnels au sein même de ce réseau. Les parcours de socialisation permettent donc au socialisateur de rencontrer de nouveaux individus intéressants (rencontres aléatoires) pour constamment enrichir et renouveler son réseau personnel, au fil du temps. De plus, lors de ses parcours, si le socialisateur rencontre un individu intéressant pour l'un de ses contacts personnels, il peut les présenter l'un à l'autre (recommandations) et ainsi favoriser les rencontres pertinentes au niveau global du réseau.

Lorsqu'un nœud i décide de socialiser, celui-ci choisit d'abord aléatoirement un nœud j présent dans le réseau pour commencer son parcours. Il navigue ensuite de nœud le plus similaire en nœud le plus similaire, puis lorsqu'il rencontre un nœud de profil identique, il effectue une rencontre et lorsqu'il rencontre un nœud qui n'est pas de même profil que lui-même, mais de profil identique à l'un de ses contacts personnels de profil différent, il effectue une recommandation.

Les rencontres aléatoires

Le mécanisme de rencontres aléatoires modélise le processus des rencontres effectuées lors d'un parcours de socialisation. On commence par échanger avec une personne, puis on passe à une autre, sans vraiment savoir si l'on va rencontrer quelqu'un avec qui l'on a des affinités. C'est en ce sens qu'on parle ici de hasard. Cependant, c'est en rencontrant des gens de cette

manière qu'on crée la possibilité d'une rencontre intéressante. Lorsqu'on rencontre un individu avec lequel on a des affinités, on crée un lien. La formation de ce nouveau lien a pour effet de créer un pont entre les communautés d'appartenance respectives des deux individus qui se sont rencontrés. On a montré, en effet, que dans nos réseaux sociaux réels, lorsque deux individus se rencontrent (se connaissent), il y a une forte probabilité que les contacts personnels de ces deux individus se rencontrent aussi. On parle ici du phénomène de transitivité, présenté au chapitre 2, qu'on observe dans nos réseaux sociaux habituels : les amis de nos amis sont aussi nos amis. Dans notre modèle, on effectue alors à une restructuration du réseau par la création de nouveaux liens d'affinités entre certains membres des communautés d'appartenance des deux individus qui se sont rencontrés. C'est dans cette optique qu'on parle d'échange de contacts lors de rencontres aléatoires. Une seule rencontre entre deux individus devient bénéfique pour plusieurs (tous les membres des communautés d'intérêt des deux individus qui se sont rencontrés). Dans cette optique, on peut voir ce phénomène comme un processus de collaboration sociale implicite.

Les recommandations

Lorsqu'un individu socialise, il peut parfois rencontrer un autre individu n'ayant pas nécessairement les mêmes intérêts que lui, mais qu'il peut recommander à l'un de ses contacts personnels qui semble avoir des intérêts en commun avec cette personne rencontrée. En contexte social, ceci se produit selon différents scénarios. On peut, par exemple, discuter avec un individu et lui dire, tout simplement, qu'il aurait intérêt à rencontrer telle ou telle personne qu'on connaît ou qu'on vient de rencontrer, on peut aussi le présenter directement à l'une de nos connaissances, en les introduisant personnellement l'un à l'autre, etc. Dans tous les cas, ce qui nous intéresse, ici, est qu'on crée un nouveau lien entre ces deux individus recommandés l'un à l'autre.

Dans cette optique, de la même manière que pour les rencontres aléatoires, une recommandation a pour effet de joindre les communautés d'appartenance des deux individus recommandés l'un à l'autre en favorisant la création de nouveaux liens entre tous les membres des deux communautés ainsi réunies. Par contraste aux rencontres aléatoires, le principe de

recommandation peut être vu comme un processus de collaboration sociale explicite dans lequel on partage volontairement nos contacts personnels pour s'entraider les uns les autres : on effectue des recommandations pour mettre en contact des individus similaires et l'on profite aussi des recommandations des autres qui nous dirigent vers des individus intéressants.

D'un point de vue formel, une rencontre ou une recommandation consiste à faire une restructuration du réseau. Dans le cas d'une rencontre, on relie le sous-réseau local du nœud qui socialise au sous-réseau local du nœud rencontré tandis que dans le cas d'une recommandation, on relie ensemble les deux sous-réseaux des nœuds recommandés l'un à l'autre. À la section 3.3, nous parlerons de fusion de communautés pour expliquer la manière dont nous allons traiter ce phénomène d'un point de vue fonctionnel.

Les rencontres et les recommandations lors d'un parcours de socialisation ont donc pour effet de rapprocher les individus de profils similaires dans le réseau, de les regrouper ensemble au sein de communautés d'intérêt.

3.2.2.4 Règles de mise à jour des relations

Cette règle a pour but de modéliser le fait qu'au cours du temps, les intérêts d'un individu peuvent changer. Parfois, un individu se découvre de nouveaux intérêts et en abandonne d'autres sans toutefois quitter le réseau. Celui-ci peut alors se retrouver au sein d'une communauté d'intérêt qui ne répond plus à ses besoins. Celui-ci cherche alors de nouvelles relations qui sont plus proches de ses intérêts, une nouvelle communauté d'appartenance.

Formellement, dans notre modèle, cette règle se traduit tout simplement par une déconnexion (le noeud quitte son ancienne communauté) suivie d'une connexion (le nœud recherche une nouvelle communauté d'appartenance selon ses nouveaux intérêts).

La section suivante présente la mise en œuvre des règles d'évolution de notre modèle de socialisation. Nous verrons que les algorithmes proposés tiennent aussi compte de considérations pratiques dans l'implémentation de ces règles, toujours dans l'optique de

l'application éventuelle de ce modèle pour la réalisation, par exemple, d'un système collaboratif d'échanges d'informations ou de socialisation automatique.

3.3 Fonctionnement du modèle de socialisation

Avant d'expliquer en détail nos algorithmes d'évolution du réseau qui implémentent les règles de notre modèle de socialisation, nous discutons d'abord des choix que nous avons faits quant à la structure des communautés d'intérêt (et plus globalement du réseau complet) dans le but de favoriser la navigabilité de nos réseaux.

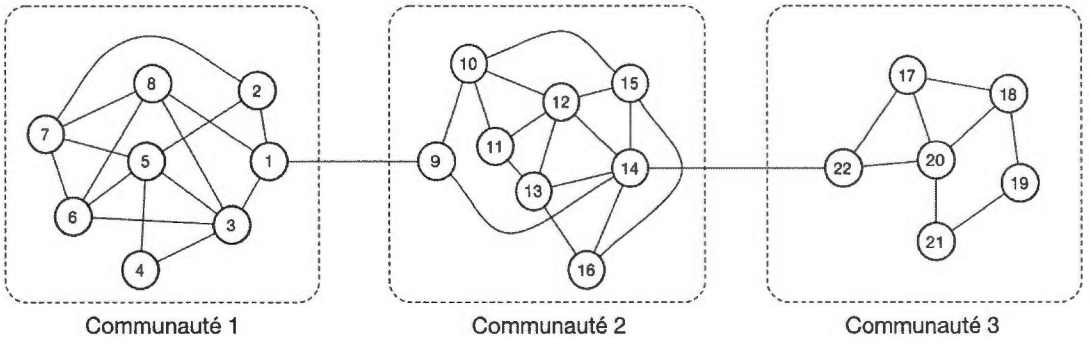
3.3.1 Structure du réseau et navigabilité

3.3.1.1 Type de regroupement pour la formation de communautés

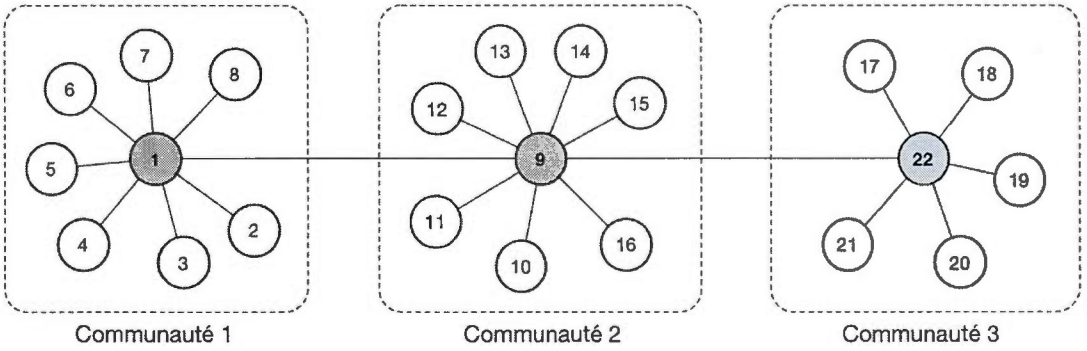
La première question qui se pose quant à l'implémentation de notre modèle de socialisation porte sur les regroupements d'individus au sein du réseau. En effet, divers types de configurations sont possibles. Dans la littérature, on a souvent observé que les individus, dans les réseaux sociaux, avaient tendance à former des sous-groupes fortement connectés, très cohésifs, mis en évidence par un fort taux de *clustering* (Girvan et Newman, 2002 ; Mislove, 2009). La figure 3.1 (a) illustre un exemple de telles communautés structurales très cohésives et fortement transitives. On a aussi observé des communautés en forme d'étoiles (Kumar et al., 2006) dans lesquelles un individu central est relié à tous les autres individus de sa communauté. La figure 3.1 (b) montre les mêmes communautés structurales qu'en (a), mais dont les individus sont regroupés en forme d'étoile. Dans les deux cas, on distingue clairement les communautés, au niveau structural, en comparant la densité des liens intra-communautés avec la densité des liens extra-communautés : il y a beaucoup plus de liens à l'intérieur des communautés que de liens qui relient les communautés entre elles.

Dans le cas qui nous occupe, pour favoriser la navigabilité dans le réseau lors des parcours de connexion et de socialisation, nous désirons, évidemment, obtenir un graphe *connexe*, dans lequel il existe un chemin entre n'importe quelle paire de nœuds. De plus, on le veut *non dense* pour qu'il ne génère pas trop de chemins possibles et *de petit diamètre* pour s'assurer de

l'existence de chemins de courte distance entre les nœuds. Pour obtenir une densité faible, il faut que le graphe comporte le moins de liens possible, et ce, tout en demeurant connexe.



(a)

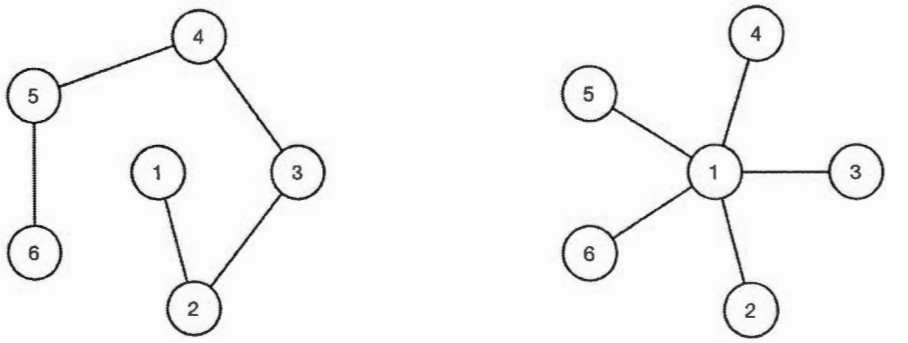


(b)

Figure 3.1 Structures de communautés (a) fortement transitives, et (b) en forme d'étoile.

On voit clairement que le fort taux de *clustering* dans les communautés illustrées à la figure 3.1 (a) ne favorise pas une faible densité. Il y a en effet beaucoup plus de liens dans le réseau de la figure 3.1 (a) que celui de la figure 3.1 (b). Une densité élevée, comme on l'a déjà expliqué, ne favorise pas la navigabilité au sens où, même si un court chemin existe, il devient difficile de le trouver parmi tous les choix possibles qu'offrent les nombreux liens. Pour s'en convaincre, regardons encore la figure 3.1 (a). Pour aller du nœud 5 au nœud 9, par exemple, on peut prendre le chemin 5 - 2 - 1 - 9 ou le chemin 5 - 2 - 7 - 8 - 6 - 3 - 1 - 9, ce dernier étant doublement plus long que le premier. Ceci n'est qu'un exemple et l'on comprend

aisément que les possibilités de longs chemins sont multiples. Dans le graphe de la figure 3.1 (b), cependant, pour relier les mêmes nœuds, on n'a qu'un seul choix de chemin : 5 - 1 - 9. Donc, en privilégiant une faible densité, on diminue la redondance des chemins et l'on augmente ainsi nos chances de trouver le bon. Cependant, une faible densité ne garantit pas la présence de courts chemins dans le réseau. Pour illustrer ce fait, la figure 3.2 montre deux petits réseaux de densité identique, mais dont les diamètres diffèrent grandement. On voit que le réseau de la figure 3.2 (b) possède un diamètre beaucoup plus petit que celui de la figure 3.2 (a), bien que les deux réseaux soient de même densité.



(a) densité = 0.33, diamètre = 5 (b) densité = 0.33, diamètre = 2

Figure 3.2 Densité et distance dans deux graphes de même taille.

La structure étoilée semble donc pouvoir favoriser à la fois, et la faible densité, et la grande proximité entre les nœuds. Nous avons donc choisi cette structure de base pour former nos communautés. Ces étoiles sont en fait composées d'un nœud pivot (un *hub*) auquel sont rattachés tous les autres membres de cette communauté (les nœuds simples). Nous parlerons alors de regroupement autour de nœuds pivots pour la formation de nos communautés d'intérêt, que nous expliquerons plus en détail aux articles 3.3.2 et 3.3.3.

Nous devons mentionner ici que nous venons de négocier une propriété structurale nécessaire pour favoriser la navigabilité d'un point de vue fonctionnel, au détriment d'une caractéristique désirable dans l'optique de modélisation sociocognitive. En effet, le *clustering* est une propriété structurale qu'on observe dans la plupart des réseaux sociaux et nous tentons de le minimiser au possible pour des raisons pratiques.

Cependant, outre l'absence de *clustering*, ce type de regroupements, autour de nœuds pivots, n'est pas invraisemblable au niveau sociocognitif. En effet, comme on l'a vu au chapitre 2, plusieurs réseaux sociaux réels ou virtuels possèdent une structure dont la distribution des degrés suit une loi de puissance, ce qui suppose la présence de ce genre de composantes étoilées. Par exemple, la figure 2.3 illustre un réseau social réel (de contacts sexuels) qui, nous le verrons, ressemble beaucoup à l'allure qu'aura notre réseau.

De plus, l'efficacité de ce type de regroupement n'est pas surprenante. Comme on l'a vu au chapitre 2, la présence de nœuds pivots, dans un réseau, permet l'accès rapide à plusieurs autres nœuds, possiblement pertinents (en fonction du contexte). Dans notre cas, ces nœuds pivots permettent l'accès rapide aux autres nœuds de même profil qui y sont rattachés (aux autres membres de la communauté) et il s'ensuit, comme nous le verrons au paragraphe 3.3.2.2, que les parcours dans le graphe pourront être limités à ces seuls nœuds clés. Mais tout d'abord, dans la section qui suit, nous voulons préciser davantage le type d'algorithme que nous utilisons pour effectuer nos parcours de connexion/socialisation dans le graphe.

3.3.1.2 Type de parcours dans le réseau

Le problème de recherche dans les graphes a été beaucoup étudié et l'on dispose d'un certain nombre d'algorithmes. Les stratégies les plus connues sont les parcours de recherche en largeur d'abord (*breadth-first search*), en profondeur d'abord (*depth-first search*), et qui choisissent le meilleur d'abord (*best-fit search*). On dira qu'un algorithme est complet s'il assure de trouver une solution et qu'il est optimal s'il assure de trouver le meilleur chemin. On peut aussi déterminer sa complexité (en temps et en mémoire) en fonction de b, le facteur de branchement (le nombre maximum de voisins d'un noeud), et d, la profondeur du graphe.

La stratégie de recherche en profondeur d'abord (*depth-first search*)

Cette technique sert à parcourir un graphe non orienté en développant toujours le nœud le plus profond : on choisit d'abord un nœud de départ, puis à partir de ce nœud, on explore une branche du graphe le plus loin possible. Si l'on ne trouve pas de nœud qui correspond à ce que l'on cherche sur cette branche, on revient sur ses pas (mécanisme de retour arrière en cas

d'échec) et l'on recommence alors l'exploration d'une autre branche, et ainsi de suite. L'algorithme de recherche en profondeur d'abord n'est ni complet ni optimal (il peut boucler ou explorer des chemins trop profonds), mais il est en général plus simple et efficace. En effet, la complexité est alors seulement O(bd) en mémoire, mais demeure élevée en temps avec O(b^d). Il est possible d'optimiser cet algorithme en limitant la profondeur a priori (si on en a une idée) pour éviter les bouclages et les mauvais choix initiaux. Concernant les cycles, on peut les éviter en notant les nœuds déjà visités (sur le même chemin ou dans le passé) ce qui peut se révéler coûteux en mémoire. Il existe un certain nombre de variantes plus ou moins complexes et efficaces. Par exemple, on peut utiliser la profondeur d'abord en augmentant progressivement la profondeur limite (c'est alors complet et optimal).

La stratégie de recherche en largeur d'abord (*breadth-first search*)

Cet algorithme sert aussi à parcourir un graphe non orienté, mais en développant les nœuds les moins profonds en premier : on choisit un nœud de départ puis, on explore les voisins immédiats de ce nœud, puis les voisins de ces voisins et ainsi de suite, jusqu'à ce qu'on trouve un nœud qui correspond à ce qu'on cherche ou que l'on ait visité tous les nœuds du graphe sans succès. Cet algorithme utilise généralement une file pour conserver les nœuds qui restent à parcourir, à mesure qu'il traverse le graphe :

```
Début
   Trouvé = faux
   Choisir un noeud de départ i
   Enfiler i
   Tant que pas Trouvé et que la file n'est pas vide faire
      Défiler un noeud k et l'examiner

      Si k correspond au noeud recherché alors
         Résultat = k
         Trouvé = vrai
      Sinon
         Enfiler tous les voisins immédiats de k qui n'ont pas encore
         été visités.
      Fin si

      Si la file est vide et que pas Trouvé alors
         Résultat = nul  (le noeud recherché n'a pas été trouvé)
      Fin si
   Fin tant que
   Retourner Résultat
Fin
```

Il est à noter que pour effectuer une recherche en profondeur, on n'aurait qu'à utiliser une pile au lieu d'une file dans l'algorithme précédent.

La technique de recherche en largeur d'abord est complète et optimale, mais de complexité prohibitive : $O(b^d)$ en temps et mémoire (souvent impraticable en mémoire). En effet, on garde en mémoire tous les chemins partiels en cours.

Les stratégies de recherche qui choisissent le meilleur d'abord (*best-first search*)

Les algorithmes précédents explorent le graphe systématiquement sans information préalable, d'où leur coût exponentiel qui devient vite prohibitif. Mais il est souvent possible de choisir les meilleures branches à explorer grâce à des heuristiques de parcours. Une heuristique ne garantit pas toujours un résultat, mais elle diminue grandement le coût du parcours en moyenne, à condition d'être elle-même facile à calculer.

Plus précisément, une heuristique propose une fonction d'évaluation *f(n)* permettant de choisir, à chaque pas, les nœuds les plus prometteurs lors du parcours du graphe.

Les algorithmes gloutons ou voraces (*greedy search*) en sont un cas particulier : on choisit à chaque pas la branche de moindre coût, minimisant par exemple la distance apparente vers le but donné. La recherche s'apparente donc à la profondeur d'abord avec retour arrière. Les algorithmes gloutons ne sont ni complets ni optimaux, mais c'est une approche simple et leur efficacité est souvent raisonnable.

L'algorithme A* est un autre exemple de ce genre de recherche. Il cherche à minimiser, à chaque pas, le coût du chemin déjà parcouru *g(n)*, plus le coût estimé du chemin restant vers le but *h(n)*. La fonction d'évaluation est donc $f(n) = g(n) + h(n)$. Comme on minimise le chemin déjà parcouru, cette stratégie s'apparente à la largeur d'abord.

Si *h(n)* ne surévalue pas le coût restant, l'algorithme A* est complet et optimal, mais sa complexité $O(b^d)$ reste élevée (surtout en mémoire) si le facteur de branchement effectif n'est

pas assez bas. Il est donc crucial d'avoir une bonne fonction d'estimation pour diminuer la combinatoire. Il existe des variantes plus perfectionnées (IDA*, SMA*) pour diminuer le coût en mémoire, mais elles sont plus compliquées et pas toujours efficaces.

En bref, la recherche dans les graphes est en général un problème computationnel difficile, dont le coût se révèle souvent prohibitif en pratique. Mais il existe des méthodes connues permettant de traiter le problème le mieux possible, et dont le comportement est maintenant bien compris.

Dans cette optique, nos parcours de socialisation (et de connexion) s'apparentent fortement aux algorithmes gloutons, mais sans retour arrière. La fonction d'évaluation *f(n)*, dans notre cas, est la fonction de similarité des profils : on choisit le nœud le plus similaire comme prochain nœud à parcourir puis on continue, en profondeur.

Comme nous venons de le voir, le coût des parcours dans un graphe augmente exponentiellement avec la taille du réseau (en nombre de liens et de cœur). Pour cette raison, nous avons choisi de ne pas faire de retour arrière afin de limiter encore plus la longueur des parcours dans le réseau. Ainsi, nos parcours de socialisation se limitent à l'exploration d'une seule branche du graphe et ne seront donc pas nécessairement toujours fructueux : parfois, un nœud qui socialise ne fera aucune rencontre ou recommandation lors de son parcours dans le réseau parce que le chemin qu'il aura choisi ne l'aura pas conduit vers un individu dans le réseau qu'il aurait pu rencontrer ou recommander (en supposant qu'il en existe un). Lorsqu'un nœud socialise, il ne sait pas, a priori, s'il existe dans le réseau, un individu qu'il pourrait rencontrer ou recommander. Il se peut que non. Dans ce cas, avec retour arrière, le nœud se verrait parcourir tout le réseau, ce qui risquerait d'être beaucoup trop coûteux dans de grands réseaux. Nous verrons qu'en réalité, avec la structure particulière de notre réseau, ce ne serait pas tout à fait le cas, mais tout de même, les parcours seraient plus longs avec retour arrière.

En considérant que la socialisation est un processus collaboratif et que donc, celle des uns profite aussi aux autres (par les fusions de communautés), nous pensons que même si

quelques socialisations ne sont pas fructueuses, d'un point de vue global, celles qui le sont seront suffisantes pour réunir les individus en communautés d'intérêt.

À l'article 3.3.2, nous expliquons notre modèle de socialisation qui regroupe les nœuds autour de nœuds pivots pour la formation des communautés d'intérêt au sein du réseau. À l'article 3.3.3, nous proposons une légère variante du premier type de regroupement, qui est une amélioration du premier, tant sur le plan fonctionnel que sociocognitif. Dans ce qui suit, nous voyons donc, de manière détaillée, les attributs des nœuds du réseau, les propriétés structurales à maintenir et les algorithmes d'évolution du réseau qui implémentent les règles d'évolution du modèle, pour les deux variations de regroupements proposées.

3.3.2 Regroupements autour de nœuds pivots simples

3.3.2.1 Attributs des nœuds et définitions

Ici, on tente de regrouper les nœuds autour d'un pivot de même profil. On doit donc pouvoir distinguer un nœud simple d'un nœud pivot. Cette information se retrouve sous la forme d'un attribut booléen associé à chaque nœud. La désignation des nœuds pivots se fait automatiquement, selon nos algorithmes, par l'intermédiaire de cet attribut.

Attribut des nœuds :

`profil` : Contient le profil d'intérêts du nœud – un nombre entier pour le moment.

`estPivot` : Indique si le nœud est un pivot ou non (valeur booléenne).

Définitions :

Pivot : Nœud qui a été désigné comme étant un pivot (`estPivot` a la valeur "vrai").

Nœud simple : Nœud qui n'est pas un pivot (`estPivot` a la valeur "faux").

3.3.2.2 Propriétés structurales à maintenir dans le réseau

La figure 3.3 montre un regroupement dans ce type de réseau à pivots simples. Le cercle plus grand au contour gras représente un nœud pivot et les cercles plus petits, des nœuds simples de même profil que le nœud pivot auquel ils sont rattachés. Le premier nombre de l'étiquette de chaque nœud indique le numéro d'identification unique de ce nœud et le nombre suivant, entre parenthèses, indique son profil numérique. Les traits discontinus représentent des liens vers d'autres nœuds pivots de profil différent, vers d'autres regroupements.

On voit donc que le pivot 7 de profil 6 a cinq voisins immédiats (nœuds simples) de même profil et que chaque nœud simple possède un seul voisin immédiat de même profil (le nœud pivot) et quatre voisins indirects de même profil qu'ils peuvent trouver facilement en passant par le pivot. Ce regroupement est donc composé de six nœuds et forme la communauté de chacun de ces nœuds.

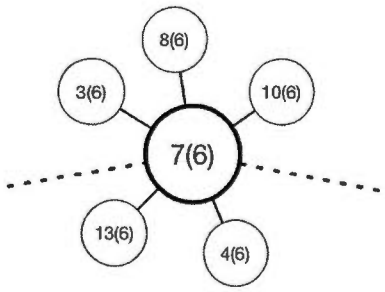


Figure 3.3 Regroupement autour d'un nœud pivot.
Le nœud pivot, ici, est le nœud 7(6) où 7 est le numéro d'identification du nœud et 6 représente son profil. Ce pivot a des voisins simples de profil identique (profil 6) et des liens vers des voisins de profils différents représentés par les traits discontinus. La communauté de profil 6 est donc formée du pivot et de tous ses voisins simples du même profil.

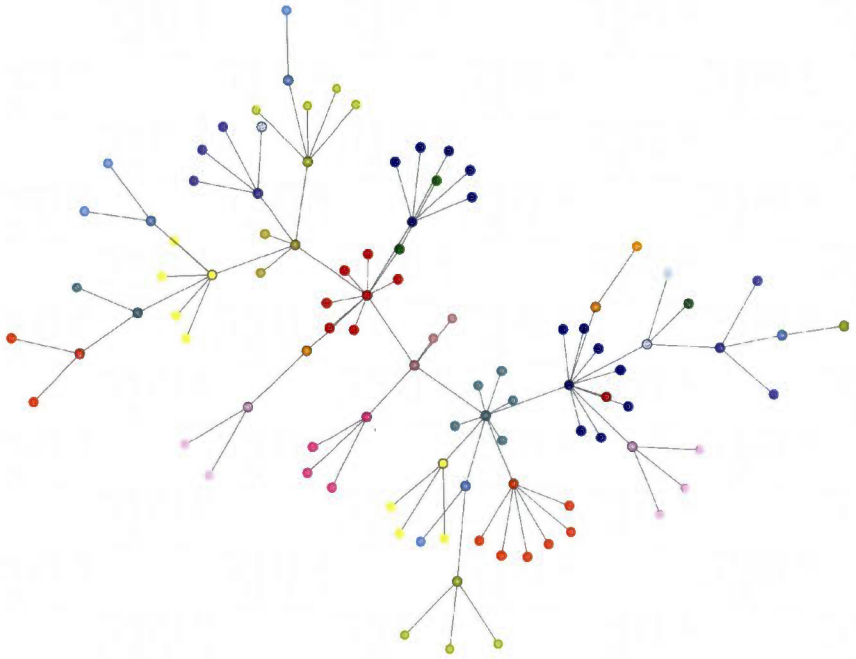
Nos algorithmes d'évolution du réseau doivent donc générer et maintenir un réseau qui a les propriétés structurales suivantes :

- Le réseau est connexe.

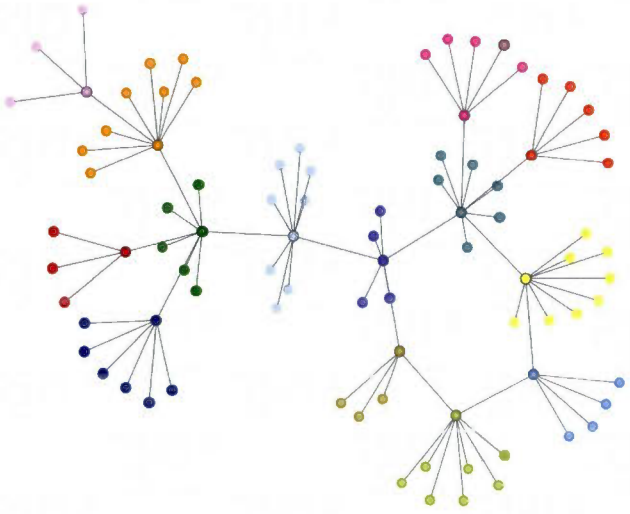
- Le réseau contient des nœuds pivots qui sont soit reliés à des nœuds simples de même profil, soit reliés à des nœuds pivots de profils différents. Un nœud pivot peut donc accéder directement à tous les membres de sa communauté.
- Le réseau contient des nœuds simples qui n'ont qu'un seul lien avec un nœud pivot de même profil. Les nœuds simples connaissent donc directement leur seul voisin pivot et tous les autres membres de leur sous-réseau de façon indirecte, en passant par le pivot puis aux autres nœuds simples rattachés à ce pivot.
- Les voisins pivots attachés à un pivot ne sont jamais de même profil entre eux. Par exemple, le pivot 1(2) ne pourrait pas avoir comme voisins pivots les voisins 5(3) et 7(3).

Dans un réseau possédant de telles caractéristiques structurales, il suffit de parcourir les nœuds pivots pour avoir accès à tous les autres nœuds du réseau. En effet, pour chaque nœud pivot, tous les voisins immédiats qui ne sont pas des pivots (nœuds simples) sont de même profil que le pivot et tous les voisins immédiats qui sont des pivots sont de profils différents. Cette propriété permet de limiter la longueur des parcours dans le réseau à ces seuls nœuds pivots. En effet, lorsqu'on rencontre un nœud pivot, on rencontre du même coup tous ses voisins immédiats de même profil (qui sont des nœuds simples). On remarque, aussi, que les nœuds d'une même communauté peuvent se visiter sans équivoque, de manière optimale et sans avoir à comparer de nouveau leurs profils d'intérêts. En effet, un nœud pivot sait que sa communauté est composée de tous ses voisins immédiats qui ne sont pas des pivots et un nœud simple sait que sa communauté est composée de son seul voisin immédiat (le nœud pivot) et de tous les voisins immédiats simples de ce pivot.

Un réseau parfait, dans lequel tous les nœuds de même profil sont regroupés ensemble, est un réseau où le nombre de pivots est égal au nombre de profils différents dans le réseau. Dans cette optique, nos algorithmes ont pour but d'amener et de maintenir le réseau le plus près possible de cet état lors des restructurations causées par les connexions, déconnexions, mais surtout par les rencontres et recommandations. La figure 3.4 montre deux exemplaires de tels réseaux. Le premier réseau (a) est un réseau qui n'a pas atteint son état de perfection et dans lequel, donc, on retrouve plusieurs pivots pour un même profil. Le second réseau (b) est un réseau parfait dans lequel on ne rencontre qu'un seul nœud pivot par profil. Les deux réseaux respectent cependant les caractéristiques structurales ci-dessus mentionnées.



(a) réseau imparfait



(b) réseau parfait

Figure 3.4 Exemples de réseaux à pivots simples.

Cette figure montre deux réseaux à pivots simples dans lesquels chaque profil est représenté par une couleur différente et où les pivots sont représentés par les nœuds de contour noir. Le premier (a) est un réseau dont les nœuds de même profil sont partiellement regroupés entre eux. Par exemple, on remarque deux regroupements de nœuds jaunes, deux regroupements de nœuds bleu foncé, etc. Le second réseau (b) montre un réseau dit parfait, dans lequel tous les nœuds de même profil sont regroupés ensemble, autour d'un seul pivot pour chaque profil. On y compte 15 regroupements autour de 15 pivots de couleurs différentes.

Plus le réseau tend vers son état de perfection, moins il y a de pivots dans le réseau et plus les parcours de connexion et de socialisation seront courts puisque l'on ne traverse que les nœuds pivots. Évidemment, le nombre de pivots varie aussi en fonction du nombre de profils différents dans le graphe.

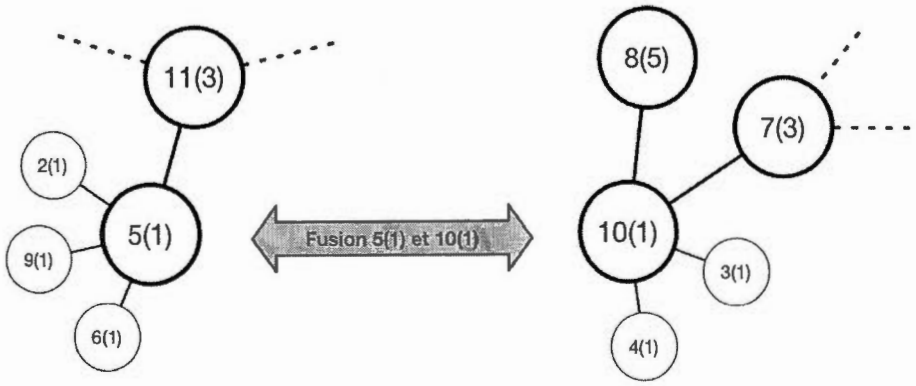
Tous les nœuds pivots ne sont pas nécessairement parcourus lors d'un trajet dans le graphe. En effet, selon l'algorithme de parcours que nous avons choisi (sans retour arrière), on remarque que lorsqu'on visite un pivot qui possède plusieurs liens vers des pivots de profils différents, on doit faire un choix entre les liens que l'on n'a pas encore parcourus, et l'on ne revient pas sur nos pas pour explorer les autres possibilités. Par exemple, dans la figure 3.4 (b), si l'on amorce un parcours sur le pivot vert foncé, on a quatre possibilités pour le prochain nœud à parcourir et celui qu'on choisit détermine la branche qu'on explore. Si l'on choisit le pivot rouge, le parcours s'arrête, si l'on choisit le pivot bleu-gris pâle, le parcours se poursuit sur le pivot bleu, où l'on a un autre choix à faire, et ainsi de suite, jusqu'à ce qu'il n'y ait plus de choix possibles (de voisins non encore visités).

3.3.2.3 Algorithmes d'évolution du réseau

Fusionner

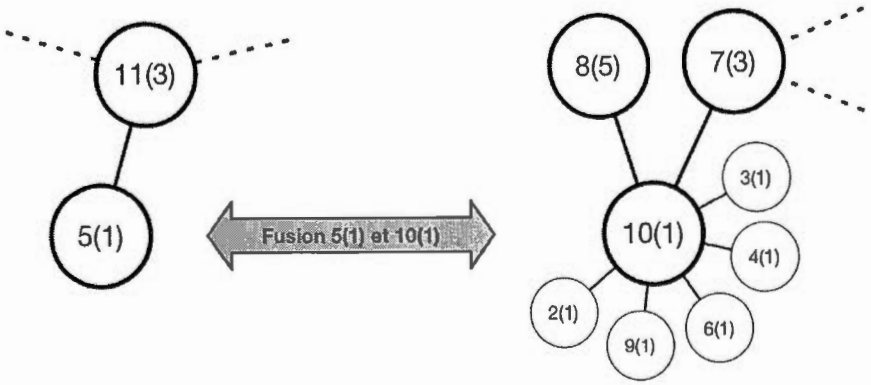
Avant d'aborder les algorithmes généraux pour l'implémentation des règles du modèle (connexion, déconnexion, socialisation et mise à jour des relations), nous allons tout d'abord détailler l'algorithme principal de fusion de deux regroupements. C'est en effet cet algorithme qui est au cœur de la restructuration du réseau causée par une rencontre ou une recommandation, lors d'un parcours de socialisation. L'idée derrière cet algorithme est de prendre deux groupes différents dans le réseau, mais de même profil, pour n'en reformer qu'un seul qui contiendra tous les nœuds de même profil des deux groupes initiaux, et ce, tout en conservant les propriétés structurales imposées. La figure 3.5 illustre les étapes principales de la fusion entre deux sous-groupes de même profil.

(1)



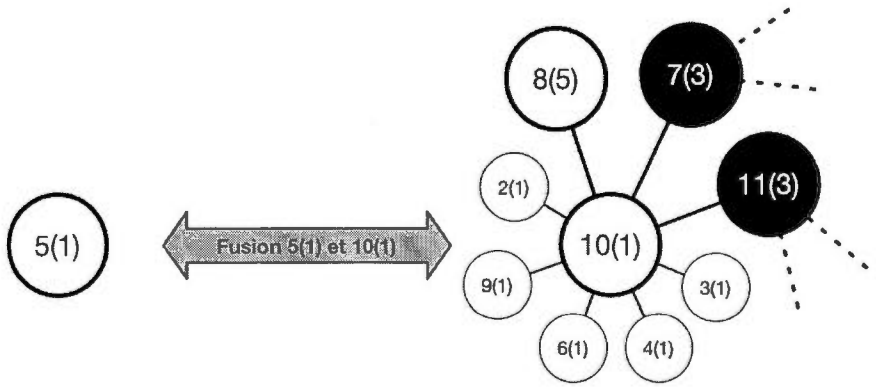
Supposons deux regroupements de profil 1 à fusionner. Le premier sous-groupe est représenté par le pivot 5(1) et le deuxième, par le pivot 10(1). Nous dirons que nous effectuons la fusion de 5(1) avec 10(1).

(2)



La fusion consiste tout d'abord à transférer les voisins simples (non pivots) de même profil du premier pivot sur le deuxième pivot. On transfère donc les noeuds simples 2(1), 9(1) et 6(1) sur le pivot 10(1).

(3)

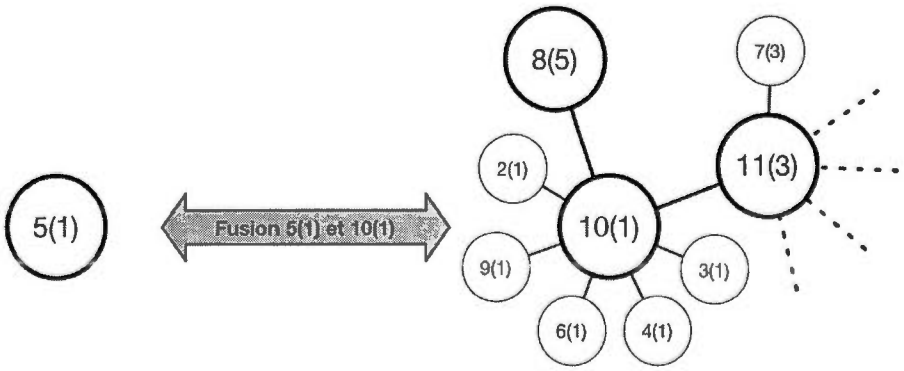


La deuxième étape consiste à transférer les voisins de profils différents sur le pivot du deuxième sous-groupe c.-à-d. sur 10(1). On transfère donc le seul voisin pivot de 5(1), qui est 11(3), sur le pivot 5(1).

Après le transfert d'un voisin pivot de profil différent, il faut toujours vérifier que les pivots qui ont reçu un nouveau lien, 11(3) et 10(1) en l'occurrence, ne se retrouvent pas avec des voisins de même profil entre eux. Dans notre cas, on voit que le pivot 10(1) possède désormais deux voisins pivots qui ont des profils identiques : 7(3) et 11(3). Lorsque ceci se produit, on coupe le lien entre 10(1) et l'un des voisins problématiques, disons 7(3), et l'on effectue de nouveau la fusion de 7(3) vers 11(3), et ainsi de suite, de manière récursive jusqu'à ce qu'on ne trouve plus de voisins de même profil entre eux.

Ce phénomène peut être vu comme une recommandation implicite induite par la rencontre explicite des individus de deux sous-groupes différents de même profil.

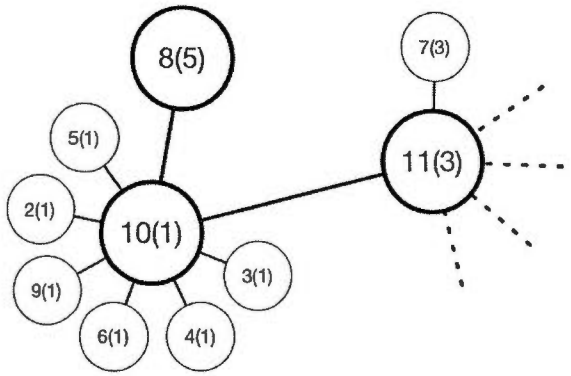
(4)



Au retour de l'appel récursif de la fusion de 7(3) vers 11(3), on voit que ces deux noeuds font maintenant partie du même regroupement. On continue notre fusion initiale.

Le pivot 5(1) n'a plus aucun voisin à transférer. Il devient donc un noeud simple et va se greffer à son nouveau pivot c.-à-d. 10(1). Le graphe suivant illustre le regroupement final après cette fusion.

(5)



Regroupement final après fusion. Tous les noeuds des sous-groupes initiaux font maintenant partie d'un même regroupement et les caractéristiques structurales ont été préservées.

Figure 3.5 Fusion de communautés dans un réseau à pivots simples.

Pour des raisons de performance, on veut minimiser le nombre de transferts de liens. Ainsi, on détermine le sous-groupe qui viendra se greffer sur l'autre en prenant celui qui contient le nœud pivot ayant le plus petit degré (le moins de voisins immédiats). Lorsque les deux pivots sont de même degré, le choix est aléatoire.

Voyons maintenant les algorithmes qui implémentent les règles de notre modèle de socialisation dans ce type de réseaux à pivots simples.

Connexion

La connexion consiste en l'ajout d'un nouveau nœud dans le réseau qui, à partir d'un nœud existant choisi aléatoirement, effectue un parcours dans le graphe. Le trajet se poursuit jusqu'à ce que le nouveau noeud trouve un nœud de profil identique auquel se connecter ou qu'il n'y ait plus de nœuds à parcourir. Si l'on trouve un pivot de profil identique, le nouveau nœud se connecte sur ce pivot en tant que nœud simple. Sinon, le nouveau noeud se connecte au pivot le plus similaire rencontré sur le trajet et devient un pivot pour son profil.

Étant donné la structure particulière de ce graphe dans lequel tous les regroupements sont représentés par des nœuds pivots, il suffit de parcourir les nœuds pivots. Puisqu'on n'a que l'information locale pour choisir le prochain nœud à parcourir, l'heuristique de parcours est

alors de choisir le pivot le plus similaire, qui n'a pas encore été visité lors de ce parcours, comme prochain noeud à parcourir. La figure 3.6 illustre une suite de connexions, selon cet algorithme, à partir d'un réseau vide. Les flèches en traits discontinus partent du nouveau nœud à connecter et pointent vers le nœud choisi aléatoirement pour amorcer le parcours de connexion. Dans le cas où le nœud choisi aléatoirement pour commencer le parcours de connexion est un nœud simple, on trouve d'abord son voisin pivot et l'on continue le parcours sur ce voisin.

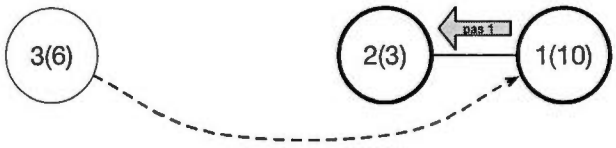
(1)

Connexion du noeud 1(10) dans un réseau vide.
Arrivée du premier noeud dans le réseau, c'est un noeud de profil 10. Ce noeud devient un pivot pour son profil.

(2)

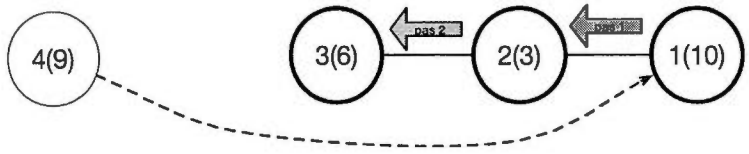
Connexion du noeud 2(3) à partir du noeud aléatoire 1(10).
Le parcours de connexion du nouveau noeud 2(3) s'arrête ici étant donné que le noeud 1(10) n'a aucun voisin. Il se connecte donc directement au noeud 1(10) et devient un pivot pour son profil.

(3)



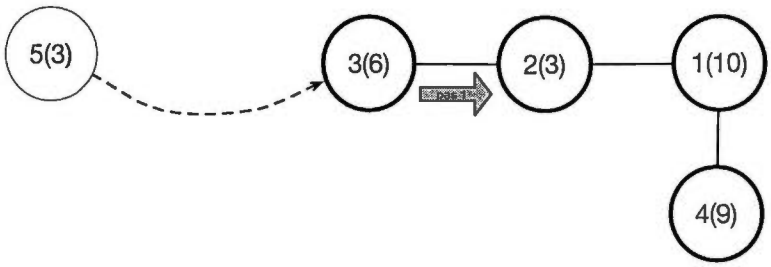
Connexion du noeud 3(6) à partir du noeud aléatoire 1(10).
Le prochain noeud à parcourir est 2(3), c'est le seul choix possible, et l'on arrive en fin de parcours. On n'a pas trouvé de noeud de profil identique donc 3(6) se connecte au noeud le plus similaire rencontré, 2(3), et devient un pivot pour son profil.

(4)



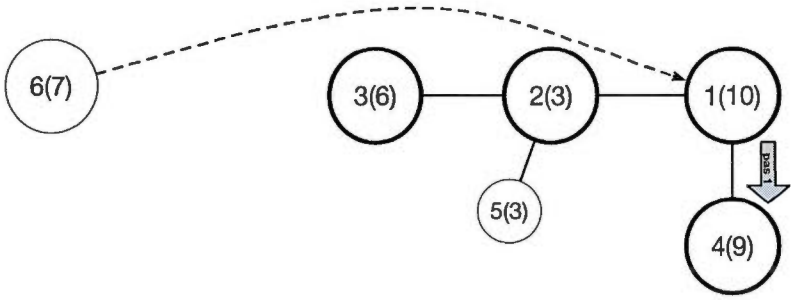
Connexion du noeud 4(9) à partir du noeud aléatoire 1(10).
Le prochain noeud à parcourir est 2(3) puis de 2(3) on passe à 3(6). On n'a pas trouvé de noeud de profil 9 donc le nouveau noeud 4(9) se connecte au noeud le plus similaire rencontré qui est 1(10) et devient un pivot pour son profil.

(5)



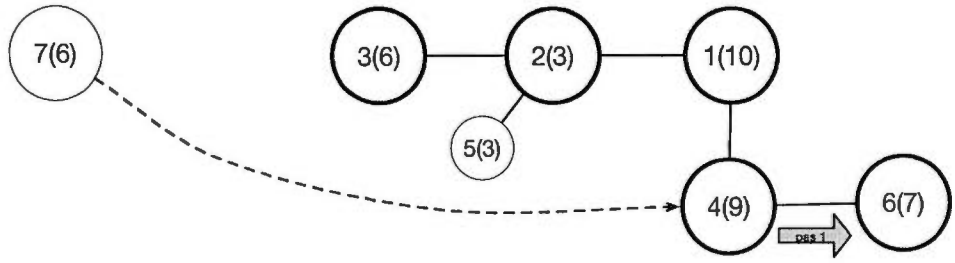
Connexion du noeud 5(3) à partir du noeud aléatoire 3(6).
Le prochain noeud à parcourir est 2(3) puis le parcours s'arrête ici, car on a trouvé un noeud de même profil. 5(3) se connecte donc au pivot 2(3) en tant que noeud simple.

(6)



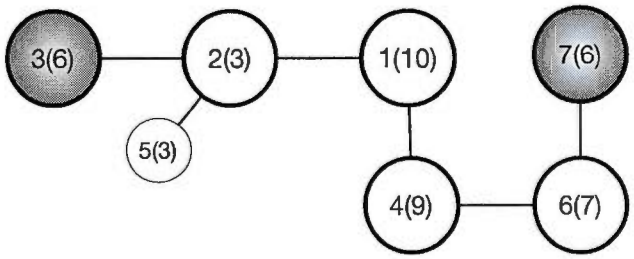
Connexion du noeud 6(7) à partir du noeud aléatoire 1(10).
Ici, on a le choix entre deux noeuds pour continuer le parcours : 2(3) ou 4(9). On choisit 4(9) puisque c'est le noeud de profil le plus similaire à 6(7) et l'on arrive à la fin du parcours. On n'a pas trouvé de noeud de profil 7 donc le nouveau noeud 6(7) se connecte au noeud rencontré le plus similaire qui est 4(9), et devient un pivot pour son profil.

(7)



Connexion du noeud 7(6) à partir du noeud aléatoire 4(9).
On a le choix entre deux noeuds pour continuer le parcours : 1(10) ou 6(7). On choisit 6(7) puisque c'est le noeud de profil le plus similaire à 7(6) et l'on arrive à la fin du parcours. On n'a pas trouvé de noeud de profil 6 donc le nouveau noeud 7(6) se connecte au noeud rencontré le plus similaire qui est 6(7) et devient un pivot pour son profil.

(8)



Fin des connexions successives.
Réseau résultant des sept connexions précédentes. On remarque deux noeuds de même profil, 3(6) et 7(6), qui n'ont pas été regroupés ensemble.

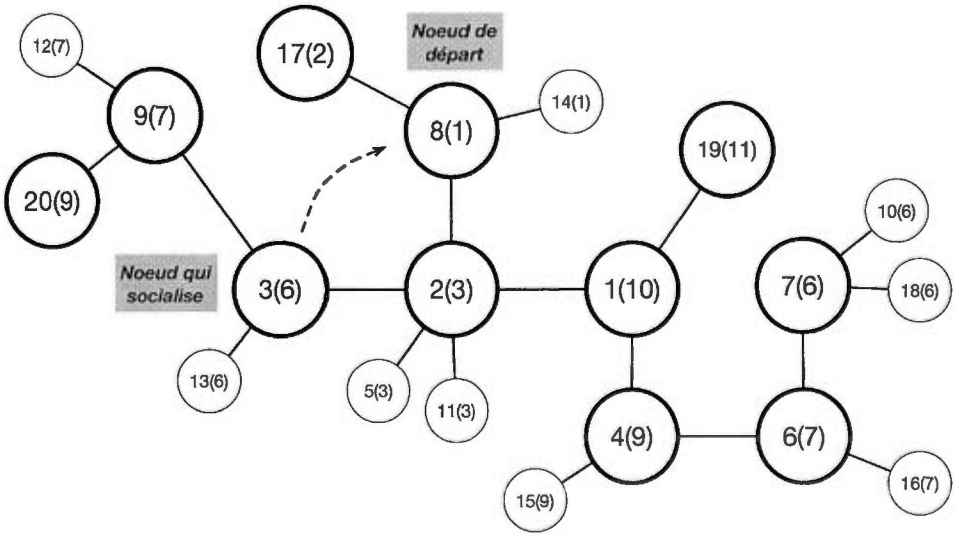
Figure 3.6 Connexions successives dans un réseau à pivots simples.

La figure 3.6 montre que notre parcours de connexion, bien que favorisant parfois les regroupements, ne suffit pas à rassembler complètement tous les noeuds de même profil. Le réseau montre en effet deux noeuds de profil 6 qui ne sont pas regroupés ensemble. En ajoutant d'autres nœuds au réseau, de cette façon, on obtiendra un réseau semblable à celui de la figure 3.4 (a) dans laquelle on retrouve plusieurs regroupements dispersés pour un même profil. En fait, le réseau de la figure 3.4 (a) a été simplement construit, de cette manière, par connexions successives. Nous verrons dans ce qui suit que les mécanismes de rencontres aléatoires et de recommandations, lors des parcours de socialisation, ont comme effet de joindre ces communautés de même profil qui sont éparpillées dans le réseau global.

Parcours de socialisation – rencontres et recommandations

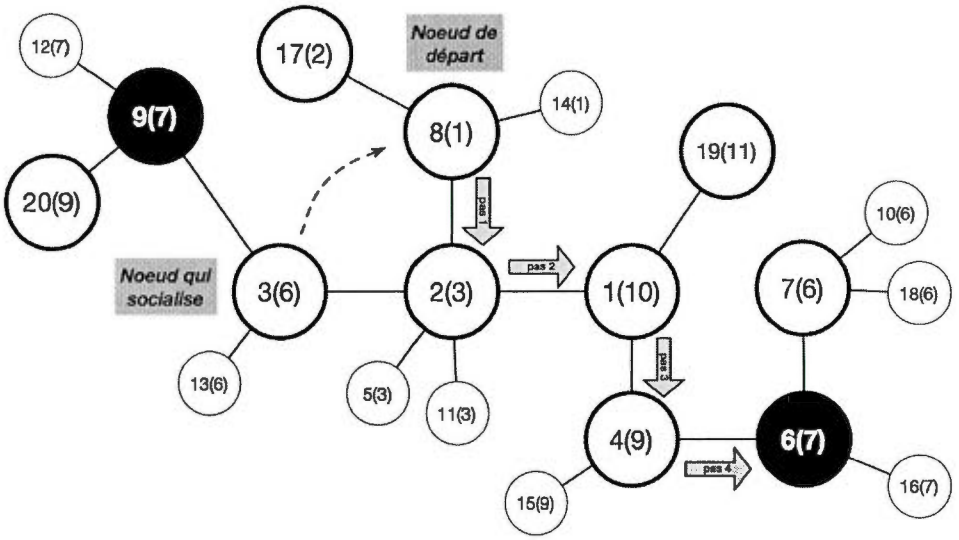
La socialisation consiste en un nœud qui socialise, qu'on nommera le socialisateur, qui effectue un parcours de socialisation dans le graphe et d'un nœud choisi aléatoirement qui sera le point de départ du parcours de socialisation. On utilise la même heuristique de parcours que pour les connexions, mais lorsqu'on rencontre un nœud de même profil que celui du socialisateur (rencontre aléatoire) ou de profil identique à l'un des voisins pivots immédiats du socialisateur (recommandation), on effectue alors une restructuration du réseau pour regrouper les deux sous-réseaux de même profil. Aussi, contrairement au cas de la connexion, un parcours de socialisation se poursuit toujours jusqu'à la fin du trajet, c.-à-d. lorsqu'il n'y a plus de possibilités pour le choix du nœud suivant. La figure 3.7 illustre un exemple de parcours de socialisation, selon cet algorithme, avec lequel on montre l'effet des fusions effectuées lors des rencontres et des recommandations.

(1)



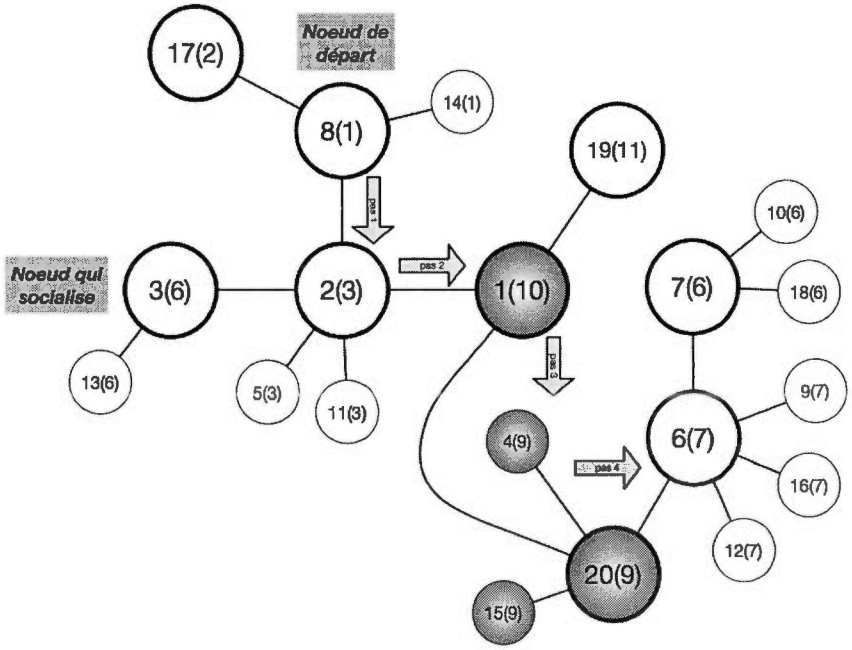
Socialisation de 3(6) à partir du noeud aléatoire 8(1).

(2)



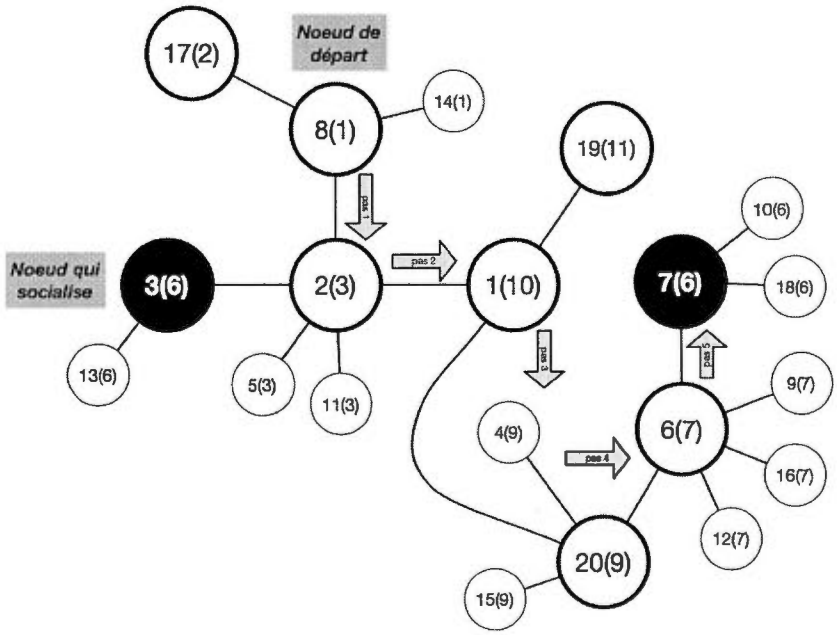
Socialisation de 3(6) à partir du noeud aléatoire 8(1) – recommandation.
De 8(1), on parcourt le graphe en choisissant toujours le noeud de profil le plus similaire au socialisateur 3(6). On visite donc les noeuds 8(1) – 2(3) – 1(10) – 4(9) – 6(7) et avant de terminer le parcours, on fait une **recommandation entre 9(7)** – voisin immédiat de 3(6) – **et 6(7)** – le noeud rencontré. On effectue donc la fusion du pivot 6(7) vers le pivot 9(7) pour rassembler les deux sous-groupes.

(3)



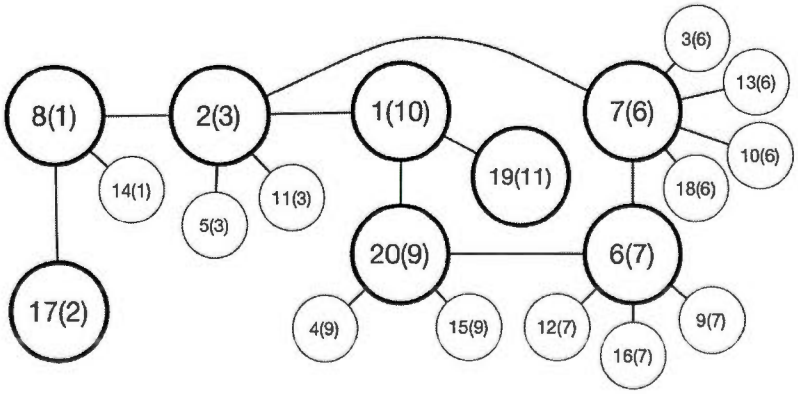
Socialisation de 3(6) à partir du noeud aléatoire 8(1) – suite.
Après la fusion de 6(7) vers 9(7) (et toutes les fusions récursives), on obtient le réseau ci-dessus. La recommandation est terminée et tous les noeuds de profil 7, qui se trouvaient auparavant dans des sous-groupes séparés, n'en forment plus qu'un seul.

(4)



Socialisation de 3(6) à partir du noeud aléatoire 8(1) – rencontre.
On continue le parcours en passant au noeud suivant, 7(6), avec lequel on fait une **rencontre.** On effectue la **fusion de 7(6) et 3(6)** pour réunir les deux sous-réseaux.

(5)



Socialisation de 3(6) à partir du noeud aléatoire 8(1) – fin.
Après cette rencontre, le parcours de socialisation se termine car il n'y a plus de noeuds (non déjà visités) à parcourir. Les deux sous-réseaux de profil 7 du réseau initial ne forment maintenant qu'un seul sous-groupe et il en va de même pour les deux sous-réseaux de profil 6 qu'on retrouvait aussi dans le réseau initial.

Figure 3.7 Rencontre et recommandation dans un réseau à pivots simples.

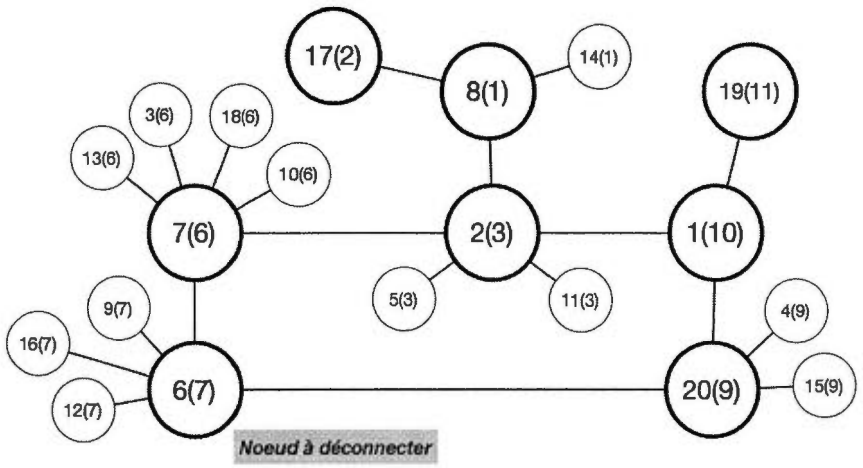
La figure 3.7 illustre un exemple possible de parcours de socialisation et montre à quel point la socialisation d'un seul nœud profite à plusieurs. En effet, plusieurs nœuds qui ne socialisaient pas se retrouvent maintenant dans un voisinage augmenté de nouveaux contacts et donc potentiellement plus riche. Bien entendu, un nœud qui socialise n'effectue pas toujours un parcours aussi fructueux et il arrive qu'on ne fasse ni rencontre ni recommandation, mais d'un point de vue global, les rencontres et recommandations réalisées s'avèrent efficaces au niveau de toute la collectivité.

Nous avons montré un exemple de socialisation d'un noeud pivot. Lorsque c'est un nœud simple qui décide de socialiser, on fait tout simplement comme si c'était son pivot immédiat qui socialisait.

Déconnexion

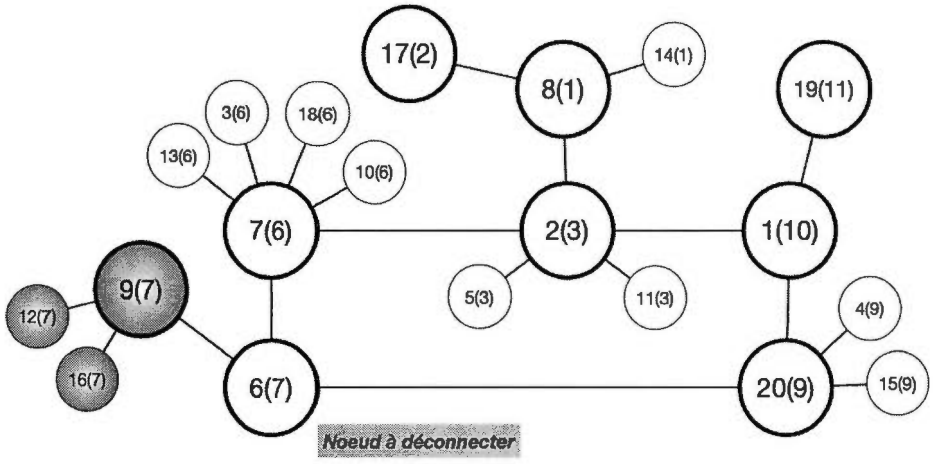
Lors d'une déconnexion, on désire évidemment conserver les propriétés structurales du réseau, mais aussi éviter de défaire les regroupements de mêmes profils qui y sont déjà présents. La déconnexion d'un nœud simple ou d'un pivot qui n'a qu'un seul voisin (qui est obligatoirement un autre pivot sinon le graphe ne serait pas connexe) est triviale puisque ces types de nœuds ne possèdent qu'un seul lien. On les supprime tout simplement du réseau. Pour les autres cas, l'idée générale derrière la déconnexion d'un pivot est de transférer tous ses voisins vers un autre nœud avant de le supprimer du graphe. La figure 3.8 illustre les deux cas de figure pour la déconnexion d'un nœud pivot.

(1)



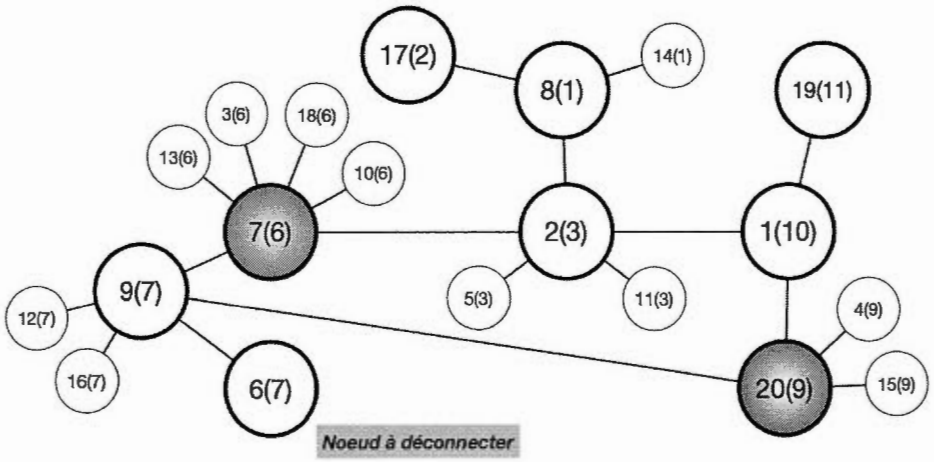
Cas 1 : Le pivot à déconnecter a des voisins pivots et des voisins simples de même profil. Déconnexion du pivot 6(7) .

(2)



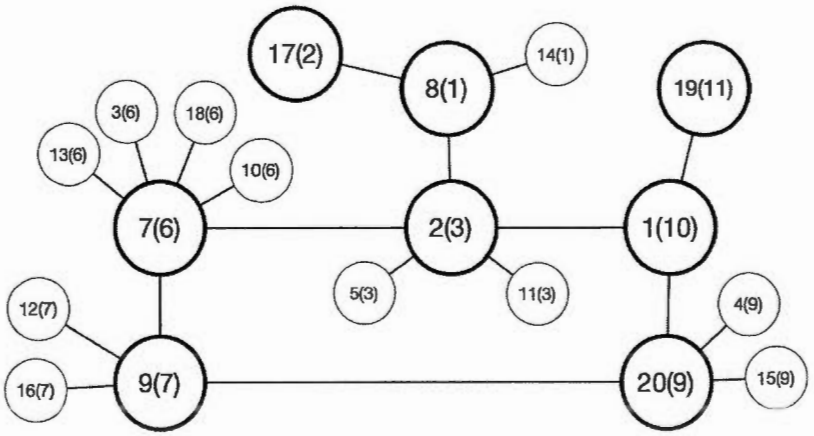
Déconnexion du pivot 6(7) – cas 1 (suite).
On choisit d'abord aléatoirement un voisin simple de même profil du noeud à déconnecter, par exemple le voisin 9(7), et l'on en fait le nouveau pivot pour ce regroupement. On transfère ensuite tous les autres voisins simples de 6(7), c.-à-d. les voisins 16(7) et 12(7) sur le nouveau pivot 9(7).

(3)



Déconnexion du pivot 6(7) – cas 1 (suite).
On transfère ensuite tous les voisins pivots de 6(7) qui sont de profils différents, c.-à-d. 7(6) et 20(9) sur le nouveau pivot 9(7). Comme dans le cas de la fusion, il faut toujours vérifier, lors du transfert d'un voisin pivot, si les deux noeuds qui reçoivent un nouveau lien se retrouvent avec des voisins de même profil entre eux. Si oui, on doit les fusionner. Dans cet exemple, ce n'est pas le cas : les noeuds 9(7), 7(6) et 20(9) n'ont pas de voisins de même profil entre eux après le transfert.

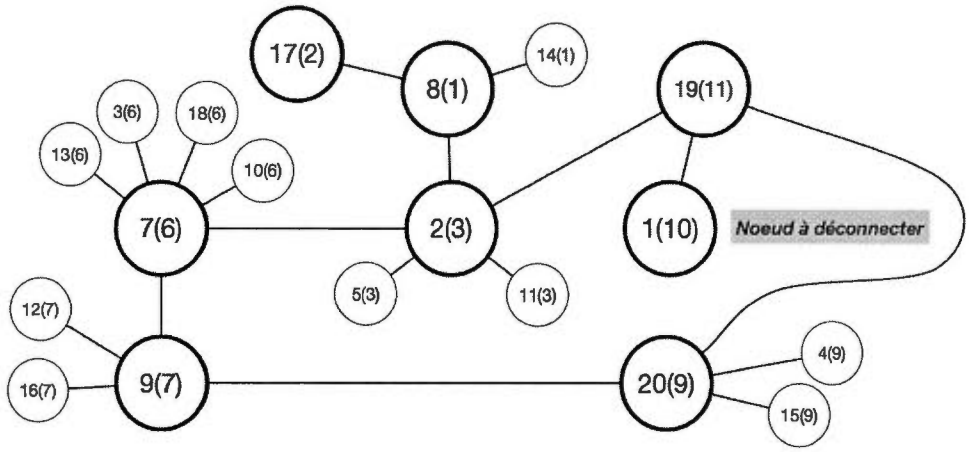
(4)



Déconnexion du pivot 6(7) – cas 1 (fin).
Après le transfert de tous les voisins, on supprime le noeud 6(7) du réseau. La déconnexion est terminée et les regroupements sont restés intacts.

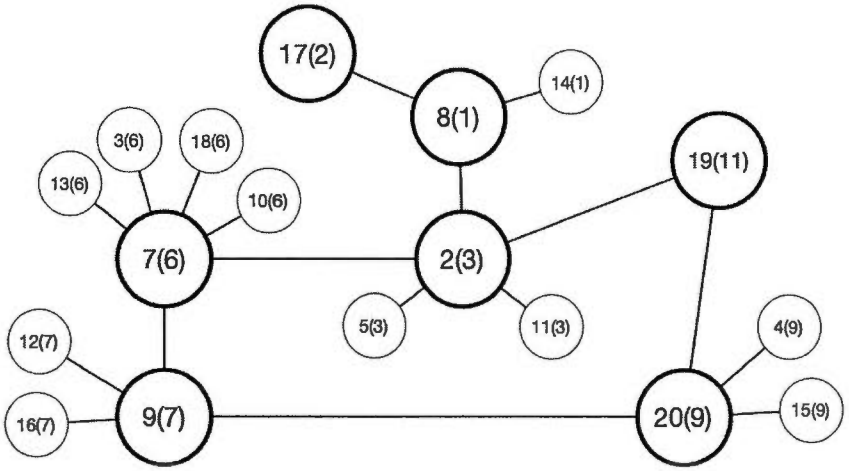
Cas 2 : Le pivot à déconnecter n'a que des voisins pivots (de profils différents).
Déconnexion du pivot 1(10) dont les voisins pivots sont 19(11), 20(9) et 2(3).
On choisit, parmi les voisins pivots, celui qui est le plus similaire au pivot à déconnecter. Dans ce cas, on fait un choix arbitraire entre 19(11) et 20(9) puisque la similarité (telle que nous la calculons) entre 19(11) et 1(10) est égale à la similarité entre 20(9) et 1(10). Disons que l'on choisit le pivot 19(11).

(5)



Déconnexion du pivot 1(10) – cas 2 (suite).
On transfère ensuite les autres voisins pivots de 1(10) c.-à-d. 2(3) et 20(9) sur le pivot 19(11) en vérifiant toujours si le transfert crée des voisins de même profil entre eux. Dans l'affirmative, on les fusionne. Ici, ce n'est pas le cas.

(6)



Déconnexion du pivot 1(10) – cas 2 (fin).
Après le transfert, on supprime le pivot 1(10) du réseau. La déconnexion est terminée et les regroupements sont restés intacts.

Figure 3.8 Déconnexions de noeuds pivots dans un réseau à pivots simples.

Mise à jour du profil

Lorsque le profil d'un nœud change ou que pour une raison ou une autre, il se retrouve dans un regroupement qui ne lui convient plus, il doit pouvoir se repositionner dans le réseau pour changer de communauté. Pour ce faire, le noeud effectue tout simplement une déconnexion, qui a pour effet de le couper de son ancien réseau local pour ensuite se connecter de nouveau au réseau, en quête de nœuds similaires ou mieux, identiques.

3.3.3 Regroupements autour de noeuds pivots chaînés

Le modèle à pivots simples que nous venons d'expliquer fonctionne bien en terme de regroupements et de parcours, mais génère des réseaux dans lesquels un petit nombre de nœuds peut se retrouver avec une multitude de liens tandis que le reste n'en aura qu'un seul. Imaginez par exemple un réseau de ce type formé de 500 nœuds dans lequel on a dix profils différents. Dans sa forme parfaite, ce réseau contiendrait dix pivots ayant en moyenne 50 liens et 450 nœuds simples n'en possédant qu'un seul.

Par analogie avec nos réseaux sociaux réels, on comprend tout de suite qu'un individu ne peut pas véritablement posséder un nombre infini de contacts ou amis. L'entretien d'une relation demande une certaine disponibilité qui, bien que variant sûrement d'un individu à l'autre, n'est certainement pas illimitée. D'autre part, d'un point de vue fonctionnel, cette façon de faire poserait un problème de performance si l'on implémentait concrètement ce modèle dans une application distribuée générant et entretenant un réseau pair-à-pair, par exemple. Dans un tel réseau, où chaque nœud est un ordinateur (qui peut être un client et un serveur), certains d'entre eux, les nœuds pivots, seraient constamment sollicités tandis que les autres, les nœuds simples, presque jamais. Comme dans le cas des individus, la disponibilité de chaque ordinateur ainsi que les ressources qui y sont associées, comme la bande passante par exemple, ne sont jamais illimitées, bien que variant d'un ordinateur à l'autre.

Dans cette optique, nous proposons donc un autre type de regroupements autour de nœuds pivots, mais qui tiendra compte de la capacité limitée de chaque individu du réseau à

entretenir des relations. Dans ce modèle, chaque nœud possède une capacité c'est-à-dire, un nombre maximum de liens qu'il peut supporter. Par conséquent, un même regroupement pourra posséder plusieurs pivots qui seront chaînés les uns aux autres. Sur chacun de ces pivots chaînés pourront alors être connectés des voisins simples de même profil, des voisins pivots de profils différents et, à la différence du premier modèle, au maximum deux pivots de profil identique (le suivant et le précédent du pivot dans la chaîne).

Pour ce faire, lorsqu'un pivot aura atteint son nombre maximum de liens (égal à sa capacité), on choisira alors, lorsque possible, un voisin simple de même profil pour en faire un autre pivot pour ce regroupement. Cet autre pivot deviendra alors le suivant du premier pivot et le premier pivot deviendra le précédent du second pivot. C'est ce nouveau pivot qui pourra dès lors accueillir de nouveaux venus de même profil au sein de la communauté.

3.3.3.1 Attributs des nœuds et définitions

Pour gérer ces nouveaux types de regroupements, les nœuds devront fournir un peu plus d'informations que dans notre modèle précédent. On doit toujours pouvoir déterminer si un nœud est un pivot ou non, mais aussi, chaque pivot doit maintenant connaître son pivot précédent ainsi que son pivot suivant. De plus, chaque nœud possède un attribut qui indique le nombre maximum de liens qu'il peut entretenir.

Attribut des nœuds :

`profil` :	Contient le profil d'intérêts du nœud – un nombre entier pour le moment.
`estPivot` :	Indique si le nœud est un pivot ou non (valeur booléenne).
`suivant` :	Indique le pivot suivant dans la chaîne de pivots ou a la valeur "nul" s'il n'y a aucun suivant.
`précédent` :	Indique le pivot précédent dans la chaîne de pivots ou bien la valeur "nul" s'il n'y a aucun précédent.
`capacité` :	Indique le nombre maximum de liens que peut avoir un nœud.

Définitions :

Pivot :	Nœud qui a été désigné comme étant un pivot (`estPivot` a la valeur "vrai").
Pivot initial :	Nœud pivot qui n'a pas de précédent (`précédent` a la valeur "nul") mais qui peut avoir un suivant. C'est le pivot en début de chaîne.
Pivot final :	Nœud pivot qui n'a pas de précédent (`précédent` a la valeur "nul") mais qui peut avoir un suivant. C'est le pivot en début de chaîne.
Pivot intermédiaire :	Nœud pivot qui a un suivant et un précédent. C'est un pivot en milieu de chaîne.
Pivot immédiat :	Le pivot sur lequel est attaché un nœud simple. On dira que ce pivot est le pivot immédiat de ce nœud simple.
Nœud plein :	Nœud dont le nombre de liens (son degré) a atteint la valeur de son attribut `capacité`.
Nœud libre :	Nœud dont le nombre de liens (son degré) est plus petit que la valeur de son attribut `capacité`.
Chaîne pleine :	Une chaîne de pivots dont tous les pivots (sauf le pivot final) sont pleins.
Chaîne libre :	Une chaîne de pivots dont au moins un des pivots (sauf le pivot final) est libre.

3.3.3.2 Propriétés structurales à maintenir dans le réseau

Regroupements

La figure 3.9 montre un regroupement par pivots chaînés. Les traits discontinus indiquent des liens vers des pivots de profils différents et les liens fléchés relient un pivot à son pivot suivant dans la chaîne de pivots. Le pivot 14(6) est le pivot initial, les pivots 7(6) et 1(6) sont des pivots intermédiaires et le pivot 5(6) est le pivot final de cette chaîne de pivots qui contient quatre pivots. On dira que 7(6) est le suivant de 14(6) et que 14(6) est le précédent de 7(6), que 1(6) est le suivant de 7(6) et que 7(6) est le précédent de 1(6), etc.

Dans ce modèle de regroupements, un nœud doit avoir une capacité au moins égale à 3 : un nœud, lorsqu'il devient un pivot, doit pouvoir posséder au minimum un lien vers un

précédent, un lien vers un suivant et un autre lien vers un pivot de profil différent (pour conserver la connexité du graphe).

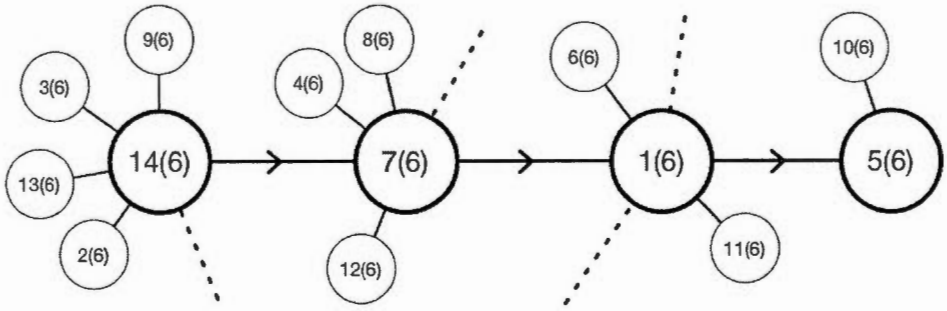


Figure 3.9 Regroupement autour de nœuds pivots chaînés.

On remarque aussi que, selon ces caractéristiques structurales, un nœud, peu importe lequel, peut toujours trouver tous ces voisins immédiats et indirects de manière optimale et sans avoir à comparer son profil. Il n'a qu'à parcourir la chaîne de pivots de suivant en suivant et de précédent en précédent. Lorsqu'il traverse un des pivots de la chaîne, il peut atteindre les voisins simples reliés à ce pivot en un seul pas, comme dans le modèle précédent.

De plus, les parcours dans une telle structure peuvent encore se limiter aux nœuds pivots. Par contre, le nombre des nœuds pivots risque d'être plus élevé que dans notre modèle précédent. Pour limiter le chaînage excessif des nœuds pivots, nous introduisons une autre propriété structurale à maintenir : tous les pivots précédents doivent être pleins (chaîne pleine). De cette manière, la longueur des chaînes de pivots est toujours minimale. Par exemple, dans la figure 3.9, les pivots 14(6), 7(6) et 1(6) doivent être pleins, c.-à-d. que leur degré doit être égal à leur capacité.

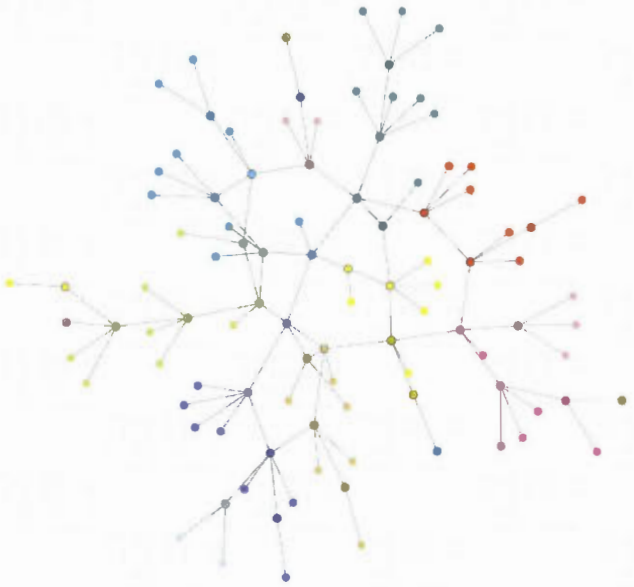
Nos algorithmes d'évolution du réseau doivent donc générer et maintenir un réseau qui a les propriétés structurales suivantes :

- Le réseau est connexe.

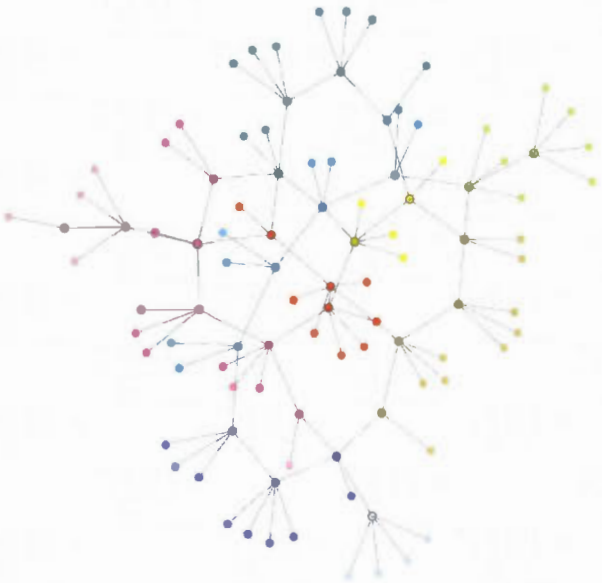
- Le réseau contient des nœuds pivots qui sont soit reliés à des nœuds simples de même profil, soit reliés à des nœuds pivots de profils différents, soit reliés à au plus deux pivots de même profil c.-à-d. le suivant et/ou le précédent.
- Le réseau contient des nœuds simples qui n'ont qu'un seul lien avec un nœud pivot de même profil.
- Tous les pivots qui sont les précédents d'un autre pivot sont toujours pleins.
- Le degré de chaque nœud ne dépasse jamais sa capacité.
- Les voisins pivots de profils différents (donc ni le suivant ni le précédent) attachés à un pivot ne sont jamais de même profil entre eux.

Un réseau de ce type sera dit "parfait" lorsque tous les nœuds de même profil sont regroupés ensemble sur une seule chaîne dans le réseau. Le réseau contient exactement autant de chaînes que de profils différents présents dans le réseau. De ce point de vue, les mécanismes de connexion, de déconnexion, de mise à jour des relations et surtout de socialisation ont comme objectif de réaliser au possible cet état de perfection. La figure 3.10 illustre deux réseaux à pivots chaînés qui respectent les propriétés structurales mentionnées ci-dessus. Le premier (a) est un réseau imparfait comportant plusieurs chaînes pour un même profil. Par exemple, on y trouve deux chaînes de nœuds turquoise, deux chaînes de nœuds bleus, etc. Le second (b) est un réseau ayant atteint son état de perfection dans lequel on ne retrouve qu'un seul regroupement par profil (dix chaînes pour dix profils différents).

Comme dans le premier modèle, plus le réseau tend vers son état de perfection, moins il y a de pivots dans le réseau et plus les parcours de connexion et de socialisation seront courts puisque l'on ne parcourt que des nœuds pivots. Évidemment, le nombre de pivots variera aussi en fonction du nombre de profils différents dans le graphe.



(a) réseau imparfait



(b) réseau parfait

Figure 3.10 Exemplaires de réseaux à pivots chaînés.

Chaque profil est représenté par une couleur différente. Les pivots sont représentés par les nœuds de contour noir. Le réseau (a) montre un réseau imparfait dans lequel on observe plus d'une chaîne de même couleur, par exemple, deux chaînes jaunes, deux chaînes turquoise, etc. Le réseau (b) illustre un réseau parfait dans lequel tous les nœuds de même couleur se trouvent sur la même chaîne dans le réseau.

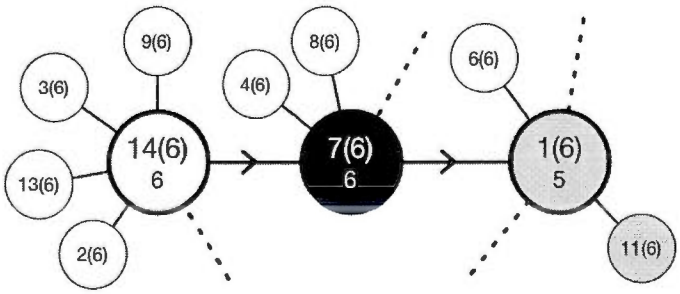
3.3.2.3 Algorithmes d'évolution du réseau

L'introduction du chaînage ainsi que de la contrainte sur la longueur minimale des chaînes de pivots complexifie évidemment les algorithmes du modèle qui doivent maintenant conserver des propriétés structurales plus sévères lors des restructurations du réseau. Avant d'aborder les algorithmes qui implémentent les règles du modèle, nous allons tout d'abord détailler quelques algorithmes importants qui seront à leur tour utilisés dans l'implémentation des règles générales (connexion, déconnexion, socialisation, mise à jour des relations).

Remplir une chaîne

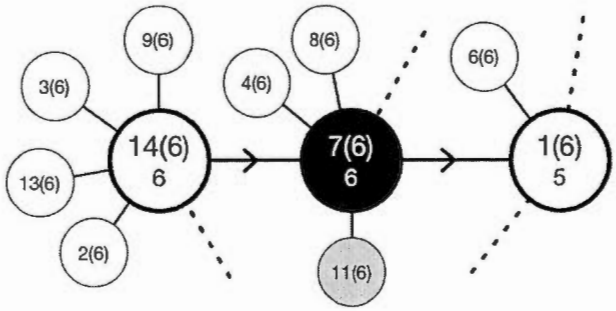
Advenant le cas où un pivot initial avec suivant ou un pivot intermédiaire deviendrait libre (degré < capacité) au cours d'une restructuration, on doit le remplir pour maintenir la contrainte que tous les nœuds qui sont des précédents (qui ont donc un suivant) doivent être pleins. Pour ce faire, on transfère les voisins du pivot en fin de chaîne ou le pivot final lui-même, s'il n'a aucun voisin, vers le pivot à remplir. La figure 3.11 illustre les cas possibles pour remplir une chaîne libre. L'étiquette de chaque nœud indique le numéro d'identification unique du nœud suivi de son profil, entre parenthèses puis juste en dessous, un troisième nombre indique la capacité de ce nœud (qui est illustrée ici pour les nœuds pivots seulement).

(1)



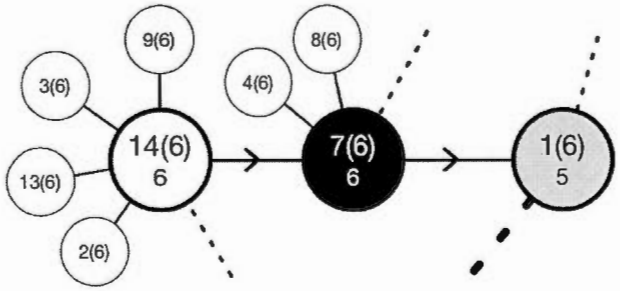
Cas 1 : Le pivot en fin de chaîne possède des voisins simples de même profil.
Remplissage du noeud 7(6).
On remarque ici que le noeud intermédiaire 7(6), de capacité égale à 6, ne possède que 5 liens. Puisque le pivot final, 1(6), a des voisins simples de même profil, on prend l'un de ces voisins, disons 11(6), et on le transfère tout simplement sur 7(6).

(2)



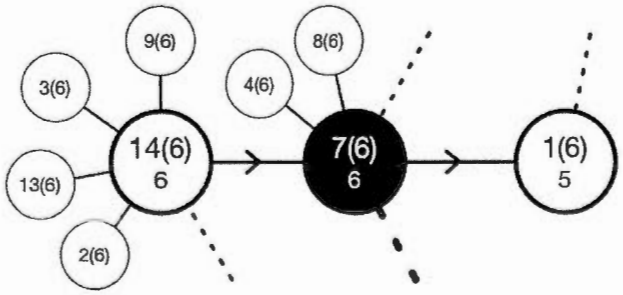
Remplissage du noeud 7(6) – cas 1 (fin).
Le noeud 7(6) est maintenant plein et la chaîne est de longueur minimale.

(3)



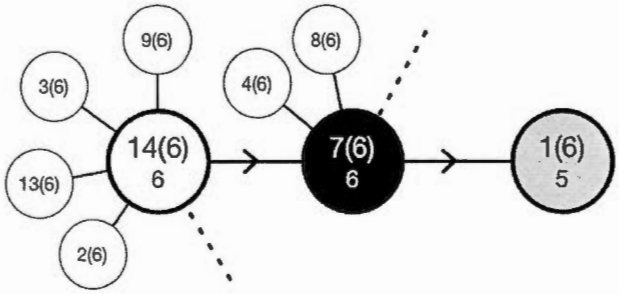
Cas 2 : Le pivot en fin de chaîne ne possède que des voisins de profils différents (illustrés ici par les traits discontinus).
Remplissage du noeud 7(6).
Dans ce cas, puisque le pivot de fin de chaîne n'a aucun voisin simple de même profil, on transfère un de ses voisins de profil différent sur le noeud à remplir, 7(6).

(4)



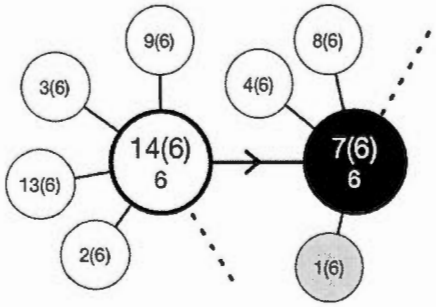
Remplissage du noeud 7(6) – cas 2 (fin).
Le noeud 7(6) est maintenant plein. Il est à noter que, comme dans le cas du premier modèle, chaque fois qu'on transfère un noeud pivot de profil différent sur un autre pivot, on doit vérifier que les deux noeuds ayant reçu ce nouveau lien ne se retrouvent pas avec des voisins de même profil entre eux. Dans l'affirmative, on doit les fusionner. Nous verrons la fusion pour ce modèle un peu plus loin dans cette section (voir fig. 3.13).

(5)



Cas 3 : Le pivot en fin de chaîne n'a aucun voisin.
Remplissage du noeud 7(6).
Comme le pivot en fin de chaîne n'a plus aucun voisin, on en fait un noeud simple (qui n'a ni suivant ni précédent) et le noeud 7(6) devient le nouveau pivot de fin de chaîne qui n'a plus de suivant.

(6)



Remplissage du noeud 7(6) – cas 3 (fin).
Le noeud 7(6) n'est pas plein, mais il est devenu le pivot final de la chaîne (n'est donc le précédent d'aucun pivot et n'a pas à être plein). La chaîne initiale est devenue plus courte, de longueur minimale.

Figure 3.11 Remplissage de chaînes libres dans un réseau à pivots chaînés.

Cet algorithme peut être rappelé récursivement sur chaque pivot jusqu'à ce que la chaîne soit pleine. Pour alléger la description des algorithmes qui suivent, nous ne mentionnons pas les appels de cet algorithme et tenons pour acquis qu'il s'exécute chaque fois que c'est nécessaire.

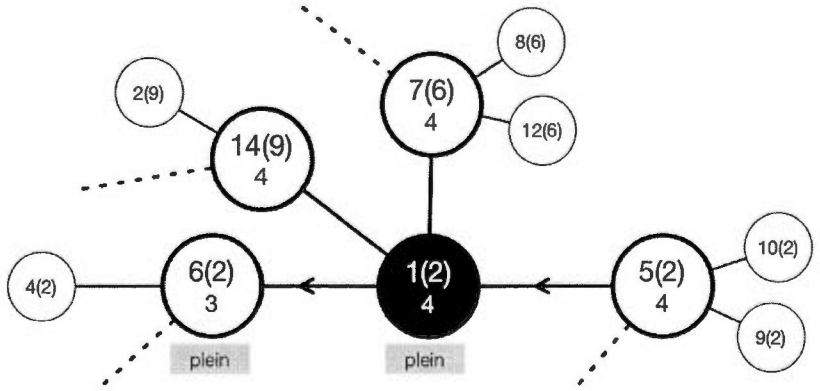
Libérer une place

Un autre mécanisme dont nous aurons besoin pour implémenter le modèle tout en maintenant les propriétés structurales du réseau est un algorithme servant à libérer ou créer une place libre sur un nœud pivot (ou un autre noeud de même profil faisant partie du même

regroupement), d'un profil donné, pour que celui-ci puisse recevoir une connexion (d'un autre nœud de ce même profil à greffer à ce regroupement). L'important, ici, est de trouver un pivot libre de même profil qui fait partie du même regroupement que le nœud à libérer.

Pour libérer une place sur un pivot plein (qui n'est pas en fin de chaîne), on regarde si le pivot en fin de chaîne est libre. Si oui, le cas est trivial, on utilisera simplement ce pivot final libre pour recevoir la connexion. Si, par contre, le pivot final est plein, deux cas peuvent se présenter : 1) le pivot final possède au moins un voisin de même profil ou 2) le pivot final n'a que des voisins pivots de différents profils (on ignore ici le précédent). Dans le premier cas, on crée une nouvelle place sur la chaîne en transformant un voisin simple de même profil en un nouveau pivot de la chaîne et en utilisant ce nouveau pivot pour la connexion. Dans le deuxième cas, on libère une place sur le pivot final en transférant un de ses voisins de profil différent vers un autre de ses voisins de profil différent. La figure 3.12 illustre ces deux cas typiques lors de la libération d'une place sur un pivot.

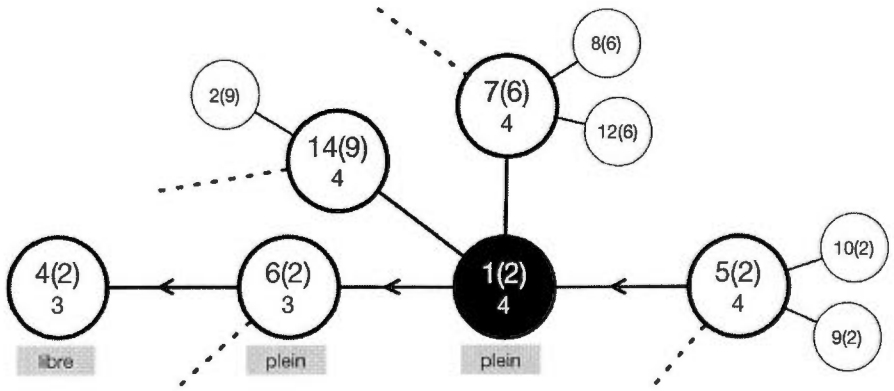
(1)



Cas 1 : Le pivot final est plein, mais possède des voisins simples de même profil. Libération d'une place sur le pivot 1(2).
Le noeud 1(2) est plein. On cherche le pivot de fin de chaîne (en suivant les flèches) et l'on trouve le pivot 6(2). 6(2) est plein, mais il possède un voisin de même profil 4(2). On prend donc ce voisin pour en faire un nouveau pivot pour ce regroupement qui devient donc le noeud libre pouvant maintenant recevoir une connexion.

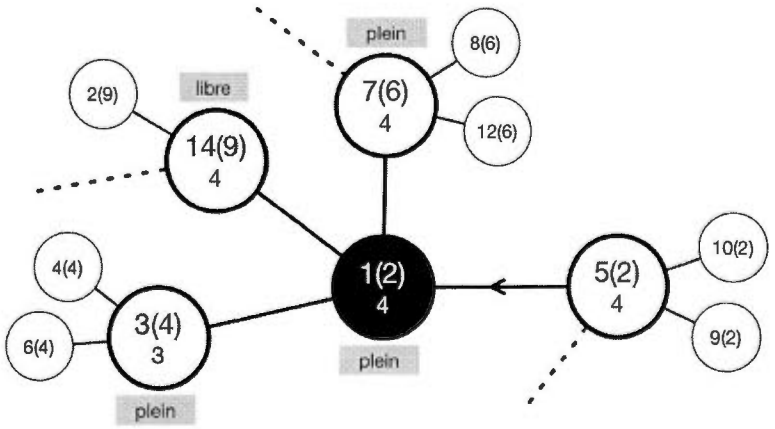
Notons que lorsqu'on a plusieurs voisins simples pour le choix du nouveau pivot, on choisit celui ayant la plus grande capacité.

(2)



Libération d'une place sur le pivot 1(2) - cas 1 (fin)
On a créé une place libre sur la chaîne de 1(2) qui se trouve sur le nouveau pivot de fin de chaîne, 4(2).

(3)

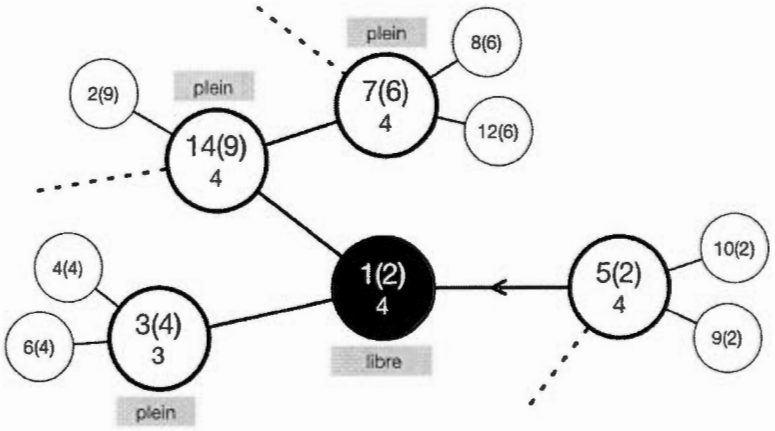


Cas 2 : Le pivot final est plein et n'a que des voisins pivots de profils différents.
Libération d'une place sur le noeud 1(2).
Le pivot 1(2) est déjà le pivot de fin de chaîne, mais n'a aucun voisin de même profil pour chaîner un pivot supplémentaire. On cherche alors un voisin de profil différent qui est libre, par exemple, le pivot 14(9) pour lui transférer un autre voisin de profil différent, par exemple, 7(6).

Si, parmi les voisins pivots de profils différents du noeud 1(2) on ne trouve pas de voisin libre sur lequel transférer un autre voisin, on rappelle récursivement l'algorithme de libération d'une place sur un de ces voisins.

Encore une fois, chaque fois qu'on assigne un nouveau lien entre deux pivots, il faut vérifier si cela crée des voisins de même profil entre eux et si oui, on les fusionne (voir figure 3.13 pour la fusion dans un réseau à pivots chaînés).

(4)



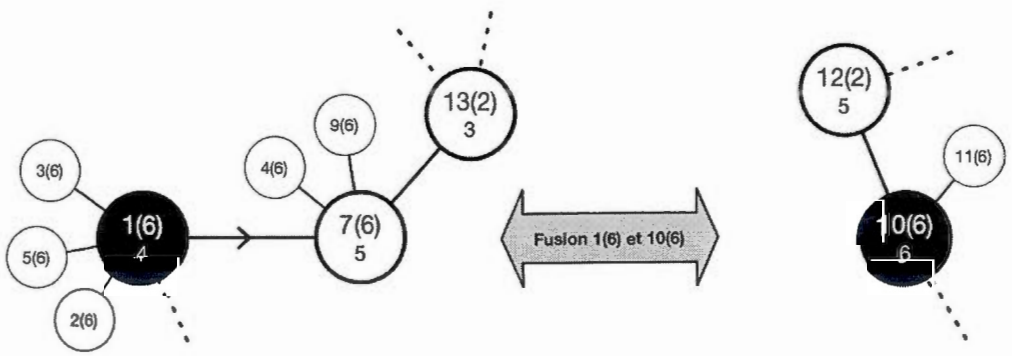
Libération d'une place sur le noeud 1(2) - cas 2 (fin).
On a supprimé le lien entre 7(6) et 1(2) et l'on a ajouté un lien entre 7(6) et 14(9). Le noeud 1(2) est maintenant libre et peut recevoir une connexion.

Figure 3.12 Libération de places dans un réseau à pivots chaînés.

Fusionner

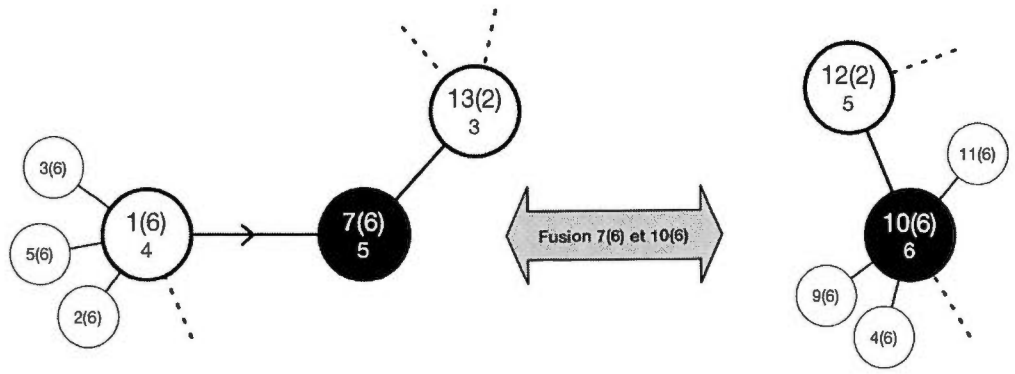
La fusion demeure l'algorithme principal lorsqu'il s'agit d'effectuer les restructurations induites par les rencontres et les recommandations. Comme dans notre premier modèle, l'idée est de fusionner deux regroupements de même profil pour qu'ils n'en forment qu'un seul. La figure 3.13 montre comment s'effectue une fusion entre deux chaînes de même profil. La fusion se fait à partir d'un nœud pivot d'une chaîne vers un autre nœud pivot d'une autre chaîne de même profil.

(1)



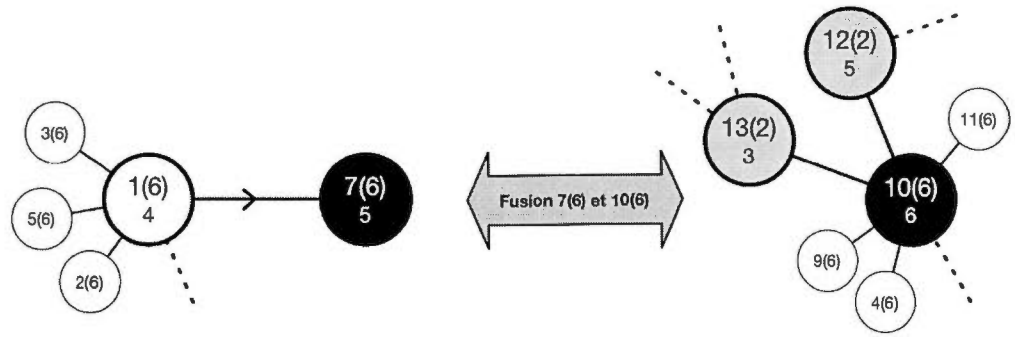
Fusion de 1(6) vers 10(6).
On veut fusionner la chaîne de 1(6) avec la chaîne de 10(6). La fusion doit débuter sur les pivots de fin de chaîne qui sont 7(6) et 10(6).

(2)



Fusion de 7(6) vers 10(6) – suite.
On transfère d'abord les voisins simples de même profil de 7(6) sur 10(6), tant que 10(6) a de la place. On transfère donc les noeuds 4(6) et 9(6) sur 10(6). 10(6), dont la capacité est de 6, a encore de la place puisqu'il ne possède que 5 liens après le transfert.

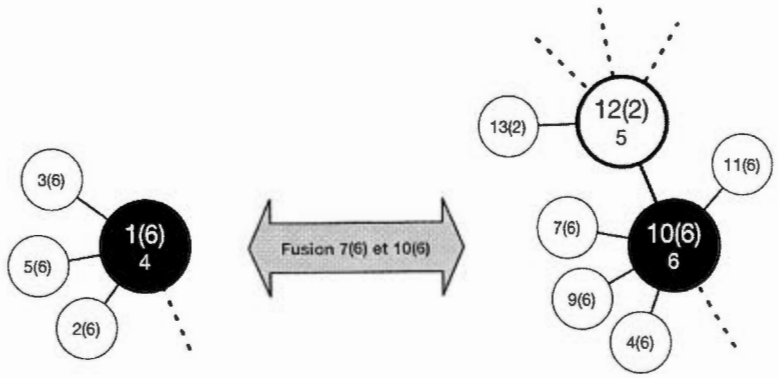
(3)



Fusion de 7(6) vers 10(6) – suite.
On transfère ensuite les voisins pivots de profils différents de 7(6) sur 10(6). Dans notre exemple, il n'y en a qu'un seul, 13(2).

On vérifie si la réception d'un nouveau lien entre les deux noeuds pivots concernés, 13(2) et 10(6), a créé des voisins de même profil entre eux. On remarque que c'est le cas du noeud 10(6) qui a maintenant deux voisins de profil 2 : 13(2) et 12(2). Dans ce cas, on coupe le lien entre, par exemple, 13(2) et 10(6) puis on rappelle récursivement la fusion sur 13(2) et 12(2).

(4)

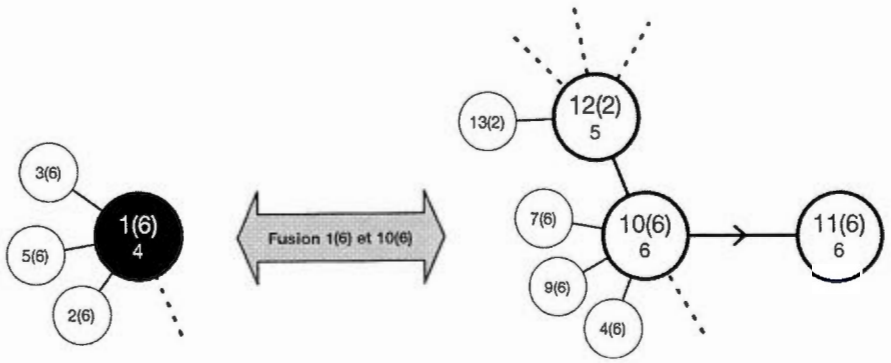


Fusion de 7(6) vers 10(6) – fin.
Au retour de l'appel récursif de la fusion de 13(2) vers 12(2), on remarque que ces deux noeuds font maintenant partie du même regroupement.

On continue la fusion entre 7(6) et 10(6). 7(6) n'a plus de voisins à transférer, mais il possède un précédent. On fait de 7(6) un noeud simple, sans précédent, et on le transfère sur 10(6) – qui a encore de la place.

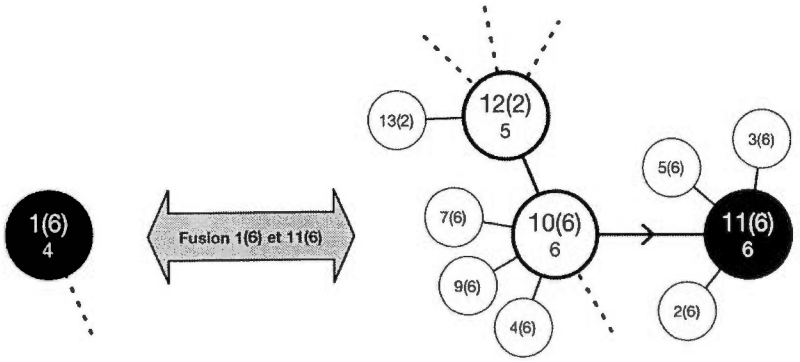
On rappelle ensuite la fusion sur le noeud qui était le précédent de 7(6), c.-à-d. 1(6), vers 10(6).

(5)



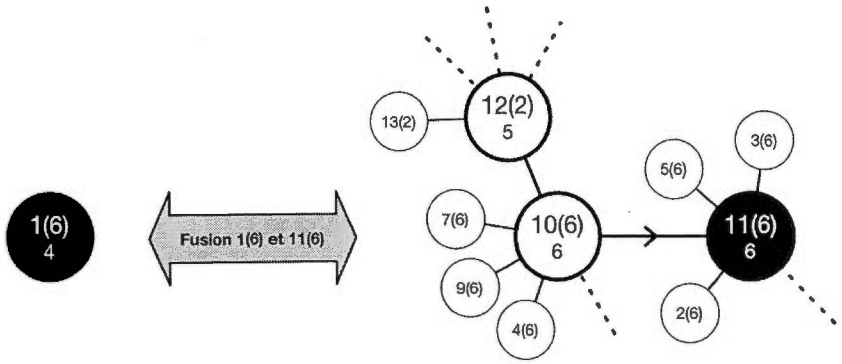
Fusion de 1(6) vers 10(6).
On veut transférer les voisins simples de même profil de 1(6) sur 10(6), mais celui-ci est plein. On doit **libérer une place** sur la chaîne de 10(6). On crée un nouveau pivot en utilisant le voisin simple 11(6). La fusion se poursuit donc sur ce nouveau pivot final.

(6)



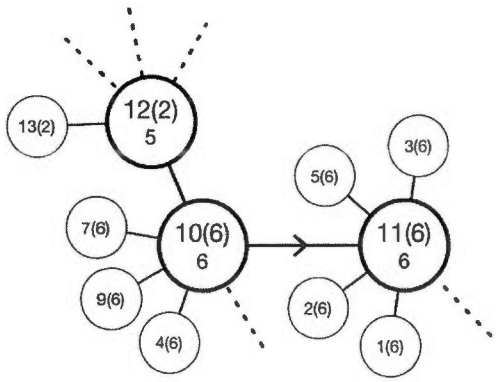
Fusion de 1(6) vers 11(6).
On transfère les voisins simples de même profil de 1(6), c.-à-d. 3(6), 5(6) et 2(6), sur le nouveau pivot final 11(6). Après le transfert, 11(6) a encore de la place.

(7)



Fusion de 1(6) vers 11(6) – suite.
On transfère les voisins de profils différents (dans notre cas, il n'y en a qu'un seul, représenté par le trait discontinu) tout en fusionnant les voisins de même profil entre eux lorsqu'il y a lieu.

(8)



Fusion de 1(6) vers 11(6) – fin.
Une fois le transfert terminé, on fait de 1(6) un noeud simple qui va se greffer au pivot 11(6). Les deux chaînes à fusionner n'en forment plus qu'une seule.

Figure 3.13 Fusion de communautés dans un réseau à pivots chaînés.

Voyons maintenant comment les trois algorithmes précédents sont utilisés dans l'implémentation des règles générales de notre modèle de socialisation.

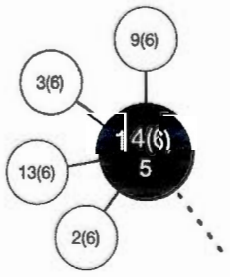
Connexion

La connexion dans un réseau à pivots chaînés est très similaire à celle de notre premier modèle à la différence près que, lors du parcours de connexion, lorsqu'on choisit le meilleur pivot sur lequel se connecter (de profil identique ou le plus similaire rencontré), on doit se rendre sur le pivot au bout de la chaîne de ce pivot choisi (puisque tous les précédents sont nécessairement pleins). Si le pivot final est libre, le cas est trivial, on connecte tout simplement notre nouveau nœud sur ce pivot, de la même manière que dans notre premier modèle. Sinon, on doit libérer une place (voir l'algorithme de libération d'une place, fig. 3.12) sur ce pivot final pour ensuite effectuer la connexion.

Déconnexion

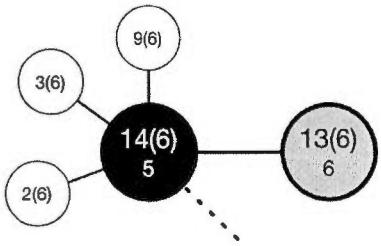
Comme dans notre premier modèle, la déconnexion d'un nœud simple est triviale puisqu'il n'a qu'un seul lien le reliant à son pivot immédiat. Dans ce cas, on le supprime tout simplement. Pour ce qui est de la déconnexion des pivots, on doit considérer plusieurs cas de figure selon que l'on déconnecte un pivot initial (avec ou sans suivant), un pivot intermédiaire ou un pivot final. Cependant, dans tous les cas, l'idée générale reste la même : désigner un nœud tiers qui recevra les liens du nœud à déconnecter, et ce, tout en maintenant les propriétés structurales du réseau et en conservant les regroupements de nœuds déjà présents dans le graphe. La figure 3.14 montre comment déconnecter un pivot initial dans différentes conditions.

(1)



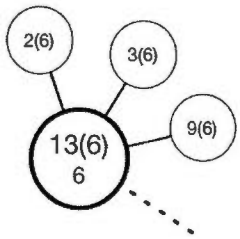
Cas 1 : Déconnexion d'un pivot initial sans suivant, avec voisins de même profil. Déconnexion de 14(6).

(2)



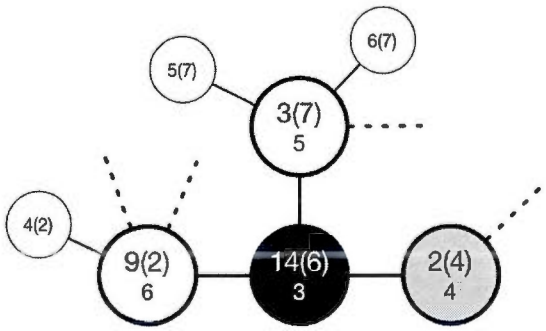
Déconnexion de 14(6) – cas 1 (suite).
Dans ce cas, on choisit un noeud simple de même profil, par exemple, 13(6), et l'on en fait un noeud pivot qui remplacera le pivot à déconnecter. Le choix du noeud simple est arbitraire, mais pour des raisons de performance, on choisira le noeud simple de capacité maximale pour réduire le chaînage.

(3)



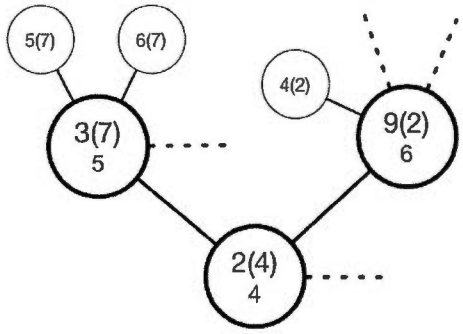
Déconnexion de 14(6) – cas 1 (fin).
On transfère tous les voisins simples de même profil du pivot 14(6) sur le nouveau pivot 13(6) ainsi que tous les voisins pivots de profils différents (en fusionnant les voisins de même profil entre eux s'il y a lieu). Finalement, on coupe le noeud 14(6). La déconnexion est terminée.

(4)



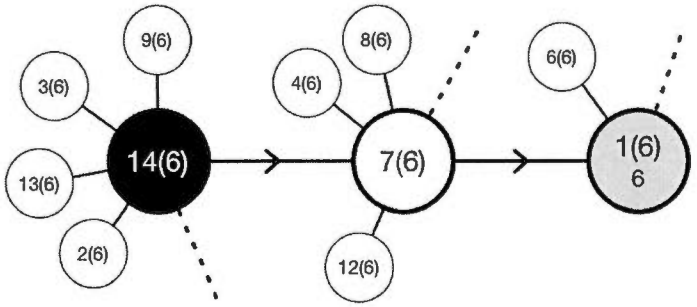
Cas 2 : Déconnexion d'un pivot initial sans suivant, sans voisins simples de même profil. Déconnexion de 14(6).
On choisit un voisin pivot de profil différent qui est libre (ou sur lequel on libère une place), par exemple 2(4), pour remplacer le noeud à déconnecter.

(5)



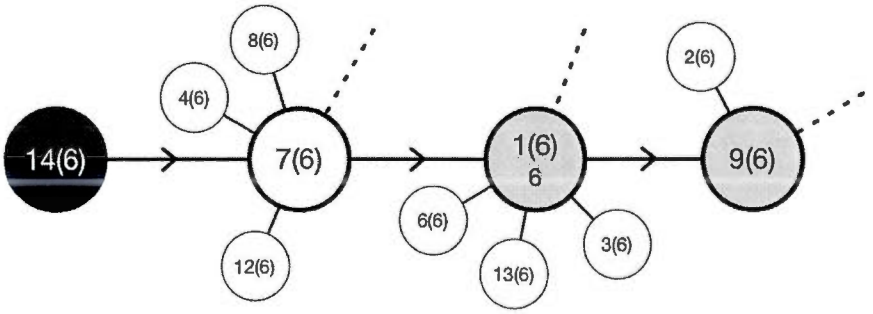
Déconnexion de 14(6) – cas 2 (fin).
On transfère tous les voisins pivots de 14(6) sur le noeud choisi 2(4), en libérant de la place si nécessaire et en fusionnant les voisins de même profil entre eux s'il y a lieu. Puis, on coupe le noeud 14(6). La déconnexion est terminée.

(6)



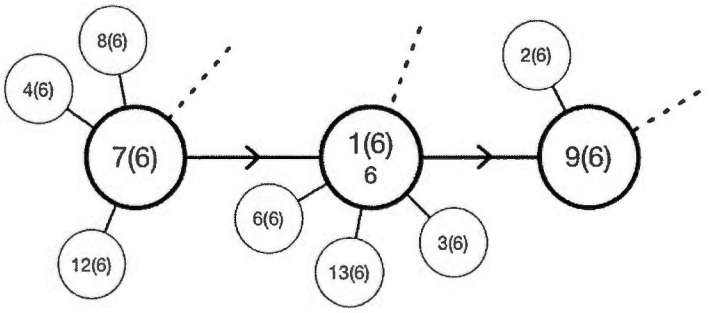
Cas 3 : Déconnexion d'un pivot initial qui a un suivant.
Déconnexion de 14(6).
On trouve d'abord le pivot final de la chaîne de 14(6) qui est 1(6).

(7)



Déconnexion de 14(6) – cas 3 (suite).
On transfère les voisins simples de même profil du pivot 14(6) sur 1(6) – en libérant de la place lorsqu'il y a lieu (dans cet exemple, on chaîne un pivot supplémentaire pour faire de la place sur la chaîne). On transfère ensuite les voisins pivots de profils différents sur 1(6) – en libérant de la place et en fusionnant les voisins de même profil entre eux lorsqu'il y a lieu.

(8)

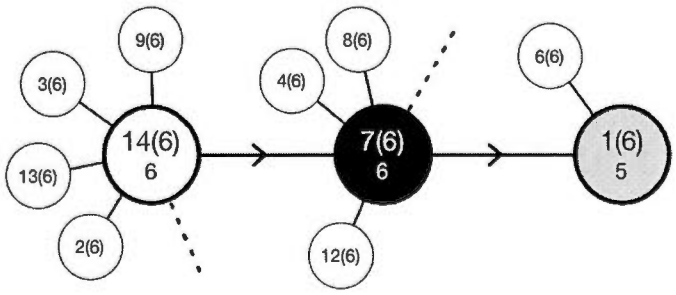


Déconnexion de 14(6) – cas 3 (fin).
On supprime le noeud 14(6). La déconnexion est terminée. On remarque cependant que le nouveau pivot initial de cette chaîne, 7(6), n'est pas plein et possède un suivant. Après cette déconnexion, on doit donc exécuter l'algorithme de remplissage d'une chaîne libre sur le noeud 7(6) qui aura pour effet de transférer le noeud 2(6) sur le noeud 7(6).

Figure 3.14 Déconnexions de pivots initiaux dans un réseau à pivots chaînés.

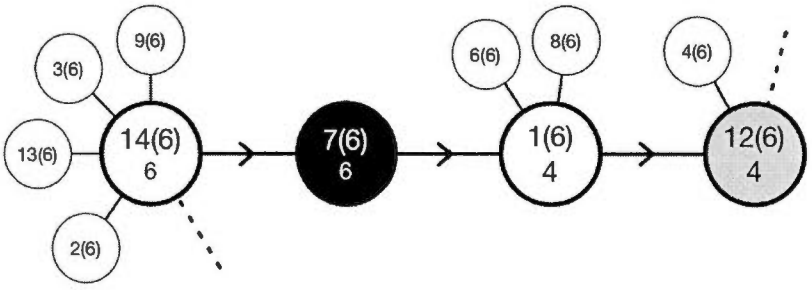
La figure 3.15 illustre l'algorithme de déconnexion d'un pivot intermédiaire.

(1)



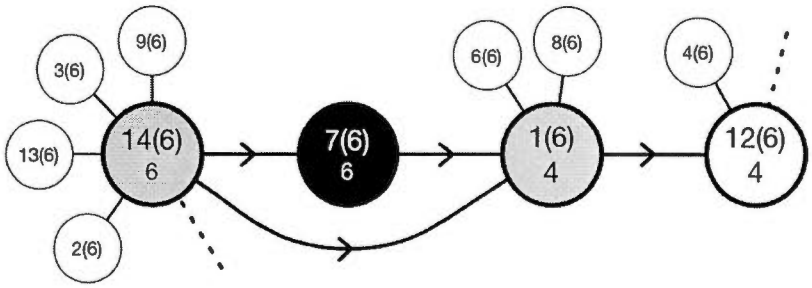
Déconnexion du pivot intermédiaire 7(6).
Choisir le pivot suivant du noeud à déconnecter 7(6), c.-à-d. le pivot 1(6).

(2)



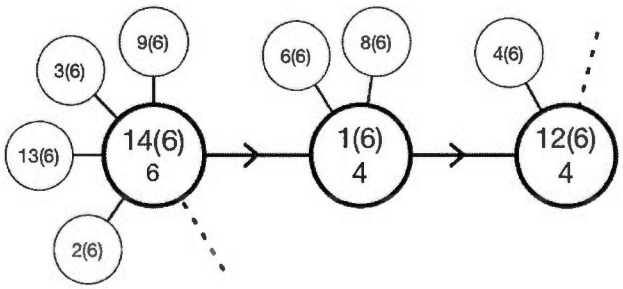
Déconnexion du pivot intermédiaire 7(6) – suite.
Transférer tous les voisins simples de même profil, de 7(6) vers 1(6), en libérant de la place si nécessaire. Puis, transférer tous les voisins pivots de profils différents, de 7(6) vers 1(6), en libérant de la place et en fusionnant les voisins de même profil s'il y a lieu.

(3)



Déconnexion du pivot intermédiaire 7(6) – suite.
Créer un lien entre le précédent de 7(6) et le suivant de 7(6). Le précédent de 7(6) devient le précédent de 1(6) et le suivant de 7(6) devient le suivant de 14(6).

(4)

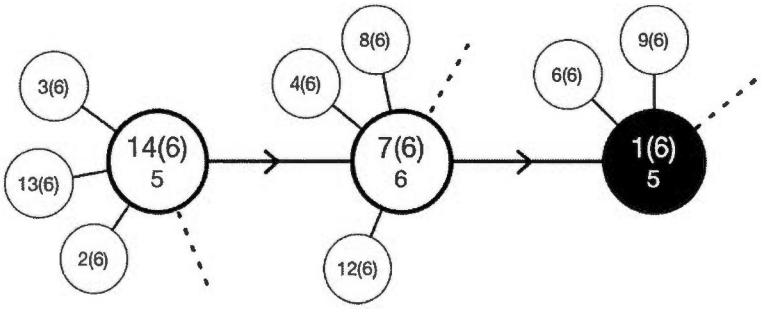


Déconnexion du pivot intermédiaire 7(6) – fin.
Supprimer le pivot 7(6). La déconnexion est terminée.

Figure 3.15 Déconnexion d'un pivot intermédiaire dans un réseau à pivots chaînés.

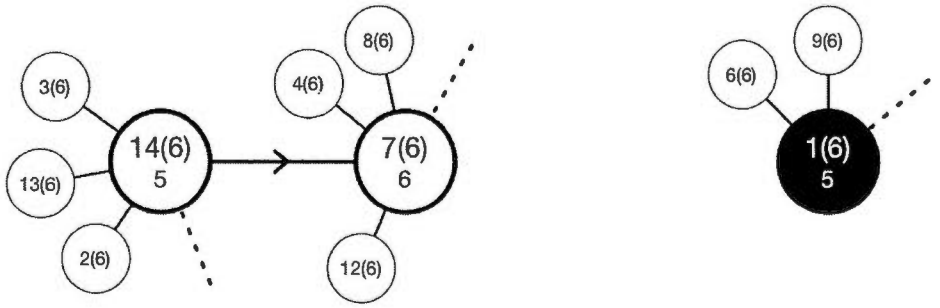
La figure 3.16 explique la déconnexion d'un pivot de fin de chaîne.

(1)



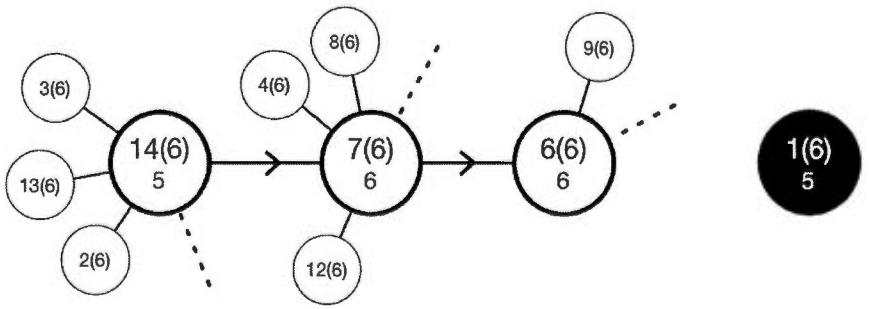
Déconnexion du pivot final 1(6).

(2)



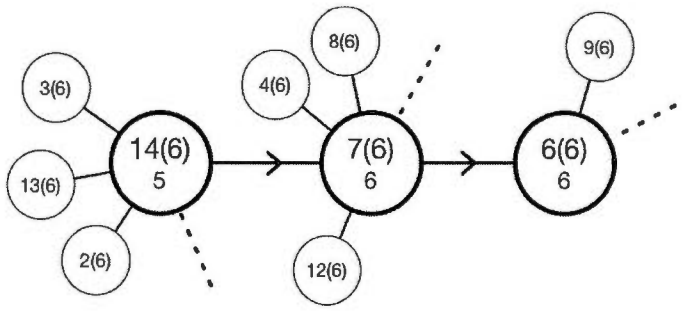
Déconnexion du pivot final 1(6) – suite.
Supprimer d'abord le lien entre le pivot final à déconnecter, 1(6), et son pivot précédent, 7(6).

(3)



Déconnexion du pivot final 1(6) – suite.
Transférer tous les voisins simples de même profil de 1(6) vers 7(6), en libérant de la place si nécessaire. Puis, transférer tous les voisins pivots de profils différents de 1(6) vers 7(6), en libérant de la place et en fusionnant les voisins de même profil s'il y a lieu.

(4)



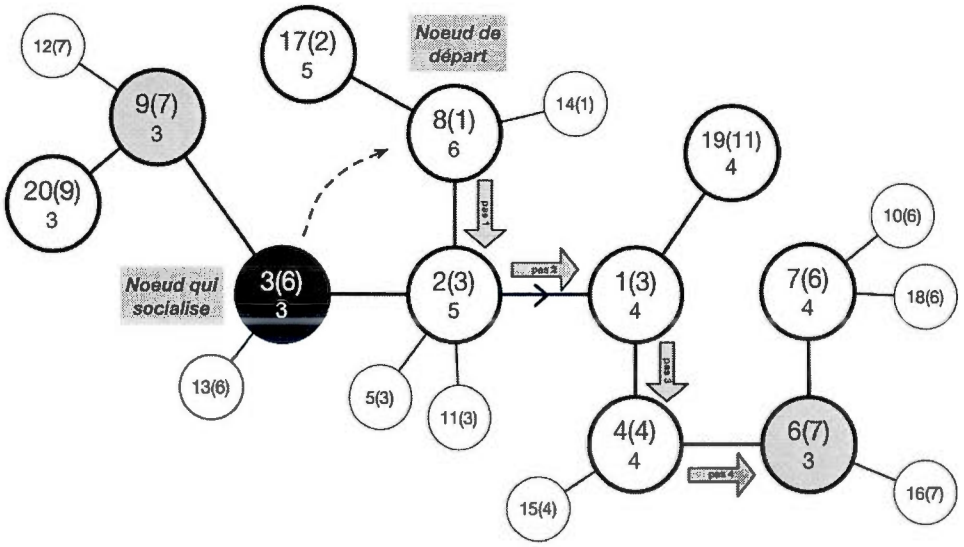
Déconnexion du pivot final 1(6) – fin.
Supprimer le pivot 1(6). La déconnexion est terminée.

Figure 3.16 Déconnexion d'un pivot final dans un réseau à pivots chaînés.

Parcours de socialisation – rencontres et recommandations

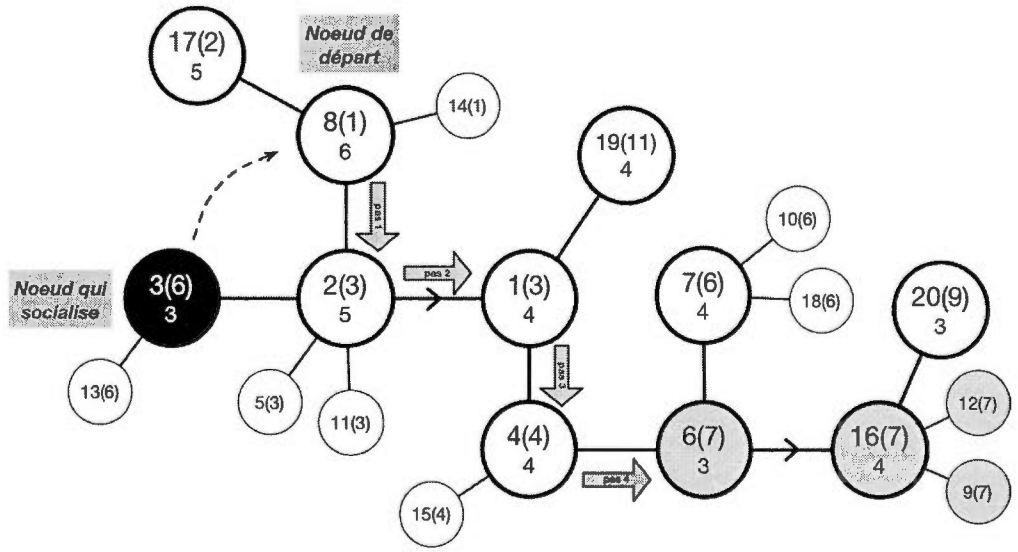
Les parcours de socialisation sont effectués de la même manière que dans le premier modèle présenté. Seule la façon d'opérer la fusion lors de rencontres et recommandations diffère (voir l'algorithme de fusion, fig. 3.13). La figure 3.17 illustre un exemple de parcours de socialisation dans lequel on effectue une recommandation, puis une rencontre.

(1)



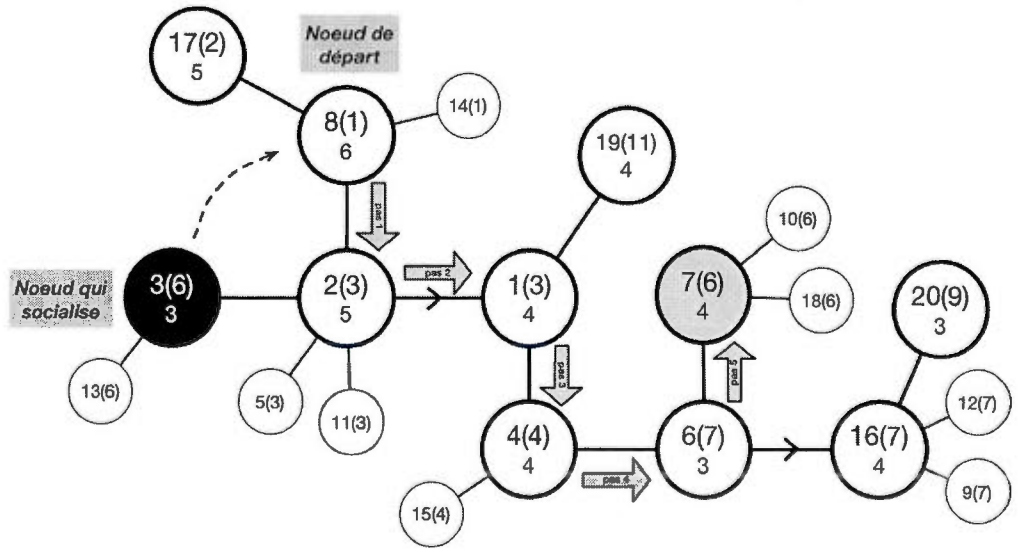
Socialisation du pivot 3(6) à partir du noeud aléatoire 8(1) – recommandation.
On effectue un parcours de socialisation en choisissant toujours le pivot de profil le plus similaire au noeud qui socialise. Lorsqu'on arrive au pivot 6(7), on effectue une recommandation avec le voisin immédiat 9(7) du socialisateur. On coupe ensuite e lien entre 9(7) et 3(6) puis on **fusionne** 9(7) avec le noeud rencontré 6(7).

(2)



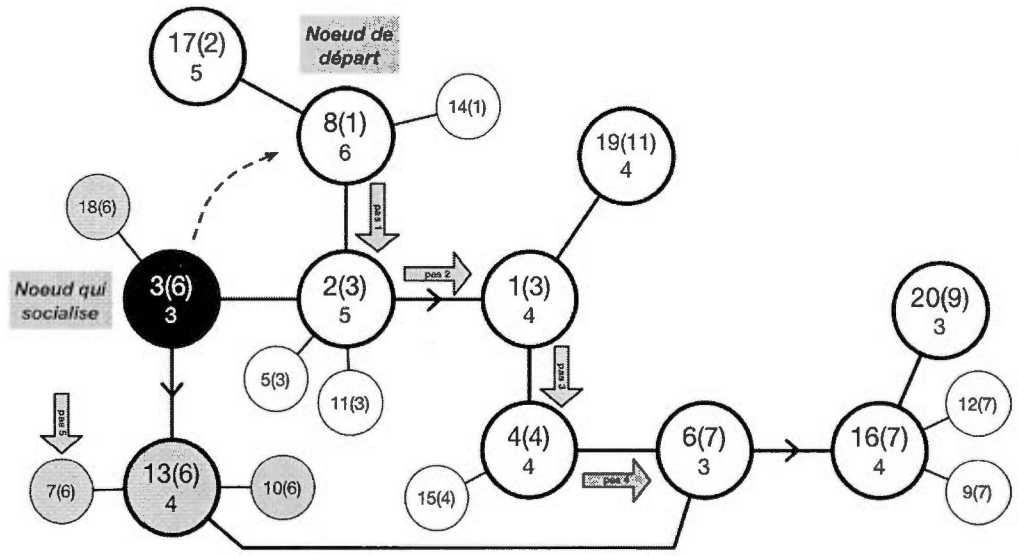
Socialisation du pivot 3(6) à partir du noeud aléatoire 8(1) – suite.
Après la recommandation, on remarque que la chaîne de 9(7) et la chaîne de 6(7) n'en forment plus qu'une.

(3)



Socialisation du pivot 3(6) à partir du noeud aléatoire 8(1) – rencontre.
On continue le parcours de socialisation où l'on était rendu en passant au noeud 7(6) et l'on fait une rencontre. On **fusionne** donc le pivot rencontré 7(6) avec le noeud qui socialise 3(6).

(4)



Socialisation du pivot 3(6) à partir du noeud aléatoire 8(1) – fin.
Après la rencontre (fusion), tous les noeuds de la chaîne de 7(6) sont maintenant sur la chaîne de 3(6). On continue le parcours de socialisation. Comme le dernier pivot rencontré est devenu un noeud simple, on continue le parcours à partir de son pivot immédiat 13(6). Les deux voisins pivots sont 3(6), le noeud qui socialise, et 6(7) qu'on a déjà traversé donc le parcours s'arrête ici et la socialisation est terminée.

Encore une fois, on voit que les restructurations causées par la recommandation puis la rencontre ont l'effet de regrouper les noeuds de même profil.

Figure 3.17 Rencontre et recommandation dans un réseau à pivots chaînés.

Mise à jour des relations

Comme dans notre premier modèle, la mise à jour des relations correspond à une déconnexion, pour séparer le nœud d'un regroupement qui ne lui convient plus, suivie d'une connexion, pour lui donner la chance de trouver une autre communauté d'intérêt de profil similaire à son nouveau profil.

Dans la section qui suit, nous présentons une manière plus plausible de représenter et comparer les profils d'intérêt des acteurs dans le réseau et les conséquences qui en découlent sur nos algorithmes d'évolution.

3.4 Considérations sur les profils d'intérêts

3.4.1 Deux individus ne sont jamais identiques

3.4.1.1 Ajout d'une limite d'équivalence

Jusqu'à maintenant, par souci de clarté et de simplicité, nous avons décrit les règles de notre modèle de socialisation en utilisant des nombres entiers pour représenter les profils d'intérêts des utilisateurs. Selon la méthode que nous avons utilisée pour déterminer la similarité (voir par. 3.2.1.2), lorsque les nœuds ne sont pas de profil parfaitement identique, ils sont similaires à différents degrés : la similarité entre deux nœuds de même profil est égale à 1 (similarité parfaite) tandis que la similarité entre un nœud de profil 4 et un nœud de profil 5 (similarité = 0.5) est plus grande que la similarité entre un nœud de profil 4 et un nœud de profil 6 (similarité = 0.3). Avec cette façon de faire, deux profils sont donc soit parfaitement identiques, soit plus ou moins similaires.

Toutefois, dans la réalité, peu importe la manière d'extraire les profils d'intérêts des utilisateurs, il est peu probable que deux individus aient des profils d'intérêts exactement identiques. On peut penser, par exemple, à deux individus qui apprécient la musique classique. Il se peut que chacun d'entre eux aime, par exemple, Brahms, Schumann et Mozart, mais qu'un seul des deux aime Beethoven. Dans ce cas, on ne peut pas affirmer que leurs profils d'intérêts sont parfaitement identiques. On pourrait cependant dire qu'ils sont très similaires (à un degré élevé) et qu'en réalité, ils gagneraient à se rencontrer.

Dans ce contexte, où aucun acteur n'est susceptible d'être parfaitement identique à un autre acteur dans le réseau, nos algorithmes de regroupements, tels que décrits à la section précédente, ne regrouperont rien du tout. Pour adapter notre modèle à ce contexte plus vraisemblable, nous allons ajouter un paramètre à notre méthode de calcul de la similarité entre deux profils : une valeur entre 0 et 1 qui détermine une *limite d'équivalence* inférieure au-delà de laquelle on considérera que deux profils sont parfaitement identiques dans le cadre de nos algorithmes. Par exemple, si ce paramètre est défini avec la valeur 0.8, la fonction de similarité considérera que, lorsque la similarité entre deux profils est supérieure ou égale à

0.8, les deux profils comparés sont en fait identiques, et la fonction retournera plutôt la valeur 1 (qui indique une similarité parfaite). Dans le cas contraire, elle retournera la valeur réelle de la similarité calculée (< 0.8). La valeur de ce nouveau paramètre détermine donc à quel point deux individus doivent être semblables pour être considérés identiques dans notre modèle de regroupement.

Aussi, dans notre modèle à profils numériques, si le nœud *a* est de profil identique au nœud *b* et au nœud *c*, on peut conclure que les nœuds *b* et *c* sont aussi de profil identique. Il n'en est plus de même lorsqu'on considère maintenant que deux profils sont dits identiques s'ils ont une valeur de similarité plus grande ou égale à une certaine limite (limite d'équivalence). Dans ce cas, on ne peut plus conclure que le profil de *c* est égal au profil de *b* lorsqu'on a que le profil de *a* est égal au profil de *b* et de *c*. Supposons qu'un profil d'intérêts est composé d'une liste d'intérêts (A, B...) et que la valeur de la similarité est déterminée en fonction du nombre d'intérêts communs des profils comparés. Dans ce cas, avec une valeur de la limite d'équivalence égale à 0.8 (similaire à 80 % = identique), deux profils composés de dix intérêts chacun devraient en avoir au moins huit en commun pour être considérés comme identiques.

La figure 3.18 montre un exemple où l'affirmation ($a = b$ et $a = c$) => $b = c$ n'est pas vraie. La figure 3.18 (a) illustre trois profils d'intérêts, P1, P2 et P3, chacun étant composé d'une liste de dix intérêts, représentés par des lettres majuscules. La figure 3.18 (b) montre le nombre d'intérêts qu'ont en commun les trois profils. On remarque que P1 possède huit intérêts en commun avec P2 ainsi qu'avec P3. Ils sont donc considérés comme identiques lorsque notre valeur limite est égale à 0.8. Cependant, P2 et P3 n'ont que six intérêts en commun et ne sont donc pas considérés comme identiques. Ainsi (P1 = P2 et P1 = P3) n'implique pas que P2 = P3. Cependant, si notre valeur limite avait été fixée à 0.6, les trois profils auraient été considérés comme identiques, car ils ont tous les trois 6 intérêts en commun.

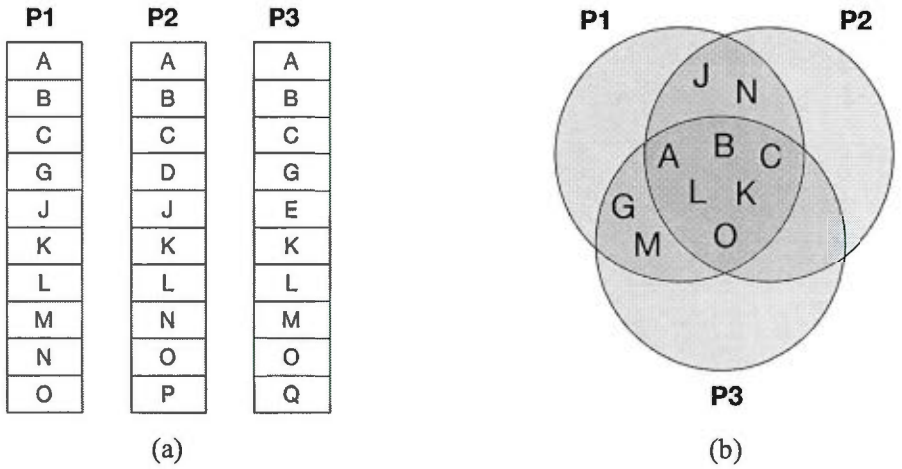


Figure 3.18 Comparaison de la similarité entre trois profils d'intérêts P1, P2 et P3.

Évidemment, plus la valeur de la limite d'équivalence est petite, moins les profils considérés parfaitement similaires le sont en réalité et donc moins les regroupements de notre modèle seront cohésifs. Ceci n'est pas nécessairement une mauvaise chose. En fait, cette limite permet simplement d'ajuster dans quelle mesure on aimerait que les membres d'une même communauté se ressemblent les uns les autres au sein du réseau.

Dans certains contextes, on voudra privilégier une très forte ressemblance tandis que dans d'autres, une plus grande ouverture. Par exemple, les membres d'un réseau social composé de fanatiques de divers chanteurs populaires, par exemple, aimeraient se retrouver au sein d'une communauté regroupant des fanatiques du même artiste (donc très similaires au niveau d'un intérêt spécifique) tandis que les membres d'un réseau composé d'amateurs de musique, en général, pourraient apprécier découvrir de nouveaux chanteurs, groupes et genres musicaux au sein d'une communauté d'individus plus variés.

Nous dirons désormais que deux profils sont identiques ou de similarité parfaite lorsque la similarité entre deux nœuds est plus grande ou égale à la valeur de la limite d'équivalence utilisée par la fonction de similarité, quelle qu'elle soit.

3.4.1.2 Désignation d'un représentant par communauté d'intérêt

Pour pallier les problèmes causés par la limite d'équivalence, dans les algorithmes du modèle, nous avons ajouté un attribut supplémentaire aux nœuds du réseau, qui différencie le profil personnel d'un nœud (son attribut `profil`) de son profil de groupe. Ce nouvel attribut, `profilGroupe`, prend la valeur du profil d'un seul nœud, désigné à l'intérieur d'un regroupement comme étant le représentant de ce groupe. Chaque nœud a donc un profil personnel, mais connaît de plus le profil du représentant de la communauté à laquelle il appartient. Lorsqu'un nœud ne fait partie d'aucun regroupement (un nœud qui se connecte) ou lorsqu'il est l'unique membre de son regroupement, son profil de groupe est le même que son profil personnel. Aussitôt qu'un nœud se greffe à une chaîne de pairs identiques (une communauté), son attribut `profilGroupe` prend la valeur du profil représentant ce groupe.

En conséquence, nous ferons dorénavant la comparaison des profils sur les profils de groupe plutôt que sur les profils personnels. Un nœud qui socialise, socialise désormais en fonction du profil de groupe. Chaque nœud d'un regroupement est donc considéré comme parfaitement identique au nœud qui représente cette communauté. Il s'ensuit que, comme discuté plus haut, certains nœuds d'un même regroupement peuvent ne pas être parfaitement identiques entre eux (en regard de la limite d'équivalence utilisée). On peut cependant supposer qu'une grande partie d'entre eux seront tout de même assez similaires. Il est à noter que l'introduction de profils de groupes n'empêche pas l'utilisation d'une fonction de similarité exacte, comme dans le cas des profils numériques. La valeur de la limite d'équivalence devrait être fixée à 1, dans ce cas.

Plus précisément, nous avons choisi de désigner le premier nœud (pivot) d'un regroupement comme représentant de son sous-groupe. Dans notre modèle (à pivots chaînés), le représentant du groupe est donc toujours le premier pivot de la chaîne, le pivot initial. Chaque fois qu'un nœud simple se connecte à un pivot de profil identique ou qu'il devient un nouveau pivot de la chaîne, il enregistre le profil de groupe du nœud auquel il vient de se connecter. Ainsi se propage le profil de groupe à mesure que des nœuds viennent s'y greffer.

Dans certains cas, de nouveaux mécanismes de propagation du profil de groupe sur une chaîne doivent être établis, mais les idées principales derrière les algorithmes décrits à la section précédente restent les mêmes. Pour ne donner qu'un exemple, lorsqu'un pivot initial se déconnecte et qu'il passe la main à un autre nœud de sa chaîne pour devenir le nouveau pivot initial, le profil de groupe du nouveau pivot initial doit être propagé sur tous les nœuds de même profil qui sont sur cette chaîne. Il arrive aussi parfois qu'au cours de cette propagation, des noeuds pivots de profils différents attachés à l'un des pivots de cette chaîne deviennent identiques (par comparaison basée sur les profils de groupes) au nouveau profil de groupe ainsi propagé. Pour respecter la propriété structurale qui stipule qu'un pivot ne peut être connecté qu'à au plus deux autres pivots de même profil (le suivant et/ou le précédent), lorsque cela se produit, on doit fusionner la chaîne de ce pivot devenu identique à la chaîne sur laquelle on propage le nouveau profil.

Un changement de représentant au sein d'un regroupement, comme dans le cas de la déconnexion d'un pivot initial, altère donc le profil du groupe et permet ainsi à des acteurs qui n'étaient pas parfaitement similaires au représentant précédent, mais qui sont identiques au nouveau représentant de venir se connecter au groupe. Cette dynamique est fortement désirable pour permettre à des nœuds d'un regroupement de rencontrer d'autres nœuds du réseau ayant un profil personnel identique, mais dont le profil de groupe les a différenciés sous le règne du représentant précédent. Ainsi, tout au cours de l'évolution du réseau, les individus se promènent dans différents regroupements (selon qu'ils sont identiques ou non au représentant courant du groupe). Nous pensons que cette dynamique favorisera la rencontre éventuelle de la majeure partie des utilisateurs de même profil personnel. Nous testerons cette hypothèse au chapitre 4.

Dans ce nouveau cadre de fonctionnement, on ne peut plus parler de l'état de perfection d'un réseau au sens où nous en avions discuté dans les sections précédentes. Nous avions défini un réseau parfait comme étant un réseau dans lequel tous les nœuds de même profil font partie d'un seul et même regroupement : le réseau contient autant de regroupements que de profils différents présents dans le réseau, impliquant que tous les nœuds de même profil sont regroupés ensemble. On a maintenant une conception plus souple du concept de perfection et

l'on dira plutôt qu'un réseau est dans l'une de ses formes parfaites possibles lorsque tous les nœuds identiques sont regroupés ensemble, en fonction des profils de groupe de tous les représentants du réseau (et non des profils personnels).

Ainsi, pour divers représentants de groupes dans le réseau, on aura très possiblement des regroupements différents. Par exemple, les figures 3.19 et 3.20 illustrent deux réseaux parfaits comportant les mêmes 38 acteurs, mais dans lesquels les regroupements d'utilisateurs sont différents. Les couleurs différentes spécifient des regroupements distincts, mais non des profils spécifiques (deux regroupements identiques dans les deux réseaux illustrés peuvent ne pas être de la même couleur). Les nœuds dont le contour est noir sont les pivots.

Les tableaux 3.1 et 3. 2 listent plus clairement toutes les chaînes de noeuds se trouvant respectivement dans les réseaux illustrés aux figures 3.19 et 3.20. Chaque case des tableaux contient un sous-groupe d'individus et dans chaque regroupement, le représentant du groupe est indiqué en caractères gras.

En considérant les tableaux 3.1 et 3.2, on observe certains effets dus aux regroupements par profil de groupe. Tout d'abord, le nombre et le choix des représentants de groupe sont différents dans les deux réseaux. En fait, les deux réseaux ont quatre représentants en commun : *Herunar*, *ppuleojr*, *mustapha19* et *bpunch*. Par contre, les membres de leur sous-groupe respectif varient d'un réseau à l'autre. Par exemple, dans le réseau de la figure 3.19, *ppuleojr* (nœud orange) est le seul membre de son groupe tandis que dans le réseau de la figure 3.20, il fait partie d'une communauté de cinq membres (regroupement rose foncé). *Herunar*, dans le premier réseau, est le représentant d'un groupe de deux acteurs composé de lui-même et de *stsinc* (groupe beige) tandis que dans le deuxième réseau, on a un regroupement similaire, mais qui contient un individu de plus, *codermonk* (groupe beige aussi).

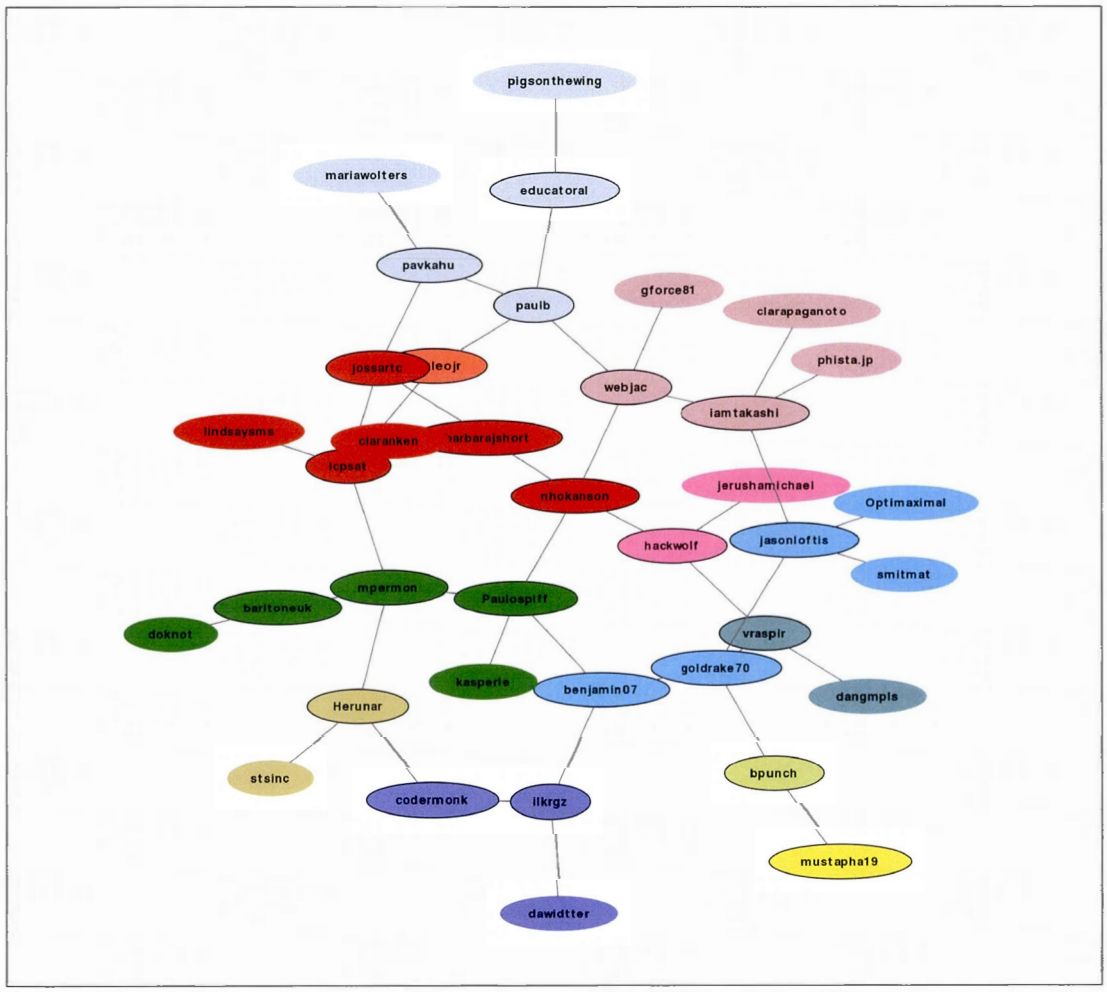


Figure 3.19 Réseau parfait composé de 38 acteurs.

Le réseau ci-dessus est dans son état de perfection, car il contient autant de chaînes différentes que de profils de groupe différents (couleurs), c.-à-d. 12.

Tableau 3.1

Regroupements d'individus dans le réseau de la figure 3.19

nhokanson barbarajshort ciaranken jossartc lcpsat lindsaysms	**webjac** gforce81 iamtakashi phista.jp clarapaganoto	**pavkahu** mariawolters paulb educatoral pigsonthewing	**Paulospiff** kasperle mpermon baritoneuk doknot	**benjamin07** goldrake70 jasonloftis Optimaximal smitmat	**ilkrgz** dawidtter codermonk
Herunar stsinc	**hackwolf** jerushamichael	**vraspir** dangmpls	**ppuleojr**	**mustapha19**	**bpunch**

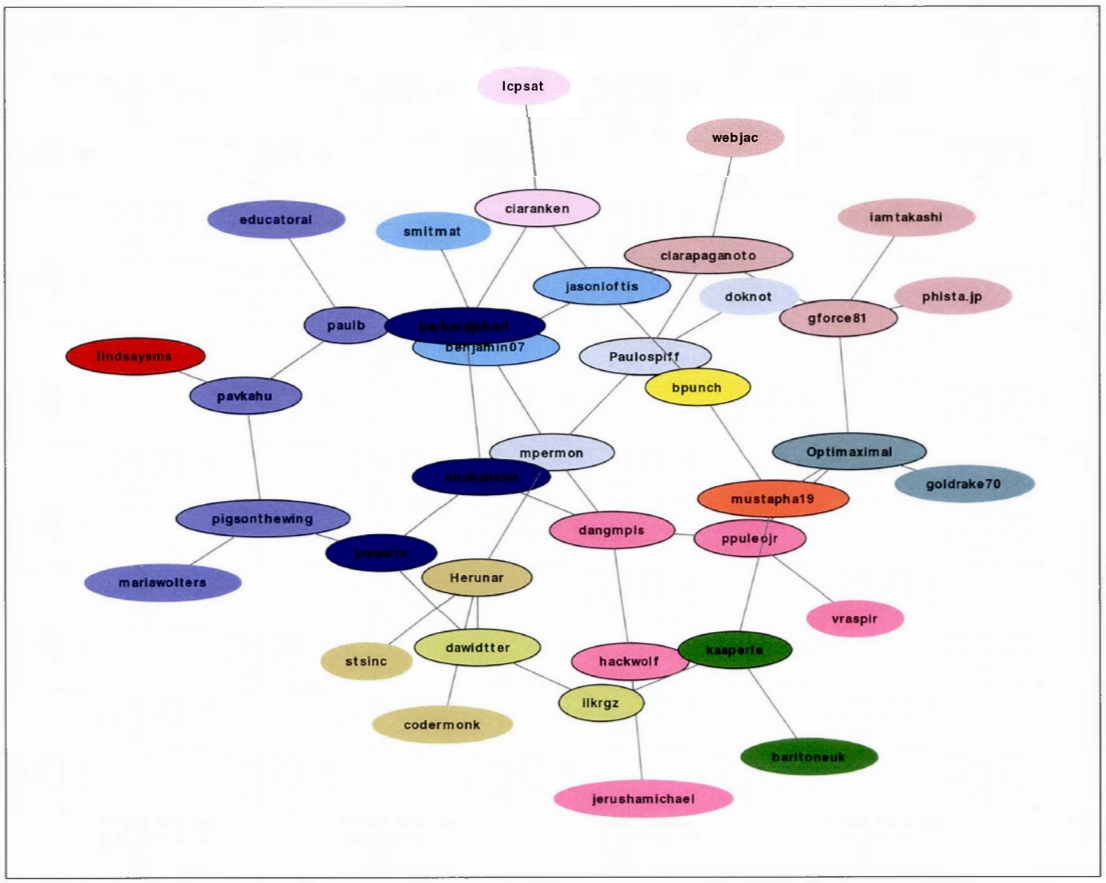


Figure 3.20 Autre configuration d'un réseau parfait composé des mêmes 38 acteurs que la figure 3.19. Le réseau ci-dessus est dans son état de perfection, car il contient autant de chaînes différentes que de profils de groupe (couleurs) différents, dans ce cas, 14.

Tableau 3.2
Regroupements d'individus dans le réseau de la figure 3.20

gforce81 phista.jp iamtakashi clarapaganoto webjac	**pppuleojr** vraspir dangmpls hackwolf jerushamichael	**paulb** educatoral pavkahu pigsonthewing mariawolters	**jasonloftis** benjamin07 smitmat	**mpermon** Paulospiff doknot	**Herunar** codermonk stsinc
jossartc nhokanson barbarajshort	**Optimaximal** goldrake70	**kasperle** baritoneuk	**ciaranken** lcpsat	**dawidtter** ilkrgz	**mustapha19**
bpunch	**lindsaysms**				

Dans les deux réseaux, donc, les regroupements ne sont pas nécessairement composés des mêmes individus. Par exemple, le regroupement de *Paulospiff* (*Paulospiff, kasperle, mpermon, baritoneuk, doknot*) dans le tableau 3.1 est séparé en deux regroupements distincts dans le tableau 3.2, celui de *mpermon* (*mpermon, Paulospiff et doknot*) et de *kasperle* (*kasperle et baritoneuk*). Dans certains cas, même s'ils ont un représentant différent, certains regroupements d'individus sont demeurés les mêmes, dans les deux réseaux. Par exemple, le sous-groupe représenté par *webjac* dans le tableau 3.1 comprend les mêmes individus que le regroupement représenté par *gforce81* dans le tableau 3.2.

Ces variations dans la formation des sous-groupes au sein du réseau sont causées par l'utilisation d'une méthode de comparaison des profils avec limite d'équivalence, et particulièrement lors des parcours de connexion : le moment de la connexion d'un nœud est le seul moment où l'on compare le profil personnel d'un nœud (avec le profil de groupe des nœuds qu'ils parcourent pour effectuer la connexion). Un nœud qui se connecte suivra un chemin dans le graphe qui le conduira possiblement vers un groupe identique à son profil personnel. Toutefois, il n'est pas exclu qu'ailleurs dans le graphe, il existe un autre regroupement, de profil différent de celui effectivement rencontré, mais tout de même identique (toujours selon la limite d'équivalence) au profil personnel du nœud qui se connecte. Ainsi, le chemin de connexion est en partie responsable du choix de la communauté à laquelle un nœud se connecte, lorsque plusieurs possibilités existent. On comprend, ici, que la dynamique du réseau (connexions, déconnexions, mises à jour, socialisations), qui favorise les restructurations, augmente aussi les possibilités de nouvelles rencontres, au sein des communautés qui sont aussi en constante évolution.

Jusqu'à présent, nous avons parlé des profils d'intérêts en des termes assez vagues. Nous avons surtout insisté sur la méthode de comparaison de profils : celle-ci, étant donné une limite inférieure entre 0 et 1, qui détermine la similarité minimale pour que deux profils soient considérés comme identiques, doit retourner une valeur de similarité qui oscille entre 0 (aucune similarité) et 1 (similarité parfaite). En effet, notre modèle de socialisation est générique au sens où peu importe la manière utilisée pour calculer le profil d'intérêts des utilisateurs, il suffit de fournir une méthode pouvant calculer la similarité entre deux profils,

comme nous venons de l'expliquer, pour que le modèle fonctionne tel que nous l'avons décrit. Pour simuler et évaluer notre modèle, toutefois, nous avons dû choisir une méthode pour calculer et comparer les profils des acteurs qui composent notre réseau. Nous expliquons cette méthode dans la section qui suit.

3.4.2 Compilation et comparaison des profils d'intérêts

3.4.2.1 Extraction des préférences des utilisateurs

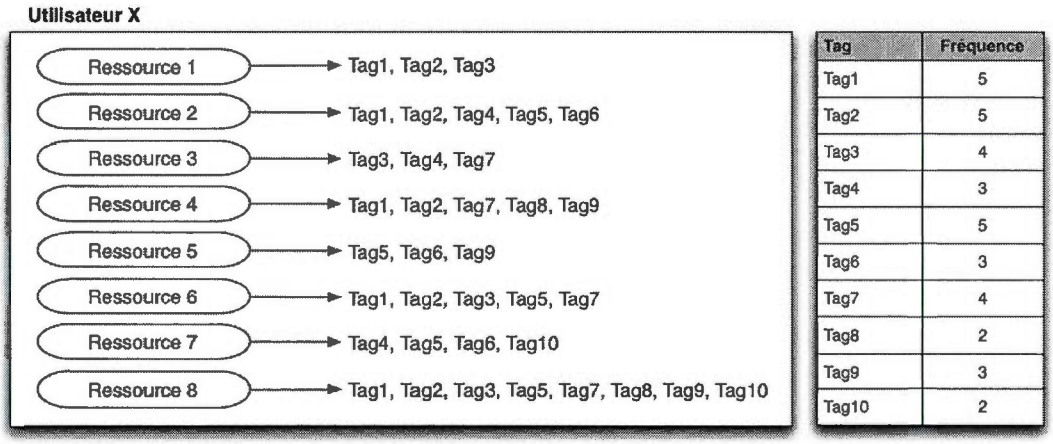
Comme nous l'avons vu au chapitre 2 (sect. 2.7), il existe déjà diverses méthodes (explicites et implicites) pour compiler le profil d'intérêts d'un utilisateur. Dans le cadre de cette thèse, cependant, le choix d'une méthode particulière est assez arbitraire. En effet, nous n'évaluerons pas l'efficacité de la méthode utilisée puisque ce n'est pas notre objectif ici et que de surcroît, dans notre contexte de modélisation, nous n'avons pas accès à des utilisateurs réels pouvant déterminer si oui ou non, les individus rencontrés (regroupés ensemble) dans le réseau étaient effectivement intéressants les uns pour les autres. Cependant, nous allons évaluer, au cours de simulations diverses, si les individus de profil identique, *étant donnée une méthode de compilation et de comparaison de profils*, se rencontrent effectivement au cours de l'évolution du réseau. Cela étant dit, nous voulons tout de même proposer une méthode vraisemblable et fonctionnelle qui pourrait s'avérer utile pour l'utilisation éventuelle de notre modèle de socialisation dans la conception d'une application collaborative de partage d'informations, par exemple.

Une approche qui nous semble raisonnable est de considérer la spécification des profils à l'aide de mots-clés. De cette manière, un profil est composé tout simplement d'une liste de mots décrivant les intérêts de chaque individu. Maintenant, la manière dont on obtient cette liste de mots-clés dépend essentiellement de l'application hypothétique qu'on ferait d'un tel modèle de socialisation, mais procure néanmoins un bon éventail de possibilités. Dans une application réelle, avec de vrais utilisateurs, on pourrait tout simplement obtenir une liste de mots-clés déclarée explicitement par chaque individu. Évidemment, dans ce cas, l'application serait dans l'impossibilité de mettre à jour les profils des individus de manière automatique et devrait s'en remettre à l'individu pour ce faire.

De manière implicite, cependant, on pourrait par exemple, extraire une liste de mots fréquents se retrouvant dans les pages Web marquées (favoris, signets, marque-pages) d'un utilisateur. Dans un système de partage de ressources textuelles, en faisant l'hypothèse que les ressources partagées par un utilisateur représentent ces intérêts, il serait encore possible de dégager une liste de mots récurrents à partir de la collection de documents partagés de chaque utilisateur. Plusieurs types de traces électroniques récupérables, laissées par les utilisateurs réels, pourraient, dans une certaine mesure, être transformées en une liste de mots-clés décrivant les goûts des utilisateurs.

Comme on l'a mentionné au chapitre 2 (art. 2.7.3), l'utilisation de tags (mots-clés) devient de plus en plus populaire pour extraire les préférences des individus. Comme le site de *tagging* collaboratif Delicious (http://delicious.com) offre une interface qui permet de récupérer certaines données publiques relatives aux ressources étiquetées de mots-clés (des URL) de différents utilisateurs, nous avons utilisé ces données pour modéliser, le plus vraisemblablement possible, les acteurs du réseau de notre modèle de socialisation. Nous avons donc récupéré une liste d'individus dans laquelle chaque individu possède une collection de ressources étant décrites par une liste de mots-clés eux-mêmes associés à une fréquence d'utilisation par cet utilisateur (voir sect. 4.2 pour plus de détails).

La fréquence associée aux mots-clés est une information qui, dans une certaine mesure, nous informe sur les préférences d'un individu. Les tags qui sont utilisés plus souvent que d'autres, par un utilisateur, sont susceptibles de dénoter des intérêts plus marqués de l'utilisateur en question. Nous utiliserons donc cette information dans la méthode de comparaison de profils, présentée au paragraphe suivant. La figure 3.21 illustre, pour un utilisateur X, un profil d'intérêts composé d'un ensemble de ressources étiquetées de mots-clés ainsi que la distribution des fréquences de tags associée à cette collection. C'est cette distribution des fréquences que nous utiliserons pour effectuer la comparaison de deux profils d'intérêts c.-à-d. pour déterminer notre fonction de similarité.



Tag	Fréquence
Tag1	5
Tag2	5
Tag3	4
Tag4	3
Tag5	5
Tag6	3
Tag7	4
Tag8	2
Tag9	3
Tag10	2

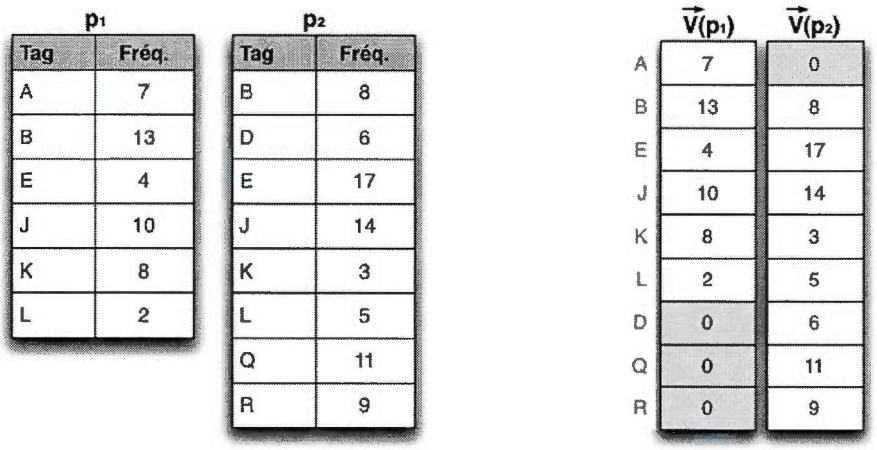
Figure 3.21 Vecteur de fréquences obtenu à partir d'une collection de ressources étiquetées de mots-clés.

3.4.2.2 Fonction de similarité dans un espace vectoriel

On utilise souvent la représentation d'un ensemble d'entités sous la forme de vecteurs dans un même espace vectoriel pour déterminer la similitude ou le niveau de correspondance entre ces entités. Dans le domaine de la recherche d'informations, par exemple, ce modèle peut être appliqué pour comparer les termes d'une requête aux termes présents dans un document dans le but de déterminer la pertinence du document étant donnée cette requête. On utilise aussi le modèle vectoriel pour classifier des entités, faire du *clustering*, etc. (voir Manning et al., 2008). Dans le cas qui nous occupe, nous voulons comparer des profils d'intérêts qui, étant décrits par une distribution de fréquences de mots-clés, se représentent bien sous la forme de vecteurs. Nous allons donc simplement appliquer le modèle d'espace vectoriel pour définir notre fonction de similarité entre deux profils.

La figure 3.22 (a) illustre deux distributions de fréquences de tags représentant deux profils différents, p_1 et p_2 où les tags sont représentés par des lettres majuscules (A, B, D...). Pour pouvoir comparer les profils, la première étape consiste à les transformer en vecteurs d'un même espace vectoriel tels qu'illustrés à la figure 3.22 (b). En faisant l'union des tags de p_1 et de p_2, on obtient un espace vectoriel commun à neuf dimensions, le vecteur $\vec{V}(p_1)$ représentant le profil p_1 et le vecteur $\vec{V}(p_2)$ représentant le profil p_2. Pour les tags qui

n'apparaissent pas dans la distribution des fréquences d'un profil, la fréquence indiquée dans le vecteur correspondant est tout simplement égale à zéro.



p_1

Tag	Fréq.
A	7
B	13
E	4
J	10
K	8
L	2

p_2

Tag	Fréq.
B	8
D	6
E	17
J	14
K	3
L	5
Q	11
R	9

(a)

	$\vec{V}(p_1)$	$\vec{V}(p_2)$
A	7	0
B	13	8
E	4	17
J	10	14
K	8	3
L	2	5
D	0	6
Q	0	11
R	0	9

(b)

Figure 3.22 Représentation vectorielle de profils d'intérêts.

Un vecteur-profil, ainsi construit à partir de la distribution des fréquences de tags, tient donc compte de l'importance relative des tags qui constituent ce profil. Dans ce contexte vectoriel, on peut maintenant calculer la distance entre les vecteurs pour en tirer un score de similarité. Plus la distance entre deux vecteurs est grande, moins ils sont similaires. Pour déterminer cette distance, nous calculons le cosinus de l'angle entre les deux vecteurs normés (normalisé sur leur longueur, leur norme) qui est égal au produit scalaire des deux vecteurs. Plus l'angle est grand entre deux vecteurs, plus ils sont distants l'un de l'autre et plus le cosinus de cet angle est petit. Pour un angle maximum de 90 degrés, le cosinus est de 0 et pour un angle minimum de 0 degré, le cosinus est égal à 1 et indique une similarité parfaite.

La similarité entre deux profils p_1 et p_2 peut donc se calculer comme suit :

$$\text{sim}(p1,p2) = \frac{\vec{V}(p_1)}{\left|\vec{V}(p_1)\right|} \bullet \frac{\vec{V}(p_2)}{\left|\vec{V}(p_2)\right|} .$$

La transformation des vecteurs en vecteurs unitaires sert à compenser l'effet du nombre possiblement distinct de tags dans chaque profil. Ce faisant, nous éliminons toute l'information indésirable concernant la longueur différente des listes de tags associés à chaque profil.

Lorsque nous calculons la similarité entre les deux profils illustrés à la figure 3.22 à l'aide de l'équation ci-dessus, nous obtenons une similarité égale à 0.60. Pour montrer l'impact des fréquences des tags sur le calcul de la similarité, modifions la fréquence du tag B dans chaque profil. Lorsque fréq(B) dans p_1 = 100 et que fréq(B) dans p_2 = 2, la similarité diminue de beaucoup pour atteindre la valeur 0.16. Le tag B est ici extrêmement important dans p1, mais pratiquement négligeable dans p_2. Cette grande différence se reflète dans le calcul de la similarité. D'un autre côté, lorsque fréq(B) dans p_1 = 100 et que fréq(B) dans p_2 = 90, la similarité augmente à la valeur 0.97. La grande popularité d'un tag dans les deux profils a pour effet de les rapprocher l'un de l'autre.

Finalement, pour tenir compte de la limite d'équivalence (discutée à la l'article 3.4.1), dans notre méthode de calcul de la similarité entre deux profils, nous ajoutons une simple condition : on calcule d'abord la similarité à l'aide de l'équation ci-dessus puis, si cette similarité est plus grande ou égale à la limite d'équivalence, la méthode retourne 1, sinon, elle retourne la valeur de la similarité effectivement calculée.

3.5 Conclusion

Dans ce chapitre, nous avons d'abord présenté notre modèle de socialisation en décrivant les règles d'évolution qui le régissent, par analogie sociocognitive. Nous avons ensuite justifié notre choix de construire les communautés d'intérêt par regroupement autour de nœuds pivots. Nous avons alors expliqué les algorithmes qui implémentent les règles d'évolution de notre modèle (connexion, déconnexion, socialisation et mise à jour), selon deux variations de ce type de regroupement, la deuxième tenant compte de la capacité limitée des individus en regard du nombre de relations qu'ils peuvent entretenir.

Bien que notre modèle s'inspire et se base clairement sur l'observation des mécanismes de socialisation au sein de nos réseaux sociaux usuels, nous avons décidé, lorsqu'un choix s'imposait, de privilégier l'aspect fonctionnel du modèle, sans toutefois trop s'éloigner d'une vraisemblance sociale que nous pensons très raisonnable du point de vue de la modélisation sociocognitive.

Puis, nous avons discuté d'une représentation des profils d'intérêts plus plausible que la représentation numérique utilisée pour expliquer nos algorithmes et nous avons détaillé les méthodes que nous utiliserons pour calculer et comparer les profils d'intérêts des acteurs qui peupleront nos simulations, au chapitre 4.

Comme nous l'avons déjà mentionné brièvement au chapitre 1, nous avons fait le choix, dans le cadre de cette thèse, de considérer des profils d'intérêts qui se concentrent autour d'un seul grand domaine d'intérêt par exemple, soit les voyages, soit la musique, soit la programmation Web, soit l'astronomie, etc. Cependant, en général, les individus possèdent plusieurs centres d'intérêt : on peut aimer à la fois les voyages, la musique et la programmation Web. Cet état de fait pose problème lorsqu'on compare les individus entre eux. En effet, si Charlotte aime la musique et l'astronomie et que Vincent aime l'astronomie et les voitures, notre méthode de comparaison, comme nous l'avons expliquée, retournera une valeur de similarité moyenne et possiblement sous la limite d'équivalence fixée. Toutefois, ces deux individus gagneraient possiblement à partager leurs intérêts musicaux.

Pour résoudre ce problème, nous pourrions par exemple, regrouper les intérêts disparates d'un individu en différents domaines. Ces regroupements pourraient se faire explicitement par l'individu ou automatiquement, à l'aide d'une méthode de *clustering*, par exemple. Les algorithmes de *clustering*, souvent utilisées en recherche d'informations, servent à regrouper ensemble les éléments qui se ressemblent, au sein d'une collection. Le but est d'obtenir des *clusters* d'éléments qui sont très similaires entre eux (forte cohérence interne), mais clairement différents les uns des autres (voir Manning et al., 2008).

On pourrait alors considérer qu'un individu a plusieurs profils d'intérêts et que lorsqu'il se connecte au réseau, il crée un lien pour chacun de ses profils et lorsqu'il exécute un parcours de socialisation, il le fait toujours par rapport à l'un de ses profils (en changeant de profil lors de diverses socialisations). Un acteur appartiendrait à autant de communautés qu'il a de profil d'intérêts. Il posséderait une collection de voisins différente pour chacun de ses profils, comme s'il possédait un avatar par domaine d'intérêts, qui participerait à la dynamique du réseau au nom du profil qu'il représente. Évidemment, cette gestion de profils multiples demanderait certains ajustements au niveau des algorithmes d'évolution du réseau cependant, les principes généraux qui régissent notre modèle de socialisation seraient les mêmes.

Au chapitre suivant, nous validons notre modèle de socialisation, par simulation, selon les hypothèses que nous avons posées au chapitre 1, du point de vue fonctionnel ainsi que dans l'optique de modélisation sociocognitive.

CHAPITRE IV

EXPÉRIMENTATIONS ET RÉSULTATS

4.1 Introduction

Dans ce chapitre, nous validons nos hypothèses de recherche en faisant l'analyse des réseaux générés par simulation de notre modèle de socialisation (avec regroupements autour de pivots chaînés).

D'un point de vue informatique/fonctionnel, nous avons fait l'hypothèse que nos mécanismes de socialisation favorisent la formation de communautés d'intérêts. De plus, nous avons suggéré qu'en contrôlant l'aspect structural du réseau, nos algorithmes d'évolution de réseaux (connexions, déconnexions, socialisations et mise à jour des relations) peuvent être implémentés de telle sorte qu'ils forment et maintiennent des réseaux qui avantagent la navigabilité c'est-à-dire l'efficacité des parcours dans le réseau : efficacité en terme de longueur et en terme de capacité à localiser les individus similaires dans le réseau.

D'un point de vue sociocognitif, nous avons fait l'hypothèse que les mécanismes de socialisation peuvent servir à expliquer, du moins en partie, comment se forment les réseaux pour en venir à posséder les propriétés structurales particulières qu'on observe dans les réseaux sociaux réels. De plus, nous avons supposé que les mécanismes de socialisation favorisent, dans une certaine mesure, la création de capital social.

Nous montrons tout d'abord de quelle manière nous avons créé les acteurs (et leur profil d'intérêts) qui serviront à peupler nos simulations. Ensuite, nous expliquons les divers paramètres du modèle qui serviront à composer différents scénarios de simulation. Plus précisément, pour vérifier nos hypothèses au niveau informatique, nous composerons des scénarios de simulation avec socialisation et sans socialisation. Nous comparerons ensuite les

résultats obtenus des deux types de scénarios pour voir dans quelle mesure nos mécanismes de socialisation favorisent effectivement la formation de communautés d'intérêts. Pour vérifier nos hypothèses du point de vue sociocognitif, d'abord, nous ne considérerons que les scénarios avec socialisation et nous analyserons les réseaux générés afin de comparer leurs propriétés structurales à celles qu'on observe dans nos réseaux sociaux habituels. Ensuite, pour tester si nos mécanismes de socialisation favorisent effectivement la création de capital social, nous comparerons encore les scénarios avec socialisation, et sans socialisation.

Finalement, nous présentons les différentes mesures (et les résultats attendus) que nous utiliserons pour valider nos hypothèses, puis nous présentons et discutons les résultats obtenus par simulation de notre modèle de socialisation.

4.2 Collecte de données pour la création de profils d'intérêts

Puisque nous testons notre modèle par simulation, nous avons besoin d'acteurs dans nos réseaux. Comme nous l'avons mentionné au chapitre 3 (art. 3.4.2), la similarité entre les acteurs est calculée en comparant leurs profils d'intérêts qui sont spécifiés par une collection de mots-clés. Pour obtenir de tels profils, nous avons récupéré de nombreux utilisateurs sur le site Web Delicious (http://delicious.com/).

Delicious est un système de *tagging* collaboratif ; un outil social (en ligne) de partage de signets (ou favoris) c.-à-d. d'URL de pages Web considérées intéressantes. Lors de la publication d'un signet, l'utilisateur doit lui assigner différents mots-clés (tags) qui décrivent le contenu du site adressé par ce signet. Ainsi, on retrouve sur Delicious des individus associés à une collection de signets partagés, dont on peut raisonnablement supposer qu'ils représentent leurs intérêts, et qui sont eux-mêmes liés avec une collection de mots-clés.

En agrégeant les mots-clés associés à tous les signets d'un utilisateur, on obtient une collection de termes, chaque terme étant associé à une fréquence obtenue en calculant le nombre de fois que cet utilisateur a utilisé ce terme pour étiqueter les signets de sa collection.

Comme nous l'avons vu, une telle collection de termes/fréquences (décrite par un vecteur) sert à représenter le profil d'un acteur dans notre modèle.

Le site Delicious offre une API qu'on peut utiliser pour interroger sa base de données via le protocole HTTP. Les résultats des requêtes sont retournés dans le format JSON, qui est un format textuel permettant de représenter l'information de manière structurée (http://www.json.org).

Par exemple, la requête suivante http://feeds.delicious.com/v2/json/tag/design?count=100 retourne une collection de 100 signets (URL) publiés ayant été étiquetés du mot-clé "design". Chaque item de la collection retournée est accompagné d'autres informations comme l'auteur de la publication, la date de publication, les autres tags associés à cette publication, etc. Divers formats de requête sont disponibles pour obtenir d'autres informations particulières, par exemple, sur les utilisateurs, les réseaux de contacts personnels des utilisateurs, les signets les plus populaires, les plus récents, etc.

Nous avons donc ciblé un certain nombre d'utilisateurs de Delicious ayant plusieurs tags en commun, pour obtenir un sous-ensemble d'utilisateurs susceptibles de montrer une certaine similarité. Nous avons répété ce processus avec des mots-clés différents (représentant des intérêts différents) pour obtenir d'autres sous-ensembles d'utilisateurs potentiellement similaires jusqu'à l'obtention d'une compilation d'utilisateurs de Delicious assez substantielle pour construire une collection de quelques centaines d'acteurs que nous utiliserons dans nos simulations.

Plus précisément, pour chaque domaine d'intérêt, nous avons d'abord déterminé une liste de mots-clés (tags) décrivant ce domaine. Ensuite, nous avons interrogé la base de données de Delicious pour obtenir un ensemble d'utilisateurs ayant étiqueté certains de leurs signets publiés d'un ou de plusieurs tags contenus dans cette liste. Puis, nous avons filtré la collection de signets de chaque utilisateur obtenu en éliminant les signets non pertinents pour le domaine d'intérêt visé (ne contenant aucun des mots-clés de la liste). Finalement , nous n'avons conservé que les utilisateurs dont la collection, après filtrage, contenait encore au

moins 25 signets. Nous avons répété ce processus sur six grands domaines d'intérêt : le design Web, l'astronomie, les réseaux sociaux, l'aquariophilie, les voyages et le Web sémantique. Par exemple, les listes suivantes montrent trois des domaines que nous avons ciblés, associés à quelques-uns des mots-clés utilisés pour décrire ces domaines.

Développement Web	**Astronomie**	**Réseaux sociaux**
jquery	astronomy	social
css	solar	network
webdesign	system	analysis
wordpress	nasa	graph
webdevelopment	planet	math
xml	exploration	sociology
html	discovery	expertise
javascript	moon	community
joomla	exploration	expert
webdesign	planetarium	collaboration
web design	telescope	knowledge
drupal	mission	community
etc.	etc.	etc.

Six domaines d'intérêt ne semblent pas beaucoup, mais étant donné la variété des mots-clés utilisés pour décrire nos domaines, les utilisateurs collectés pour un domaine ne sont pas nécessairement tous fortement similaires entre eux. Pour le domaine de l'astronomie, par exemple, certains individus s'intéressent plus particulièrement au système solaire, aux planètes et aux étoiles tandis que d'autres préfèrent l'exploration et les missions spatiales. C'est pourquoi on observe beaucoup plus de communautés d'intérêts dans nos réseaux que de domaines sur lesquels on a collecté nos utilisateurs. Notre objectif, ici, était d'obtenir une collection d'utilisateurs qui nous permette bien de faire des regroupements d'utilisateurs basés sur les mots-clés (intérêts) communs et non que ces regroupements correspondent exactement aux domaines ciblés.

Bien que nous n'abordions pas, dans ce travail de recherche, les problèmes liés à la polysémie/désambiguïsation et à la synonymie lorsqu'il s'agit de déterminer la similarité entre deux collections de termes, nous avons tout de même effectué un traitement supplémentaire de troncature (*stemming*) sur les mots-clés contenus dans les profils recueillis.

En effet, nous voulons que la comparaison entre les termes "reconnaisse" les diverses formes fléchies d'un mot. Par exemple, les mots "organisation" (au singulier) et "organisations" (au pluriel) devraient être considérés comme identiques lors d'une comparaison. Aussi, des mots de même famille comme "bibliographie" et "bibliographique" ont un sens similaire et devraient être considérés comme tels lors des comparaisons. La troncature sert à réduire les formes fléchies et les mots de même famille à une forme de base commune en coupant la fin des mots. Ainsi, les mots "organisation" et "organisations" seront réduits à la forme "organis", par exemple, et les mots "bibliographie" et "bibliographique", à la forme "bibliograph" (la forme de base obtenue pouvant varier quelque peu selon l'algorithme de troncature utilisé).

Il faut noter cependant que la troncature est un processus heuristique tout de même assez approximatif et ne fonctionne pas toujours comme souhaité : on obtient parfois une forme de base identique pour deux mots ayant pourtant des sens complètement différents. Les mots "organisme" et "organiste", par exemple, sont réduits à la forme élémentaire "organ". Dans ce cas, les deux termes seront faussement considérés comme identiques lors d'une comparaison. Néanmoins, en général, on gagne à utiliser la troncature lors de la comparaison de mots.

La lemmatisation est une autre méthode beaucoup plus performante de réduction des mots à une forme de base commune, mais celle-ci demande une connaissance préalable plus poussée de la grammaire dans la langue traitée. En gros, cette méthode ramène tous les mots à leur forme de base, comme on les retrouve dans les entrées d'un dictionnaire : les verbes sont ramenés à l'infinitif, les adjectifs sont simplifiés dans leur forme du masculin singulier, les noms sont mis au singulier, etc. Comme notre objectif n'est pas d'améliorer la performance des méthodes de comparaison et que la lemmatisation est plus complexe et plus lourde à utiliser, nous nous sommes limités, dans nos travaux, à la troncature (*stemming*).

L'algorithme de troncature le plus connu, pour traiter l'anglais, est celui de Martin Porter (Porter, 1980). Nous avons utilisé une implémentation Java de cet algorithme, fourni par le projet Snowball (dont Porter fait partie) sous licence BSD : http://snowball.tartarus.org/index.php.

4.3 Paramètres du modèle

Nous avons prévu différentes variables pour paramétrer nos simulations. Le tableau suivant fournit une explication de chacun de ces paramètres.

Tableau 4.1
Description des paramètres du modèle

Limite d'équivalence [0..1]
Détermine une limite inférieure sur la similarité entre deux profils pour que ceux-ci soient considérés comme identiques. Ce paramètre nous permet de jouer sur le niveau de structure qu'on retrouve dans nos données de simulation. Si l'on impose une limite d'équivalence égale à 0.9, par exemple, on ne trouvera pratiquement pas d'utilisateurs similaires entre eux et notre collection d'utilisateurs sera dite non structurée : lorsque personne n'a d'intérêts communs, la formation de communautés d'intérêts devient impossible. À l'opposé, une limite d'équivalence égale à 0.1 produira une collection d'acteurs très fortement structurée avec une forte probabilité de formation de communautés très fournies.
Nombre maximum d'acteurs [0..cardinalité de notre collection d'acteurs]
Ce paramètre limite le nombre d'acteurs dans le réseau, à tout moment de la simulation. La limitation du nombre d'acteurs est utile surtout pour des raisons de performance, étant donné que nous effectuons périodiquement de nombreux calculs (statistiques et mesures pour la validation) au cours des simulations. Dans plusieurs cas, la taille du réseau affecte les calculs de façon exponentielle.
Nombre minimum de voisins [3..Nombre maximum de voisins]
Ce paramètre sert à déterminer la capacité (en terme de nombre de voisins) pour chacun des acteurs du réseau : lorsqu'on crée un acteur, on lui attribue aléatoirement une valeur pour sa capacité qui se situe entre ce nombre minimum de voisins et le nombre maximum de voisins (paramètre suivant). Le nombre minimum de 3 est imposé pour le bon fonctionnement de nos algorithmes. En effet, un acteur dans le réseau doit pouvoir supporter au moins trois voisins (lorsqu'il est un pivot) : un suivant, un précédent et un lien vers une autre communauté.
Nombre maximum de voisins [Nombre min. voisins..Nombre acteurs dans le réseau - 1]
Ce paramètre sert à déterminer la capacité (en terme de nombre de voisins) des acteurs du réseau : lorsqu'un acteur est créé, on lui attribue aléatoirement une valeur pour sa capacité qui se situe entre le nombre minimum de voisins (paramètre précédent) et ce nombre maximum de voisins (paramètre suivant).

Condition d'arrêt (booléen)

Condition booléenne qui détermine le moment d'arrêt de la simulation.

Probabilité de connexion [0..1]

Probabilité qu'à chaque pas de temps, un acteur (choisi aléatoirement dans la collection d'acteurs) se connecte au réseau.

Peu importe la valeur de cette probabilité, une connexion ne s'effectue jamais lorsque le nombre d'acteurs dans le réseau est déjà égal à la cardinalité de notre collection d'acteurs (on n'a plus aucun acteur à connecter) ou qu'il est égal à la valeur donnée au paramètre **nombre maximum d'acteurs** pour cette simulation.

Probabilité de déconnexion [0..1]

C'est la probabilité qu'à chaque pas de temps, un acteur (choisi aléatoirement dans le réseau) se déconnecte du réseau (sauf si le réseau est vide).

Les acteurs déconnectés retournent dans la collection d'acteurs et redeviennent disponibles pour une connexion ultérieure.

Probabilité de socialisation [0..1]

C'est la probabilité qu'à chaque pas de temps, un acteur (choisi aléatoirement dans le réseau) effectue un parcours de socialisation.

Probabilité de mise à jour des relations [0..1]

C'est la probabilité qu'à chaque pas de temps, un acteur (choisi aléatoirement dans le réseau) effectue une mise à jour de ses relations (déconnexion suivie d'une reconnexion).

L'intervalle des valeurs possibles pour chaque paramètre est indiqué, entre crochets, à côté du nom du paramètre.

Nous décrivons maintenant les différents scénarios de simulations que nous avons composés pour valider notre modèle.

4.4 Scénarios de simulations

4.4.1 Valeurs des paramètres du modèle pour les divers scénarios

Un scénario de simulation est composé de l'ensemble des paramètres décrits ci-dessus auxquels on a assigné des valeurs précises pour étudier le comportement du modèle dans des conditions particulières. Étant donné la nature aléatoire de notre modèle, deux exécutions

d'un même scénario ne produisent pas le même réseau. C'est pourquoi nous avons choisi d'exécuter chaque scénario 100 fois pour nous assurer de la constance des propriétés étudiées.

Tout d'abord, nous voulons observer le comportement de notre modèle lorsqu'il y a plus ou moins d'acteurs qui se ressemblent dans le réseau et qui peuvent donc être plus ou moins regroupés en communautés d'intérêts. Nous voulons voir comment le niveau de structuration de notre collection d'acteurs influence le comportement du modèle. Pour ce faire, nous avons déterminé trois niveaux de structuration des données en faisant varier la valeur de la limite d'équivalence. Nous avons fixé une valeur égale à 0.2 pour avoir des données fortement structurées, une valeur de 0.6 pour des données moyennement structurées et une valeur de 0.8 pour des données faiblement structurées. Ainsi, avec une limite d'équivalence de 0.8, pour que deux acteurs soient considérés comme identiques, leur similarité doit être plus grande ou égale à 0.8, ce qui est beaucoup plus sévère qu'une limite d'équivalence égale à 0.2. Dans le premier cas (0.8), on aura moins d'acteurs similaires (parmi les acteurs de notre collection) que dans le deuxième cas (0,2). Nous avons testé différentes valeurs pour la limite d'équivalence et nous avons choisi ces valeurs parce qu'elles illustrent bien les différences dans le comportement du modèle en fonction du niveau de structuration des données.

Donc, nous avons déjà 3 scénarios de simulation différents qui utilisent la même collection d'acteurs, mais avec des limites d'équivalence de valeurs différentes : scénario 1 (données fortement structurées), scénario 2 (données moyennement structurées) et scénario 3 (données faiblement structurées).

Ensuite, comme nous voulons évaluer la performance des mécanismes de socialisation, nous désirons comparer le comportement du modèle a) lorsqu'il n'y a aucune socialisation et b) lorsqu'il se produit des socialisations, au cours de l'évolution du réseau. Nous avons donc dédoublé les 3 scénarios précédents en faisant varier leur probabilité de socialisation : une probabilité égale à 0 pour que le modèle n'effectue jamais de socialisation (scénarios a) et une probabilité égale à 1 pour qu'à chaque pas de temps, un acteur du réseau soit choisi au hasard pour effectuer un parcours de socialisation (scénarios b). Nous avons donc créé en tout 6 scénarios de simulations, résumés ci-dessous dans le tableau 4.2.

Tableau 4.2
Sommaire des scénarios de simulations

Scénario 1	**Données fortement structurées (lim. équiv. = 0.2)**
a	sans socialisation
b	avec socialisation
Scénario 2	**Données moyennement structurées (lim. équiv. = 0.6)**
a	sans socialisation
b	avec socialisation
Scénario 3	**Données faiblement structurées (lim. équiv. = 0.8)**
a	sans socialisation
b	avec socialisation

Les deux seuls paramètres dont la valeur varie dans chaque scénario sont présentés dans le tableau 4.3.

Tableau 4.3
Valeurs des paramètres qui varient d'un scénario à l'autre

Paramètre	**Valeur dans les scénarios :**					
	1a	**1b**	**2a**	**2b**	**3a**	**3b**
Limite d'équivalence	0.2	0.2	0.6	0.6	0.8	0.8
Probabilité de socialisation	0	1	0	1	0	1

Le reste des paramètres du modèle possèdent des valeurs identiques pour tous les scénarios et sont montrés dans le tableau 4.4. Chaque réseau est initialement vide et nous avons fixé une valeur pour la probabilité de connexion supérieure à la valeur de la probabilité de déconnexion, pour simuler des réseaux ouverts dont la taille augmente au cours du temps. Étant donné qu'un utilisateur peut vouloir changer de communauté d'appartenance parce que ses intérêts ont changé au cours du temps, nous avons fixé la probabilité de mise à jour des relations à 0.1 pour modéliser ce phénomène qui se produit parfois, mais pas fréquemment.

Pour modéliser le fait que chaque acteur possède une limite sur le nombre de relations qu'il peut entretenir, mais que certains sont plus sociaux que d'autres, nous faisons varier la taille maximum des réseaux personnels des acteurs entre 10 et 50. Cela signifie qu'à chaque fois qu'un acteur est créé, on lui assigne aléatoirement une valeur entre 10 et 50 qui représente le nombre maximum de voisins immédiats qu'il peut avoir à tout moment, au cours de la simulation.

De plus, étant donné la complexité computationnelle de certaines mesures que nous effectuons périodiquement en cours des simulations (à des fins d'analyse seulement), nous avons choisi de limiter à 500 le nombre d'acteurs dans le réseau. Ceci implique qu'on pourra observer, dans nos simulations, le comportement de notre modèle en période de croissance de la taille du réseau (jusqu'à 500 acteurs) et en période de constance de la taille du réseau (autour de 500 acteurs). En effectuant des tests préliminaires, nous avons remarqué que les réseaux atteignent leur taille maximale autour du pas de temps 1000. Dans cette optique, nous avons fixé la condition d'arrêt de la simulation à l'atteinte du pas de temps 2000. De cette manière, nous obtenons une période de croissance et de constance de longueur à peu près similaire. Notons ici que dans la période de constance de la taille du réseau, il se produit toujours des connexions et déconnexions sauf que les connexions se produiront moins souvent étant donné qu'une déconnexion devra avoir eu lieu avant qu'une connexion puisse advenir de nouveau. La collection d'acteurs que nous utilisons dans nos simulations comporte 553 acteurs, mais un maximum de 500 d'entre eux pourra être présent, au même moment, dans le réseau.

Tableau 4.4
Valeurs des paramètres communs à tous les scénarios

Paramètre	**Valeur dans tous les scénarios**
Nombre maximum d'acteurs	500
Nombre minimum de voisins	10
Nombre minimum de voisins	50
Condition d'arrêt	L'atteinte de 2000 pas de temps
Probabilité de connexion	0.8
Probabilité de déconnexion	0.3
Probabilité de mise à jour	0.1

4.4.2 Fonctionnement général des simulations

Les simulations sont effectuées de manière séquentielle. À tous les pas de temps, il peut se produire :

Une déconnexion

La déconnexion se produit selon la probabilité de déconnexion déterminée dans les paramètres du scénario de simulations, et seulement si le réseau n'est pas vide. Lorsqu'une déconnexion a lieu, on choisit au hasard, un acteur que l'on déconnecte du réseau, en exécutant notre algorithme de déconnexion (voir par. 3.3.2.3) et l'on conserve cet acteur dans une autre collection (la collection des acteurs déconnectés).

Une connexion

La connexion se produit selon la probabilité de connexion déterminée dans les paramètres du scénario et lorsque le réseau n'a pas atteint sa taille maximale (aussi fixée dans les paramètres du scénario). Lorsqu'une connexion se produit, on doit choisir un acteur à ajouter dans le réseau. Cet acteur est sélectionné au hasard soit dans notre collection initiale d'acteurs, soit dans la collection des acteurs déconnectés. Dans les deux cas, l'acteur choisi est retiré de la collection dans laquelle il a été sélectionné. On choisit la collection dans laquelle sélectionner notre acteur à connecter de cette manière : si la collection d'acteurs déconnectés est vide, choisir l'acteur à connecter dans la collection initiale sinon, choisir l'acteur à connecter dans la collection initiale selon la probabilité P donnée par :

$$P(\text{coll. init.}) = \frac{\text{nombre d'acteurs restants dans la collection initiale}}{\text{nombres initial d'acteurs dans la collection initiale}}.$$

De cette manière, on favorise d'abord le choix de l'acteur dans la collection initiale tant que celle-ci est assez grande, mais à mesure que les acteurs déconnectés s'accumulent (et que le nombre d'acteurs dans la collection initiale diminue), les chances de choisir l'acteur dans la collection des acteurs déconnectés augmentent.

On exécute ensuite notre algorithme de connexion (voir par. 3.3.2.3)

Une socialisation

La socialisation se produit selon la probabilité de socialisation déterminée dans les paramètres du scénario et seulement lorsque le réseau contient plus d'un acteur. Lorsque la socialisation se produit, on choisit aléatoirement dans le réseau, un acteur qui sera le socialisateur, et un autre acteur qui sera le point de départ du parcours de socialisation du socialisateur. On exécute ensuite l'algorithme de socialisation (voir par. 3.3.2.3).

Une mise à jour des relations

La mise à jour s'effectue selon la probabilité de mise à jour des relations déterminée dans les paramètres du scénario. Lorsque la mise à jour se produit, on sélectionne aléatoirement un acteur dans le réseau et l'on exécute notre algorithme de mise à jour des relations (voir par. 3.3.2.3).

À la section suivante, nous présentons les mesures que nous utiliserons pour valider notre modèle de socialisation.

4.5 Mesures utilisées et résultats attendus

Pour valider nos hypothèses du point de vue fonctionnel (informatique) et sociocognitif, nous utiliserons, entre autres, certaines des mesures présentées au chapitre 2 (art. 2.3.1). Notons que toutes les mesures utilisées ont été intégrées au modèle à des fins de vérification et d'analyse seulement et sont tout à fait indépendantes des algorithmes d'évolution du réseau qui régissent notre modèle.

4.5.1 Niveau fonctionnel

Les mesures présentées dans cette section serviront tout d'abord à vérifier que les propriétés structurales imposées sont maintenues dans les réseaux et à évaluer leur effet, que nous

supposons positif, sur la navigabilité. Nous présentons ensuite des mesures pour évaluer la capacité de nos mécanismes de socialisation à former des communautés d'intérêts.

4.5.1.1 Mesures de vérification des propriétés structurales imposées

Tout d'abord, après chaque connexion, déconnexion, socialisation et mise à jour des relations, nous avons inclus plusieurs procédures qui inspectent le réseau et s'assurent que les propriétés structurales imposées sont maintenues. Plus précisément, ces procédures vérifient que :

1. le réseau est connexe.
2. les pivots sont soit reliés à des nœuds simples de même profil, soit reliés à des nœuds pivots de profils différents, soit reliés à au plus deux pivots ayant le même profil c'est-à-dire le suivant ou le précédent.
3. les nœuds simples ne possèdent qu'un seul lien avec un pivot de même profil.
4. tous les pivots qui sont les précédents d'un autre pivot sont toujours pleins.
5. le degré de chaque nœud ne dépasse jamais sa capacité.
6. les voisins d'un pivot qui sont eux-mêmes des pivots de profils différents (donc ni le suivant ni le précédent) ne sont jamais de même profil entre eux.

Nous avons vérifié ces propriétés maintes et maintes fois lors d'innombrables simulations pour tester nos algorithmes et ces propriétés structurales sont effectivement maintenues tout au long de l'évolution du réseau. Lors de l'analyse des résultats, nous les tiendrons donc pour acquis.

4.5.1.2 Mesure de la densité

La densité mesure la proportion du nombre de liens dans le réseau par rapport aux nombres de liens possibles dans le réseau. Elle indique donc si le graphe est creux (densité faible) ou dense (densité élevée).

Comme nous l'avons expliqué au chapitre 3, une densité faible est désirable, car dans les réseaux creux (non dense), les possibilités de chemins différents dans le réseau sont bien moindres et les parcours risquent ainsi d'être moins longs. Évidemment, plus on a de liens dans un réseau (plus la densité est élevée), plus la distance (le plus court chemin) entre deux nœuds tend à être courte. Cependant, dans un contexte complètement décentralisé, où seule l'information locale est disponible, on ne connaît pas, a priori, le plus court chemin, et la redondance des parcours possibles, dans un réseau dense, multiplie la possibilité des longs chemins.

Nous nous attendons à ce que les propriétés structurales imposées produisent des réseaux de faible densité faible (proche de la densité minimale requise pour assurer la connexité du réseau) pour favoriser la navigabilité. Bien qu'une densité faible soit désirable, il faut cependant que de courts chemins entre les nœuds existent dans la structure du réseau. La densité à elle seule n'implique pas cette propriété. Les mesures de distance suivantes, par contre, permettent de caractériser la distance entre les nœuds dans un réseau.

4.5.1.3 Mesure du diamètre et de la distance moyenne

Le **diamètre** d'un graphe est la distance maximum parmi toutes les distances qui séparent n'importe quelles paires de nœuds dans le réseau, la distance entre deux nœuds étant la longueur du plus court chemin entre ces deux nœuds. Un petit diamètre indique que la distance entre les nœuds est petite (de courts chemins existent). On peut aussi calculer la **distance moyenne** (moyenne des distances entre toutes les paires de nœuds dans le réseau) comme indicateur de la présence de courts chemins. Nous évaluerons les deux mesures.

Nous supposons qu'à cause des propriétés structurales imposées, les mesures de distance seront relativement petites dans tous les scénarios. Cependant, nous nous attendons à ce que le diamètre et la distance moyenne des réseaux générés soient encore plus petits dans les scénarios avec socialisation que dans les scénarios sans socialisation.

4.5.1.4 Mesure du nombre de nœuds pivots

Comme nous l'avons montré au chapitre 3, les propriétés structurales imposées dans nos réseaux nous permettent de calculer une limite supérieure sur la distance parcourue dans le réseau, lors des parcours de connexion ou de socialisation. Cette limite correspond au nombre de nœuds pivots dans le réseau. Nous nous attendons à ce que les scénarios avec socialisation produisent des réseaux dont le nombre de nœuds pivots est plus petit que dans les réseaux générés par les scénarios sans socialisation. Les deux mesures suivantes serviront à calculer la capacité du modèle à réunir les individus en communautés d'intérêt.

4.5.1.5 Mesure d'homophilie

L'homophilie, dans notre cas particulier, est une mesure qui calcule la similarité moyenne des acteurs adjacents dans le réseau. Notre mesure d'homophilie calcule donc à quel point sont similaires les noeuds qui sont des voisins les uns des autres dans le réseau. C'est la moyenne de la similarité calculée pour chaque paire de sommets qui sont reliés ensemble. Cette mesure est donc un indice de la capacité de notre modèle à rapprocher les individus de profils similaires, c.-à-d. à former des communautés d'intérêt.

Nous nous attendons donc à ce que la valeur de l'homophilie soit plus élevée dans les réseaux simulés avec socialisation que dans les scénarios sans socialisation.

4.5.1.6 Mesure des rencontres effectuées

Au cours de l'évolution du réseau, de nouveaux acteurs arrivent dans le réseau, d'autres le quittent, certains socialisent et provoquent parfois la fusion de communautés, qui changent et évoluent aussi. À l'intérieur de cette dynamique, si l'on considère qu'un acteur a fait la rencontre de tous les autres acteurs avec qui il a été regroupé en communauté d'intérêt, à un moment ou à un autre, au cours de l'évolution du réseau, on aimerait savoir à quel point les acteurs de profils identiques se rencontrent effectivement par le biais de la formation de ces communautés d'intérêt. C'est ce que cette mesure calcule : la proportion du nombre de

rencontres effectuées sur le nombre de rencontres qui auraient pu être effectuées en considérant les acteurs présents dans le réseau.

Nous nous attendons à ce que la valeur de cette mesure soit plus élevée dans les scénarios avec socialisation que dans ceux sans socialisation.

4.5.2 Niveau sociocognitif

Nous proposons ici des mesures qui nous permettront d'évaluer si les réseaux générés par notre modèle de socialisation possèdent des caractéristiques structurales similaires à celles qu'on observe dans nos réseaux sociaux habituels (réels ou virtuels) : l'effet des petits mondes, l'émergence de communautés structurales, la transitivité et la distribution des degrés en loi de puissance.

4.5.2.1 Mesures de distance

Les mesures de distance (diamètre et distance moyenne) nous serviront à déterminer si nos réseaux générés présentent l'effet des petits mondes comme observé dans la plupart de nos réseaux sociaux. Nous pensons que oui.

4.5.2.2 Mesure de modularité

Nous avons vu, au chapitre 2, qu'on observe généralement l'émergence de communautés structurales au sein des réseaux sociaux. Nous rappelons ici qu'une communauté structurale est un ensemble de nœuds très fortement connectés (dont la densité des liens est élevée, à l'intérieur de cet ensemble) par rapport à la densité des liens qui unissent les communautés entre elles.

La mesure de modularité sert à déterminer à quel point un partitionnement particulier en sous-ensemble de nœuds correspond à la formation de communautés au niveau structural. Nous utiliserons cette mesure pour voir si effectivement, nos regroupements par comparaison des profils (nos communautés basées sur les intérêts) se reflètent bien au niveau structural

(communautés basées sur la structure) comme on l'observe dans la réalité. On considère qu'une valeur de modularité supérieure à 0.3 indique un partitionnement significatif de communautés structurales (Newman, 2004). Nous nous attendons effectivement à ce que les communautés d'intérêt formées au sein de nos réseaux se reflètent bien au niveau structural.

4.5.2.3 Coefficient de clustering

La transitivité est un autre phénomène fréquemment observé dans les réseaux sociaux réels qui indique la présence de petits cercles d'amis dans lesquels tout le monde connaît tout le monde. Nous mesurerons la transitivité en calculant le coefficient de *clustering* au niveau du réseau.

Le *clustering* est une propriété structurale désirable au niveau de la modélisation cognitive, mais que nous avons consciemment réduit au possible dans nos réseaux, au profit d'une meilleure navigabilité, d'un point de vue fonctionnel. Nous nous attendons donc à ce qu'il n'y ait pratiquement pas de *clustering* dans les réseaux générés par notre modèle de socialisation.

4.5.2.4 Mesure de la distribution des degrés

On observe, dans la réalité, que la distribution des degrés de plusieurs réseaux sociaux suit une loi de puissance. On a vu, au chapitre 2, à quoi ressemble une loi de puissance $P(d) = cd^{-\alpha}$ dans un plan cartésien. Une particularité intéressante des lois de puissance est que lorsqu'on prend le logarithme des valeurs sur les deux axes du plan cartésien, on obtient une droite (de pente négative) de la forme $y = mx + b$. Pour déterminer si la distribution des degrés de nos réseaux suit une loi de puissance, nous avons donc ajusté un modèle de régression linéaire par la méthode des moindres carrés dans la représentation cartésienne log-log de la distribution des degrés de nos réseaux. Nous utiliserons le coefficient de corrélation des droites ainsi estimées pour évaluer si la distribution des degrés des réseaux simulés ressemble à une loi de puissance. Nous nous attendons à ce que ce soit le cas.

En regard du capital social, nous avons suggéré que la socialisation est aussi une manière de créer un certain capital social au sein d'un réseau social. Nous présentons ici la mesure que nous utiliserons pour évaluer un aspect particulier du capital social dans nos réseaux.

4.5.2.5 Mesure de proximité (capital social)

Nous avons vu, au chapitre 2, que les scores des différentes centralités permettaient de mesurer à quel point la position d'un acteur dans le réseau lui procurait certains avantages c.-à-d. un certain capital social. Dans notre cas particulier, étant donné que nous avons imposé des caractéristiques structurales particulières à nos réseaux, nous ne pouvons pas utiliser, ici, ni la centralité de degré ni la centralité d'intermédiarité pour discuter du capital social. En effet, on comprend facilement que les nœuds pivots, au sein de nos réseaux, sont d'une part, les plus centraux et d'autre part, les plus intermédiaires puisqu'ils se trouvent sur beaucoup de chemins de longueur minimale entre les acteurs du réseau. Cependant, ces propriétés ont été imposées pour des considérations d'ordre fonctionnel et les nœuds pivots sont choisis automatiquement par nos algorithmes d'évolution de réseaux. Donc, le fait qu'un acteur puisse être un pivot, à un moment ou à un autre au cours de l'évolution du réseau, ne lui attribue aucun avantage particulier dans le cadre du modèle, cette position n'étant qu'une position d'ordre pratique.

Cependant, la distance entre les acteurs de nos réseaux est une propriété émergente et non une caractéristique structurale imposée. Nous pensons donc que la centralité de proximité, comme mesure du capital social, sera plus grande dans les réseaux simulés avec socialisation que dans les scénarios sans socialisation.

Nous venons d'expliquer toutes les mesures qui serviront à l'analyse de nos résultats de simulations que nous présentons dans la section qui suit.

4.6 Analyse des résultats de simulations

Dans tous nos cas d'analyse, nous avons conservé toutes nos mesures, car aucune d'entre elles ne présentait de dispersion problématique ou anormale (aucun *outlier*). Nous donnerons d'ailleurs à chaque fois, la moyenne, l'écart-type, le minimum et le maximum.

4.6.1 Analyse au niveau fonctionnel

4.6.1.1 Étude de la densité du réseau

Nous avons mesuré la densité des réseaux générés par notre modèle à chaque intervalle de 50 pas de temps. Pour chaque scénario, nous avons ensuite calculé la moyenne de la densité, sur les 100 simulations effectuées par scénario pour chaque intervalle de temps. La figure 4.1 montre l'évolution de la densité moyenne au cours du temps, pour chacun des scénarios de simulation. Dans tous les cas, que les données soient structurées ou non, avec ou sans socialisation, on remarque que la densité est très faible et tend à diminuer pour ensuite se stabiliser, autour du pas de temps 1000, vers une valeur en dessous de 0.005.

Il est à noter que lors des simulations, le plafond imposé de 500 nœuds est atteint autour du pas de temps 1000, où la taille du réseau cesse de croître. On observe alors une stabilisation de la densité (dans tous les scénarios). On peut conclure que la densité des réseaux n'augmente pas lorsque la taille du réseau demeure constante.

Le tableau 4.5 indique les valeurs de densité observées au pas de temps 2000 pour les différents scénarios de simulation. On observe des valeurs légèrement plus élevées dans les scénarios b - avec socialisation - que dans les scénarios a - sans socialisation. En effet, les mécanismes de socialisation, qui provoquent la fusion de communautés, créent nécessairement de nouveaux liens.

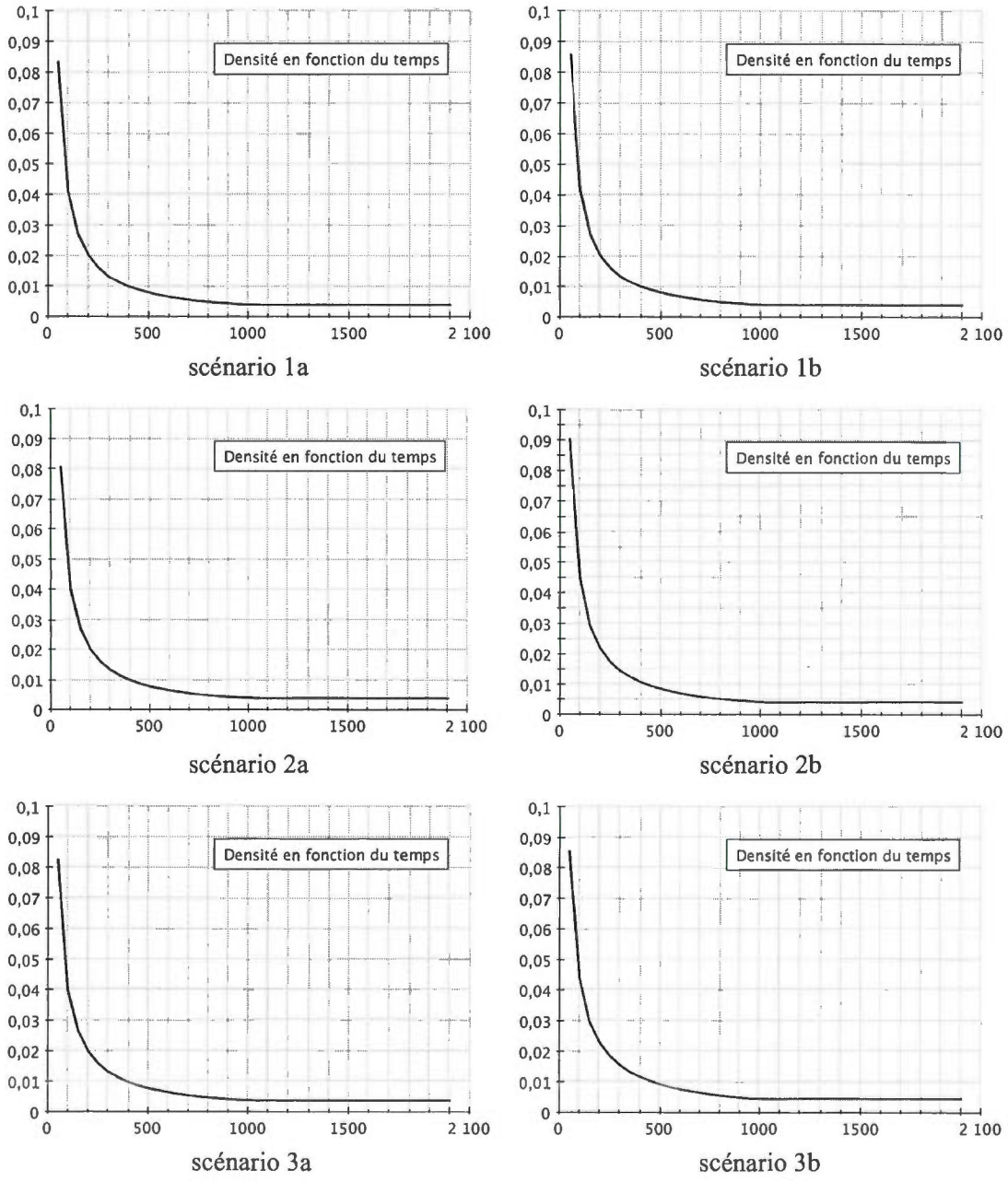


Figure 4.1 Évolution de la densité en fonction du temps.

Les scénarios 1, 2 et 3 illustrent respectivement les scénarios avec données fortement, moyennement et faiblement structurées. Les scénarios (a) montrent les simulations sans socialisation et les scénarios (b), avec socialisation.

On note aussi, dans les scénarios avec socialisation, que plus les données sont structurées, plus la densité est faible. Ceci peut s'expliquer par le fait que les réseaux avec données

faiblement structurées comportent plus de nœuds pivots (car il y a nécessairement plus de petites communautés), et qu'au niveau des nœuds pivots (et non au niveau des nœuds simples), il se crée une certaine redondance des liens, un faible taux de *clustering*, responsable de cette augmentation du nombre de liens dans le réseau. Nous reverrons ce phénomène de *clustering* au paragraphe 4.6.2.3. Néanmoins, la densité demeure très faible dans tous les cas.

Tableau 4.5
Densité pour tous les scénarios au pas de temps 2000

Scénario	Densité			
	moyenne	écart-type	minimum	maximum
1a	0.0040	0.0000	0.0040	0.0040
1b	**0.0041**	0.0000	0.0040	0.0041
2a	0.0040	0.0000	0.0040	0.0040
2b	**0.0043**	0.0001	0.0042	0.0044
3a	0.0040	0.0000	0.0040	0.0040
3b	**0.0047**	0.0001	0.0046	0.0049

Les scénarios 1, 2 et 3 représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées. Les scénarios (a) représentent les simulations sans socialisation et les scénarios (b), avec socialisation.

Nos algorithmes arrivent donc à maintenir un réseau très peu dense, ce qui est désirable pour la navigabilité. Cependant, nous devons aussi vérifier que la distance entre les acteurs du réseau n'est pas trop grande et qu'il existe des chemins de courte distance pour les rejoindre.

4.6.1.2 Calcul de la distance entre les acteurs du réseau

Tout d'abord, nous avons calculé la distance moyenne et le diamètre à chaque intervalle de 50 pas de temps. Puisqu'un diamètre (ou une distance moyenne) égal à 10 dans un réseau de 50 sommets ne signifie pas la même chose qu'un diamètre (ou une distance moyenne) de même valeur dans un réseau de 500 sommets, nous avons ensuite pondéré les valeurs obtenues par la taille du réseau, en les divisant par le nombre de nœuds dans le réseau, au moment des mesures. Puis, comme dans le cas des mesures de densité, pour chaque scénario, nous avons ensuite évalué la moyenne des valeurs sur les 100 simulations effectuées pour chaque scénario, pour chaque intervalle de temps mesuré. La figure 4.2 illustre l'évolution du

diamètre (pondéré) et de la distance moyenne (pondérée) en fonction du temps, pour chaque scénario.

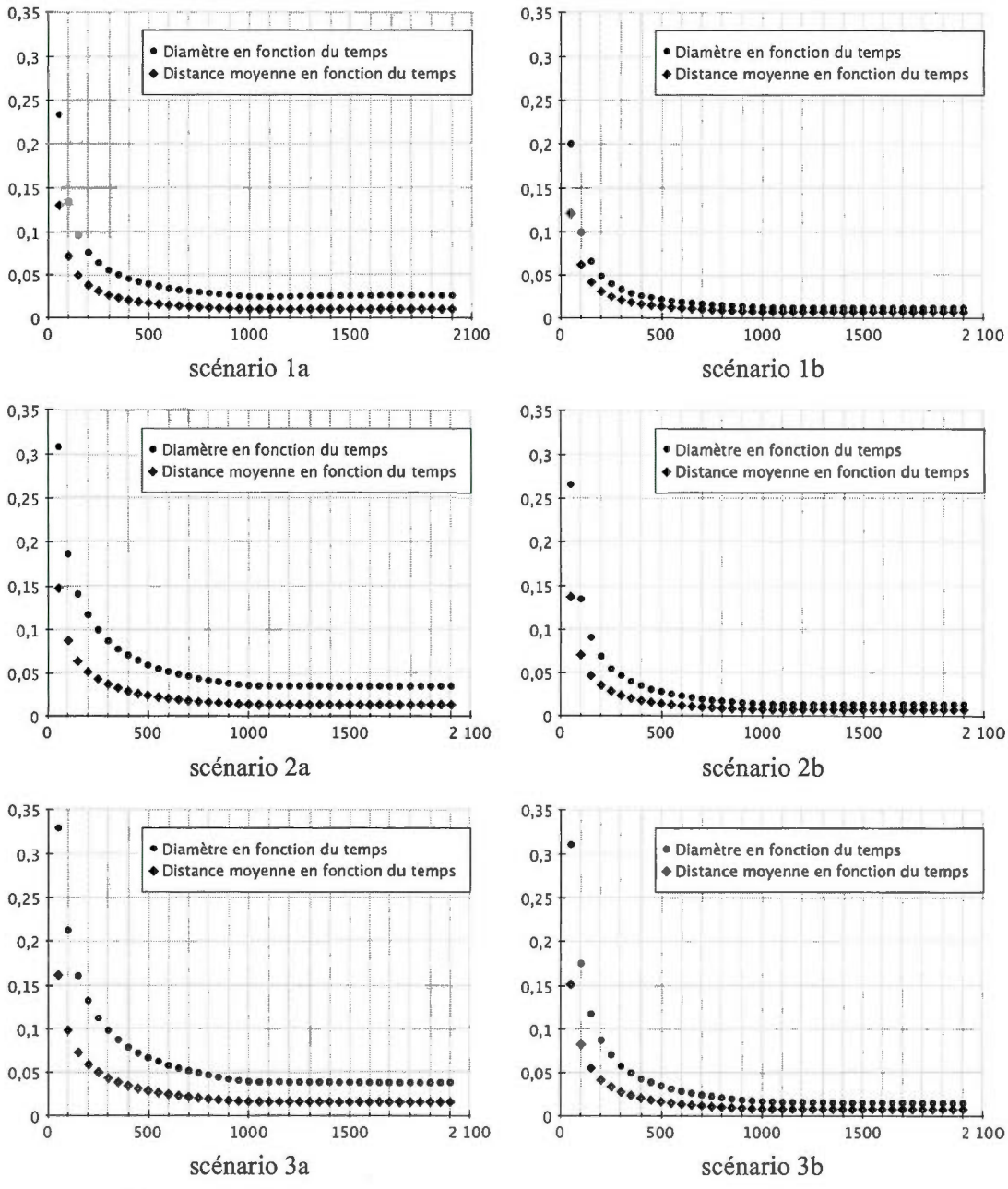


Figure 4.2 Évolution du diamètre et de la distance moyenne pondérés en fonction du temps.

Les scénarios 1, 2 et 3 illustrent respectivement les scénarios avec données fortement, moyennement et faiblement structurées. Les scénarios (a) montrent les simulations sans socialisation et les scénarios (b), avec socialisation.

Dans tous les scénarios, on note une distance qui tend à se stabiliser, encore une fois, autour du pas de temps 1000, lorsque les réseaux cessent de croître. En effet, lorsqu'on regarde nos données de simulations, on constate que le diamètre et la distance moyenne non pondérés par la taille du réseau demeurent à peu près constants tout au long des simulations. Ainsi, en période de croissance du réseau (pas de temps 0 à 1000), les valeurs pondérées diminuent et lorsque la taille du réseau se stabilise (autour du pas de temps 1000), les valeurs pondérées deviennent stables.

Par contre, on observe que dans les scénarios avec socialisation (scénarios b), le diamètre et la distance moyenne tendent à se rejoindre. Cela indique que l'écart entre les distances moyennes pour chaque paire d'acteurs dans le réseau est plus petit, que ces distances se concentrent plus autour de la moyenne, et sont moins dispersées dans le cas des scénarios avec socialisation. Les tableaux 4.6 et 4.7 montrent les valeurs moyennes et l'écart-type du diamètre pondéré et de la distance moyenne pondérée, au pas de temps 2000.

On remarque, d'abord, que les valeurs moyennes sont plus petites dans les scénarios avec socialisation (scénarios b), mais de plus, on note aussi que les écarts-types, dans ces scénarios, sont plus faibles que dans les scénarios sans socialisation (scénarios a). De plus, on observe que ces mesures de distance varient en fonction de la structuration des données : plus les données sont structurées, plus la distance entre les acteurs du réseau est faible.

Tableau 4.6
Diamètre pondéré au pas de temps 2000

Scénario	Diamètre			
	moyenne	écart-type	minimum	maximum
1a	0.0266	0.0037	0.0181	0.0382
1b	**0.0125**	0.0011	0.0100	0.0160
2a	0.0350	0.0039	0.0240	0.0462
2b	**0.0142**	0.0011	0.0120	0.0161
3a	0.0385	0.0037	0.0301	0.0481
3b	**0.0154**	0.0013	0.0140	0.0200

Les scénarios 1, 2 et 3 représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées. Les scénarios (a) représentent les simulations sans socialisation et les scénarios (b), avec socialisation.

Tableau 4.7
Distance moyenne pondérée au pas de temps 2000

Scénario	Distance moyenne			
	moyenne	écart-type	minimum	maximum
1a	0.0108	0.0008	0.0090	0.0135
1b	**0.0077**	0.0002	0.0073	0.0083
2a	0.0140	0.0012	0.0110	0.0171
2b	**0.0079**	0.0002	0.0073	0.0086
3a	0.0163	0.0012	0.0131	0.0202
3b	**0.0083**	0.0002	0.0079	0.0087

Les scénarios 1, 2 et 3 représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées. Les scénarios (a) représentent les simulations sans socialisation et les scénarios (b), avec socialisation.

Ceci n'est pas surprenant puisque plus les individus sont semblables dans le réseau, plus la possibilité de les regrouper, selon leurs intérêts, augmente et donc, plus la distance entre eux devient petite. Néanmoins, la force de structuration des données n'affecte pas dramatiquement les mesures de distance et celles-ci demeurent assez faibles dans tous les scénarios.

Nous savons maintenant d'une part que les réseaux générés par notre modèle sont de densité très faible et que de plus, la distance entre les acteurs est faible (que des chemins de courtes distances existent) et encore plus faible lorsqu'il y a socialisation. À la section suivante, nous analysons la limite supérieure sur la longueur des parcours.

4.6.1.3 Comparaison du nombre de nœuds pivots

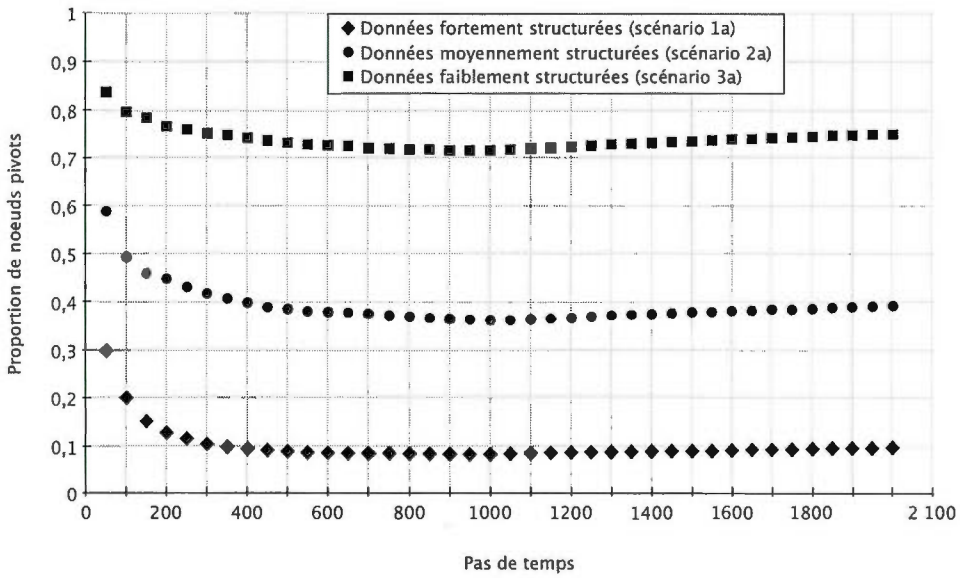
Comme nous l'avons montré au chapitre 3, étant données les propriétés structurales imposées, les sommets pivots donnent accès à tous les autres nœuds (simples) du réseau en un pas seulement. Il s'ensuit que les parcours de connexion et de socialisation peuvent se limiter à ces seuls nœuds pivots. Ainsi, moins il y a de nœuds pivots dans le réseau, moins les parcours de socialisation risquent d'être longs. Nous avons donc mesuré, à tous les intervalles de 50 pas de temps, le nombre de nœuds pivots dans les réseaux simulés.

Comme dans le cas des mesures de distance, nous avons pondéré les valeurs obtenues par la taille du réseau pour calculer la proportion des nœuds pivots par rapport au nombre total de nœuds dans le réseau. Nous avons ensuite calculé la moyenne de ces ratios sur les 100 simulations effectuées pour chaque scénario, à chaque intervalle de temps mesuré. La figure 4.3 (a) illustre l'évolution de cette proportion au cours du temps dans le cas des scénarios sans socialisation (scénarios a) et la figure 4.3 (b) montre le même phénomène, mais dans les scénarios avec socialisation (scénarios b).

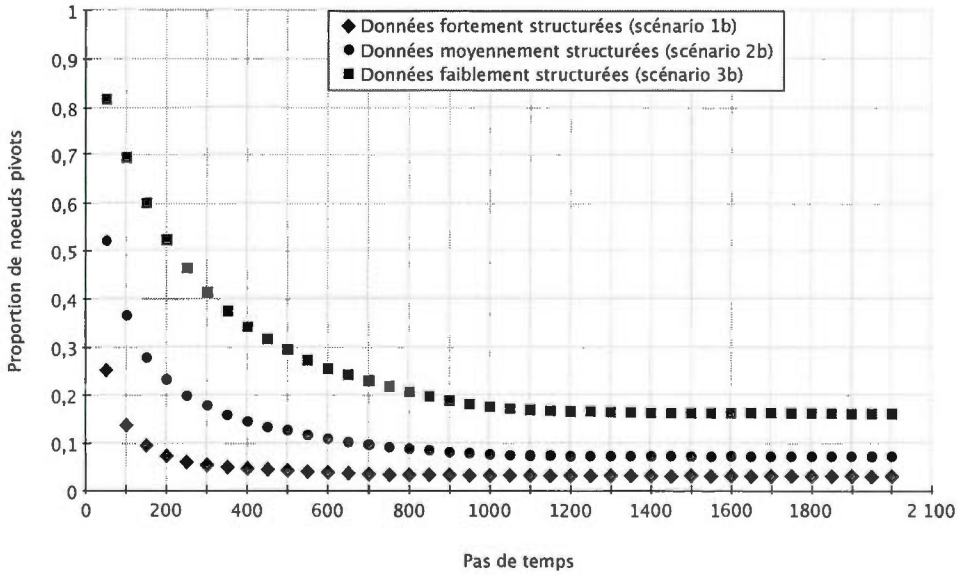
Il apparaît clairement que les mécanismes de socialisation ont pour effet de faire diminuer la proportion des nœuds pivots dans le réseau et cette différence, entre les scénarios avec socialisation et sans socialisation, est d'autant plus significative dans le cas des données peu structurées. En effet, on remarque que dans les scénarios 3 (données faiblement structurées), le scénario 3a (sans socialisation) n'affecte pratiquement pas la proportion des nœuds pivots. La valeur au pas de temps 100 (autour de 0.8) est pratiquement identique au pas de temps 2000 (autour de 0.75). Au contraire, dans les scénarios avec socialisation (scénarios b), même dans le cas de données peu structurées, on observe une diminution notable de la proportion des nœuds pivots qui passe d'une valeur approximative de 0.7 au temps 100 à une valeur autour de 0.17 au temps 2000. Il est remarquable de voir à quel point les mécanismes de socialisation tirent parti de la structure des données, si faible soit-elle.

Nous pensons que ce phénomène s'explique par le fait que les mécanismes de socialisation et la navigabilité se renforcent mutuellement : lorsqu'on socialise, on regroupe les individus entre eux, ce qui tend à diminuer (ou maintenir) la proportion des nœuds pivots dans le réseau. En effet, la fusion de deux communautés contient souvent moins (ou, au pire, le même nombre) de nœuds pivots que la somme des pivots présents dans les deux communautés initiales. Et, plus la proportion des nœuds pivots est petite, plus la navigabilité est efficace, et donc plus les nouveaux nœuds qui arrivent dans le réseau "trouvent" facilement une communauté d'appartenance (si elle existe). Ainsi, on n'augmente pas le nombre de nœuds pivots. En effet, rappelons-nous que lorsqu'un nœud ne trouve pas de communauté d'appartenance, il devient lui-même un nouveau nœud pivot dans le réseau.

Ainsi, la socialisation favorise la navigabilité qui favorise à son tour la socialisation et ainsi de suite.



(a) sans socialisation



(b) avec socialisation

Figure 4.3 Évolution de la proportion des nœuds pivots pondérée en fonction du temps.

Dans cette optique, on remarque aussi, dans les scénarios sans socialisation, que lorsque le réseau arrête de croître (autour du temps 1000), la proportion des nœuds pivots montre une petite tendance à augmenter de nouveau, indiquant clairement que les nœuds qui se connectent ne trouvent pas leur communauté d'appartenance et deviennent de nouveaux nœuds pivots dans le réseau. Au contraire, dans les scénarios avec socialisation, la proportion des nœuds pivots tend clairement à se stabiliser.

Les valeurs moyennes (au pas de temps 2000) de la proportion de nœuds pivots dans les réseaux générés par chacun des scénarios sont indiquées dans le tableau 4.8.

Tableau 4.8
Proportion des nœuds pivots pondérée au pas de temps 2000

Scénario	Proportion de noeuds pivots			
	moyenne	écart-type	minimum	maximum
1a	0.0969	0.0133	0.0600	0.1268
1b	**0.0326**	0.0026	0.0280	0.0400
2a	0.3929	0.0362	0.3026	0.4580
2b	**0.0731**	0.0061	0.0562	0.0862
3a	0.7498	0.0272	0.6660	0.8140
3b	**0.1629**	0.0075	0.1440	0.1864

Les scénarios 1, 2 et 3 représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées. Les scénarios (a) représentent les simulations sans socialisation et les scénarios (b), avec socialisation.

En regardant ce tableau, on peut calculer la différence des valeurs moyennes dans un même scénario, mais avec et sans socialisation. Notons ici que le réseau, au pas de temps 2000, contient approximativement 500 nœuds.

Dans le cas du scénario 1 (données fortement structurées), la différence des valeurs moyennes entre les scénarios a et b est de 0.0643. Dans un réseau de 500 sommets, cette valeur correspond à 32 nœuds pivots de différence. En effet, nous avons remarqué que, dans les simulations du scénario 1 avec socialisation (1b), le nombre de nœuds pivots en début de simulation, au temps 100, par exemple, est à peu près de 7 et augmente doucement pour se stabiliser autour de valeurs proches de 16 en milieu de simulation (autour du temps 1000).

Dans le cas du même scénario, mais sans socialisation (1a), le nombre de pivots débute autour de 10 et augmente continuellement jusqu'à des valeurs tournant autour de 50 en fin de simulation.

Au contraire, dans le cas des données faiblement structurées (scénario 3), la différence des valeurs moyennes entre le scénario a et le scénario b est encore plus grande : 0.5869. Dans un réseau de 500 nœuds, cette valeur représente tout de même 293 nœuds pivots de différence, soit un peu plus de la moitié du nombre total de nœuds dans le réseau. On a effectivement observé que dans les simulations sans socialisation, avec données peu structurées, le nombre de nœuds pivots débute avec une valeur aux environs de 25 et augmente, au cours du temps, jusqu'à des valeurs proches de 375 en fin de simulation. Avec socialisation, le nombre de nœuds pivots initial tourne aussi autour de 25, mais n'augmente que jusqu'à des valeurs proches de 85 qui se stabilisent en milieu de simulation.

Les scénarios avec données moyennement structurées (scénarios 2a et 2b) se situent entre les deux extrêmes avec une différence des valeurs moyennes qui est égale à 0.3198 et qui correspond à environ 160 nœuds pivots dans un réseau de 500 nœuds. Sans socialisation, on observe, au cours de la simulation, une variation du nombre de nœuds pivots passant d'environ 15 pivots en début de simulation à environ 196 pivots en fin de simulation. Avec socialisation, on passe d'environ 16 pivots à environ 36 pivots en fin de simulation.

En résumé, on constate que les mécanismes de socialisation favorisent effectivement le maintien d'une proportion faible de nœuds pivots dans le réseau, et ce, même avec des données peu structurées. Les deux mesures suivantes ne se concentrent pas sur les qualités structurales du réseau, mais plutôt, sur l'efficacité des mécanismes de socialisation à regrouper les individus de profils similaires c'est-à-dire, à créer des communautés d'intérêt.

4.6.1.4 Analyse de l'homophilie

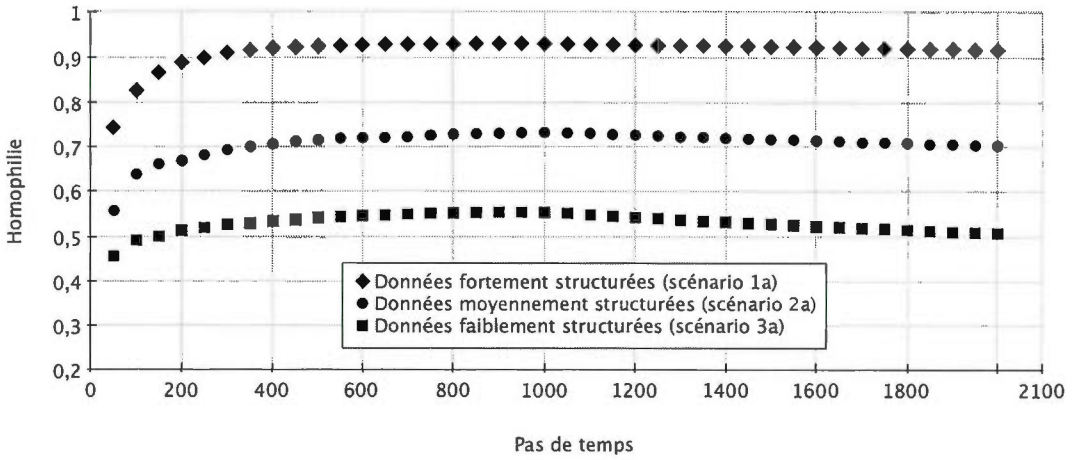
Nous avons calculé l'homophilie sur les réseaux générés par notre modèle de socialisation à tous les intervalles de 50 pas de temps. Nous avons ensuite calculé la moyenne de ces valeurs sur les 100 simulations effectuées pour chaque scénario. Nous aimerions rappeler, ici, que la

similarité entre les individus est calculée selon la valeur de la limite d'équivalence (qui a été préalablement fixée dans les paramètres du scénario). Lorsque la similarité entre deux acteurs est égale ou supérieure à cette limite d'équivalence, on considère les deux acteurs identiques et leur valeur de similarité devient donc égale à 1. Ceci pour dire que la valeur maximale de l'homophilie dans les réseaux générés est égale à 1 (si tous les acteurs dans le réseau ne formaient qu'une seule communauté d'intérêt).

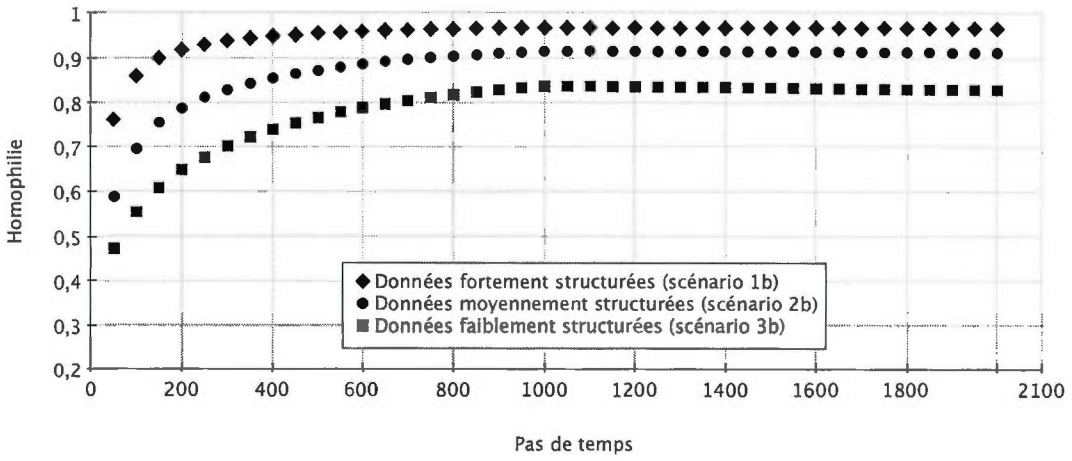
La figure 4.4 illustre l'évolution de l'homophilie moyenne au cours du temps, pour tous les scénarios. Dans tous les cas, on remarque l'atteinte rapide d'un plateau qui demeure assez stable au cours du temps, sauf dans le cas des scénarios sans socialisation (fig. 4.4 (a)) où l'on observe une petite tendance à la baisse lorsque le réseau cesse de croître (autour du pas de temps 1000).

Cette tendance négative peut s'expliquer par le fait que sans socialisation, on ne renforce pas la navigabilité (comme expliqué au paragraphe 4.6.1.3) et que lorsque des nœuds qui appartenaient à une communauté se déconnectent et se reconnectent ensuite au réseau, ils ne trouvent plus nécessairement leur ancienne communauté d'appartenance et doivent se connecter à des nœuds étrangers, ce qui a pour effet de diminuer l'homophilie dans le réseau. Évidemment, cette tendance est d'autant plus forte lorsque les données sont faiblement structurées puisque les réseaux possèdent beaucoup plus de nœuds pivots et rendent ainsi la navigation plus difficile.

Le tableau 4.9 contient les valeurs moyennes de l'homophilie au pas de temps 2000, pour chaque scénario, valeurs qui correspondent (à peu près) aux plateaux atteints lors de l'évolution des réseaux.



(a) sans socialisation



(b) avec socialisation

Figure 4.4 Évolution de l'homophilie en fonction du temps.

Premièrement, que ce soit pour les données structurées ou non, les scénarios avec socialisation montrent une homophilie plus élevée que dans le cas des scénarios sans socialisation. On remarque cependant que dans le cas des données fortement structurées (scénarios 1), cette différence est moins grande que pour les données moyennement et faiblement structurées. Ceci s'explique par le fait que la forte structuration des données, qui implique que les individus se ressemblent beaucoup, favorise naturellement l'homophilie.

Tableau 4.9
Homophilie au pas de temps 2000

Scénario	Homophilie			
	moyenne	écart-type	minimum	maximum
1a	0.9171	0.0129	0.8907	0.9516
1b	**0.9674**	0.0044	0.9566	0.9756
2a	0.7038	0.0290	0.6493	0.7780
2b	**0.9131**	0.0083	0.8913	0.9321
3a	0.5086	0.0224	0.4581	0.5760
3b	**0.8286**	0.0097	0.8042	0.8615

Les scénarios 1, 2 et 3 représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées. Les scénarios (a) représentent les simulations sans socialisation et les scénarios (b), avec socialisation.

Évidemment, nos réseaux ne peuvent pas atteindre une homophilie parfaite étant donné que nos acteurs ne sont pas tous identiques dans aucun des scénarios. Cependant, dans le cas des données fortement structurées avec socialisation (scénario 1b), on atteint une valeur moyenne tout près de 0.97. Cela indique, effectivement, la présence d'un tout petit nombre de communautés très fournies qui contiennent tous les liens de similarité égale à 1 et seulement un petit nombre de liens dont la similarité est en dessous de la limite d'équivalence et qui servent à relier les communautés entre elles. Effectivement, plus le nombre de communautés est petit, moins on a besoin de liens extra-communautés : pour relier trois communautés, il faut au moins deux liens extra-communautés tandis que pour en relier 10, il en faut au moins 9. Dans ce cas, où les données sont fortement structurées, le nombre de liens de similarité égale à 1 (liens intra-communautés) est donc très élevé par rapport au nombre de liens extra-communautés.

Au contraire, dans le cas des données faiblement structurées, avec socialisation, on atteint un plateau plus petit autour de 0.8286. En effet, selon la nature de ces données, on obtient forcément plusieurs petites communautés qui requièrent plus de liens extra-communautés pour les relier entre elles. La proportion maximum possible des liens de similarité égale à 1 (intra-communautés) est donc nécessairement plus petite. Cependant, on remarque encore une fois à quel point les mécanismes de socialisation peuvent profiter de la structure des données même lorsque celle-ci est faible. En effet, le scénario 3a (sans socialisation) produit

des réseaux dont la valeur moyenne d'homophilie est seulement de 0.5086 comparativement à son pendant avec socialisation (scénario 3b) qui génère des réseaux dont la valeur d'homophilie est beaucoup plus élevée, soit 1.6 fois plus élevée, avec une valeur de 0.8286. Les scénarios avec données moyennement structurées se situent entre ces deux extrêmes, avec une valeur moyenne d'homophilie, au pas de temps 2000, de 0.7038 sans socialisation, et de 0.9131, avec socialisation.

Les mécanismes de socialisation favorisent donc effectivement le regroupement des individus similaires en communautés d'intérêt. Voyons maintenant à quel point la formation de ces communautés d'intérêt favorise les rencontres des acteurs identiques au sein du réseau.

4.6.1.5 Examen des rencontres effectuées

Avant d'effectuer nos simulations, nous avons déterminé, pour chaque scénario, selon la valeur de la limite d'équivalence fixée pour ce scénario, tous les acteurs identiques dans notre collection d'acteurs. Plus précisément, pour chaque acteur, nous avons dressé une liste de tous les autres acteurs qu'il aurait intérêt à rencontrer. Puis, lors des simulations, à chaque intervalle de 50 pas de temps, nous avons mesuré, pour chaque acteur présent dans le réseau, la proportion donnée par le nombre d'acteurs qu'il a déjà rencontrés sur le nombre d'acteurs (présents dans le réseau au moment de la mesure) qu'il aurait intérêt à rencontrer (calculer avant la simulation). Nous avons ensuite évalué la moyenne de ces scores de rencontres individuels pour obtenir un score au niveau du réseau, à tous les intervalles mesurés. La figure 4.5 montre l'évolution du score moyen des rencontres effectuées au cours du temps, pour chaque scénario.

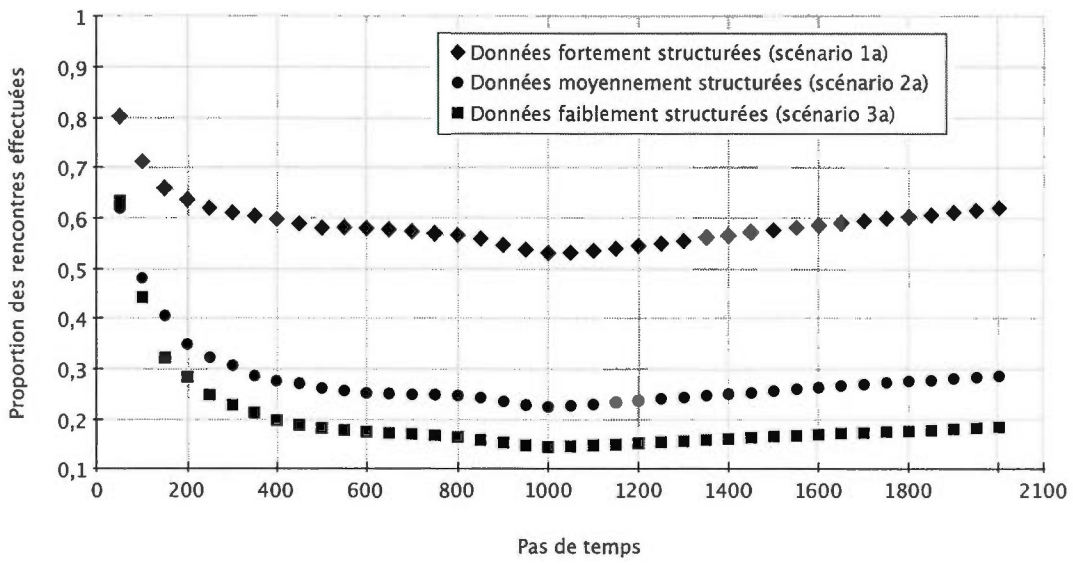
On remarque, tout d'abord, que la formation de communautés d'intérêt, par socialisation, favorise grandement les rencontres. Dans le cas des scénarios sans socialisation (fig. 4.5 (a)), on remarque, encore une fois, que le scénario avec données fortement structurées, qui induit naturellement les regroupements par intérêt, produit (et de beaucoup) de meilleurs résultats que les scénarios avec données moins structurées.

Dans tous les cas sans socialisation, on observe aussi une diminution initiale de la proportion des rencontres effectuées à mesure que la taille du réseau augmente jusqu'au pas de temps 1000, lorsque la taille du réseau cesse de croître. Cette tendance initiale indique l'inefficacité des réseaux sans socialisation à provoquer des rencontres à mesure que de nouveaux acteurs arrivent constamment dans le réseau. Cette tendance s'inverse à la hausse, cependant, lorsque la taille du réseau se stabilise (autour du pas de temps 1000). En effet, dans un réseau de taille constante, où les mêmes acteurs ne font que se déconnecter et se reconnecter, la proportion des rencontres effectuées ne peut plus diminuer.

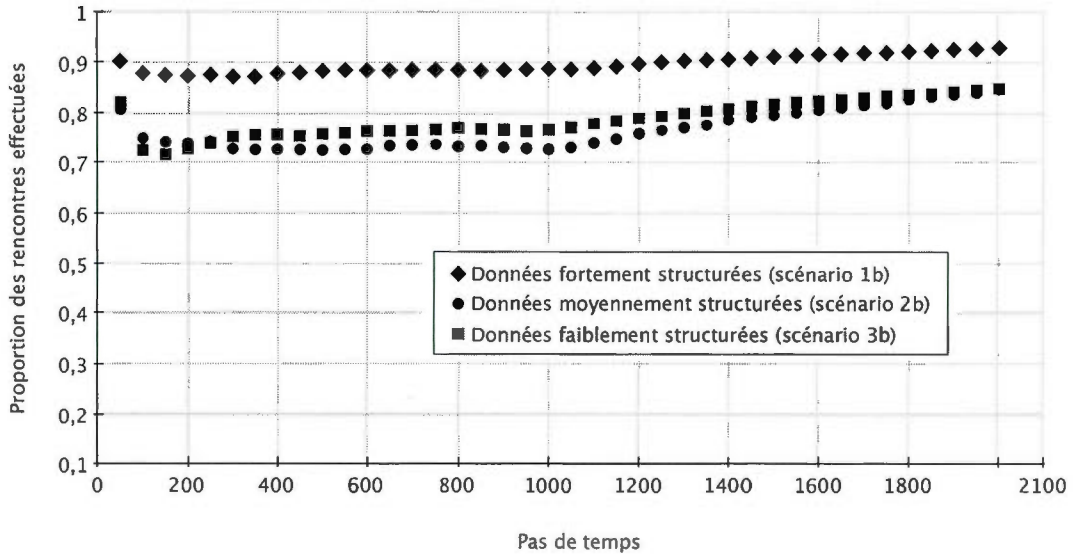
Cette tendance à la hausse, cependant, indique que les nouvelles connexions (ou reconnexions) des acteurs permettent, à un certain degré, de faire de nouvelles rencontres, et ce, seulement par l'intermédiaire des parcours de connexion. De plus, ce phénomène est un peu plus fort dans les réseaux fortement structurés : on remarque en effet que la pente positive (après le temps 1000) augmente légèrement avec le niveau de structuration des données.

Dans le cas des scénarios avec socialisation (fig. 4.5 (b)), on n'observe presque pas de tendance à la baisse en début de simulation. On peut donc penser que les mécanismes de socialisation provoquent rapidement assez de rencontres pour pallier l'augmentation continuelle de la taille du réseau, jusqu'au temps 1000.

À partir du temps 1000, on note aussi, comme dans les cas sans socialisation, une légère augmentation de la pente proportionnelle au niveau de structuration des données (pour les données faiblement et moyennement structurées). Cet effet est à peine perceptible, cependant, dans le cas du scénario avec données fortement structurées, où la valeur de la pente n'augmente que très faiblement. Cette dernière observation semble indiquer qu'avec des données fortement structurées, le taux d'évolution (la pente) de la proportion des rencontres effectuées atteint presque son maximum dès le début des simulations.



(a) sans socialisation



(b) avec socialisation

Figure 4.5 Évolution de la proportion des rencontres effectuées en fonction du temps.

Le tableau 4.10 présente les valeurs moyennes de la proportion des rencontres effectuées, en fin de simulation. Lorsqu'on regarde les valeurs moyennes de la proportion des rencontres effectuées au temps 2000, on voit clairement que les mécanismes de socialisation favorisent, et de beaucoup, le nombre de rencontres effectuées au cours de l'évolution du réseau.

Tableau 4.10
Proportion des rencontres effectuées au pas de temps 2000

Scénario	Proportion des rencontres effectuées			
	moyenne	écart-type	minimum	maximum
1a	0.6196	0.0720	0.4874	0.7578
1b	**0.9302**	0.0408	0.8346	0.9920
2a	0.2865	0.0632	0.1980	0.4992
2b	**0.8480**	0.0562	0.7053	0.9461
3a	0.1851	0.0307	0.1311	0.2695
3b	**0.8491**	0.0283	0.7761	0.9121

Les scénarios 1, 2 et 3 représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées. Les scénarios (a) représentent les simulations sans socialisation et les scénarios (b), avec socialisation.

Dans le cas des données moyennement structurées (scénario 2), le scénario avec socialisation (2b) montre une valeur moyenne de presque trois fois supérieure à son pendant sans socialisation (2a). Dans le cas des données faiblement structurées (scénarios 3a et 3b), on obtient une valeur 4.5 fois plus élevée dans le scénario avec socialisation. Pour ce qui est des scénarios avec données fortement structurées (scénario 1a et 1b), la différence est un peu moindre avec une valeur moyenne 1.5 fois plus grande pour le scénario avec socialisation.

Le score moyen observé dans le cas des données fortement structurées avec socialisation (scénario 1b) est nettement supérieur aux autres scores des scénarios avec socialisation. Curieusement, on remarque aussi qu'au pas de temps 2000, la proportion moyenne de rencontres effectuées dans le scénario 2b (données moyennement structurées avec socialisation) est pratiquement égale à celle du scénario 3b (données faiblement structurées avec socialisation). En effet, à ce pas de temps, les pentes se croisent, mais la pente légèrement plus élevée dans le scénario 2b laisse supposer que, quelques pas de temps plus tard, la valeur du scénario 2b dépassera la valeur du scénario 3b (voir fig. 4.5 (b)).

Les mécanismes de socialisation favorisent donc le nombre de rencontres effectuées au cours de l'évolution des réseaux. Nous passons maintenant à la validation de notre modèle de socialisation en tant que modèle d'évolution des réseaux sociaux.

4.6.2 Validation au niveau sociocognitif

Dans cette section, nous comparons tout d'abord les propriétés structurales observées dans les réseaux sociaux du monde réel avec celles calculées dans les réseaux générés par notre modèle de socialisation. Nous voulons ainsi déterminer dans quelle mesure les mécanismes de socialisation que nous proposons peuvent expliquer l'évolution des réseaux sociaux réels. Pour ce faire, nous ne considérons que les scénarios avec socialisation, et nous analysons la structure des réseaux générés en fin de simulation (au pas de temps 2000).

Ensuite, pour déterminer si la socialisation favorise la création de capital social, nous comparons les mesures de centralité de proximité dans les scénarios avec et sans socialisation.

4.6.2.1 Effet des petits mondes

Au chapitre 2, nous avons vu qu'une des caractéristiques largement observées dans la plupart des réseaux sociaux est l'effet des petits mondes. Comme nous l'avons déjà montré (voir par. 4.6.1.2), la distance moyenne et le diamètre dans les réseaux générés par notre modèle sont très petits. On observe donc effectivement l'effet des petits mondes dans nos réseaux.

De plus, dans la littérature, plusieurs études ont montré que la distance moyenne des réseaux sociaux réels s'apparente à celle qu'on obtiendrait dans un graphe aléatoire de même dimension (même nombre de nœuds et de liens) (Watts et Strogatz, 1998 ; Newman 2001b ; Barabási et al., 2002).

Nous avons évalué la distance moyenne d'un graphe aléatoire équivalent pour chacun des scénarios et nous avons comparé cette distance avec celle qu'on observe dans nos réseaux. La distance moyenne dans un graphe aléatoire peut se calculer comme suit (Albert et Barabási, 2002) :

$$D_{aléa} = \frac{\ln N}{\ln k},$$

où N est le nombre de nœuds dans le réseau et $k = \frac{2 \times L}{N}$ est le degré moyen (L étant le nombre de liens (bidirectionnels) dans le réseau).

Soit N, le nombre moyen de nœuds dans nos réseaux, L, le nombre moyen de liens dans nos réseaux et D, la valeur moyenne des distances moyennes dans nos réseaux (les moyennes étant calculées sur les 100 simulations effectuées pour chaque scénario). Soit $D_{aléa}$, la distance moyenne calculée pour un graphe aléatoire de N nœuds et L liens. Le tableau 4.11 montre les valeurs de N, L, D, k et $D_{aléa}$ pour chacun des scénarios. On voit que D est beaucoup plus petit que $D_{aléa}$, peu importe le niveau structuration des données : l'effet des petits mondes est plus fort que ce qu'on observe en général dans les réseaux sociaux réels.

Tableau 4.11
Comparaison de la distance moyenne avec celle d'un graphe aléatoire de même taille

Scénario	N	L	k	D	$D_{aléa}$
1b	500	509	2.036	3.8607	8.7408
2b	500	531	2.124	3.9246	8.2498
3b	500	592	2.368	4.1406	7.2091

Les scénarios 1b, 2b et 3b représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées, avec socialisation.

Ceci n'est pas surprenant étant donné que les caractéristiques structurales que nous avons imposées aux réseaux générés par notre modèle, dans une optique de performance, ont comme objectif de produire des réseaux ayant une distance moyenne la plus courte possible. Pour une modélisation sociocognitive plus fine, il faudrait alors relaxer quelque peu ces contraintes, nécessaires, cependant, au point de vue fonctionnel. On pourrait, par exemple, ne pas imposer que tous les pivots précédents, sur une chaîne de pivots, soient toujours pleins. Cela aurait pour effet de rallonger un peu les chaînes de pivots et donc, du même coup, la distance moyenne entre les nœuds du réseau.

Dans la section suivante, nous vérifions la présence de communautés structurales observées dans les réseaux sociaux réels.

4.6.2.2 Émergence de communautés structurales

Nous avons choisi aléatoirement 10 réseaux par scénario étudié sur lesquels nous avons évalué la moyenne de la mesure de modularité Q. Les valeurs obtenues sont présentées au tableau 4.12. On remarque, en effet, des valeurs de Q élevées qui indiquent une forte présence de communautés structurales dans nos réseaux : le partitionnement de nos réseaux en communautés d'intérêt se reflète très fortement au niveau structural. En général, une valeur de Q supérieure à 0.3 indique un partitionnement significatif. De plus, dans les trois scénarios étudiés, ces valeurs sont comparables aux valeurs mesurées sur divers réseaux sociaux réels présentés dans le tableau 2.2 (par. 2.3.2.4).

Tableau 4.12
Modularité au temps 2000 pour les trois scénarios avec socialisation

Scénario	**Modularité (Q) moyenne**			
	moyenne	écart-type	minimum	maximum
1b	0.7402	0.0488	0.6481	0.7944
2b	0.7826	0,0132	0.7524	0.8021
3b	0.6716	0.0069	0.6629	0.6866

Les scénarios 1b, 2b et 3b représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées, avec socialisation.

On remarque aussi que la modularité semble plus faible lorsque les données sont peu structurées. Effectivement, dans ce cas, on a un plus grand nombre de petites communautés donc la proportion des liens extra-communautés est plus grande, ce qui a pour effet de faire diminuer la valeur de modularité.

4.6.2.3 Transitivité

Dans le monde réel, nos réseaux sociaux montrent typiquement une valeur de *clustering* qui, bien que variant d'un réseau à l'autre, demeure très supérieure au coefficient de *clustering* qu'on obtiendrait pour un réseau aléatoire de même taille. On peut se référer au tableau 2.1 (par. 2.3.2.2) pour voir une comparaison du coefficient de *clustering* entre divers réseaux sociaux réels et leur homologue aléatoire.

Le coefficient de *clustering* d'un graphe aléatoire est :

$$C_{aléa} = \frac{k}{N},$$

où $k = \frac{2 \times L}{N}$ est le degré moyen, L est le nombre de liens et N, le nombre de nœuds.

Nous avons donc calculé la moyenne du coefficient de *clustering* C sur nos 100 réseaux, pour chaque scénario avec socialisation et nous avons aussi calculé le coefficient de *clustering* $C_{aléa}$ pour les graphes aléatoires équivalents. Le tableau 4.13 présente les valeurs obtenues.

Tableau 4.13
Comparaison du coefficient de *clustering* avec celui d'un graphe aléatoire équivalent

Scénario	N	L	k	C	$C_{aléa}$
1b	500	509	2.036	0.0000	0.0041
2b	500	531	2.124	0.0010	0.0042
3b	500	592	2.368	0.0039	0.0047

Les scénarios 1b, 2b et 3b représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées, avec socialisation.

Dans tous les scénarios, le coefficient de *clustering* de nos réseaux est plus faible que celui de leur graphe aléatoire correspondant. Cette différence est cependant moins marquée dans le scénario 3b (données faiblement structurées) dont le coefficient de *clustering* approche celui de son pendant aléatoire. Notre modèle ne produit donc pas des réseaux qui exhibent la propriété de transitivité qu'on observe dans les réseaux sociaux du monde réel. Cet état de fait n'est pas inattendu puisque nous avons consciemment privilégié les propriétés fonctionnelles de notre modèle au détriment, dans ce cas précis, d'une propriété désirable au niveau de la modélisation sociocognitive.

Dans nos réseaux, on remarque que le phénomène de *clustering*, si léger soit-il, ne se produit qu'au niveau des nœuds pivots. La figure 4.6 illustre trois exemplaires du graphe des nœuds pivots pour chacun des scénarios. En accord avec les valeurs du coefficient de *clustering*

indiquées dans le tableau 4.13, on note que le graphe du scénario 3b (données faiblement structurées) possède en effet plus de nœuds pivots et plus de triangles dans sa structure que le graphe du scénario 1b (données fortement structurées) qui contient très peu de nœuds pivots et ne produit presque pas de triangles.

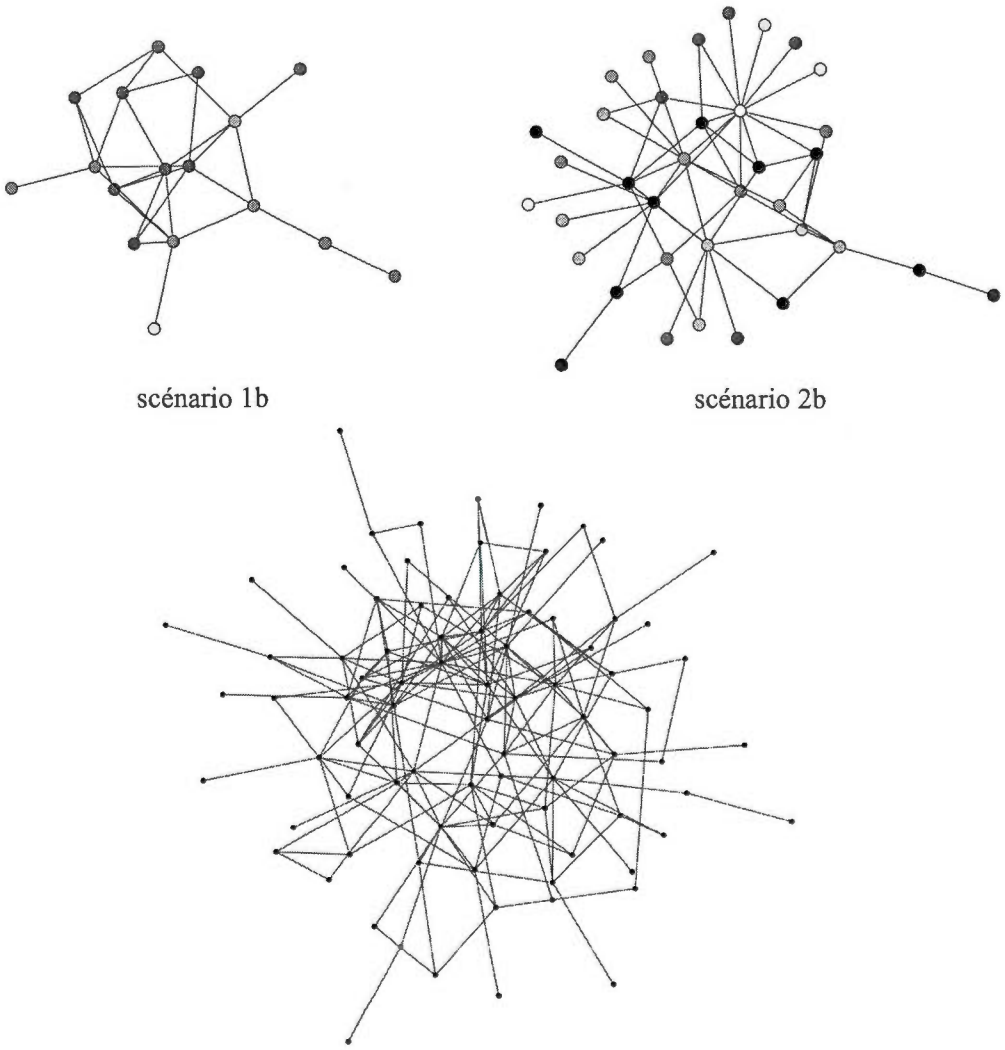


Figure 4.6 Réseaux des noeuds pivots en fin de simulation pour les trois scénarios avec socialisation. Les scénarios 1b, 2b et 3b représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées.

Cependant, dans une perspective de modélisation sociocognitive, on pourrait facilement ajouter un certain degré de *clustering* dans les réseaux générés par notre modèle, au niveau des nœuds simples. On pourrait, par exemple, imaginer un mécanisme de fermeture des triangles, similaire à celui présenté dans le modèle de Holme et Kim (art. 2.4.4). Ainsi, chaque fois qu'un nœud simple *j* se connecterait à un nœud pivot *i*, il se connecterait aussi, selon une probabilité *p*, à un autre voisin immédiat (nœud simple) de *i*. Ce mécanisme supplémentaire augmenterait la densité du réseau (non désirable d'un point de vue fonctionnel), mais produirait un certain degré de *clustering*, contrôlé par la probabilité *p*. Voyons maintenant si nos réseaux montrent une distribution des degrés qui suit une loi de puissance.

4.6.2.4 Distribution des degrés suivant une loi de puissance

Pour vérifier si la distribution des degrés de nos réseaux suit une loi de puissance, nous avons estimé un modèle de régression linéaire par la méthode des moindres carrés dans la représentation cartésienne log-log de la distribution des degrés de nos réseaux. Le tableau 4.14 montre la moyenne (sur les 100 réseaux générés par chacun des scénarios) des coefficients de corrélation associés aux droites estimées.

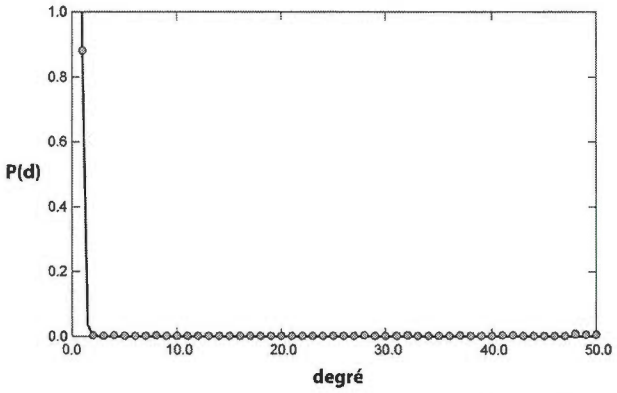
Tableau 4.14
Coefficient de corrélation des droites estimées dans les graphes log-log de la distribution des degrés de nos réseaux, pour les scénarios avec socialisation

Scénario	Coefficient de corrélation			
	moyenne	écart-type	minimum	maximum
1b	0.6570	0.1010	0.4348	0.9234
2b	0.6251	0,0550	0.4779	0.7970
3b	0.7811	0.0438	0.5972	0.8766

Les scénarios 1b, 2b et 3b représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées.

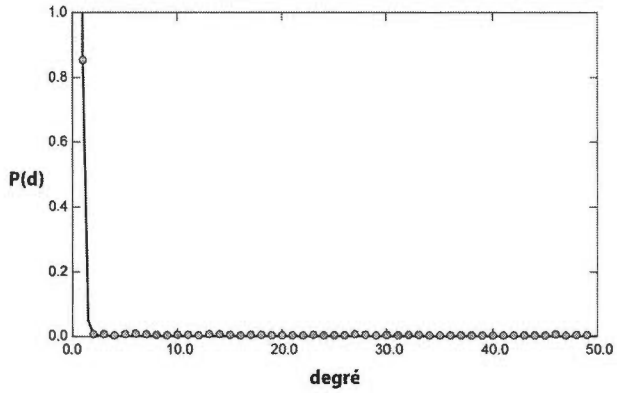
On remarque une corrélation assez forte, au-dessus de 0.6, dans tous les scénarios. La distribution des degrés de nos réseaux tend donc à suivre une loi de puissance, comme on observe dans plusieurs réseaux sociaux réels. Pour mieux visualiser les lois de puissance sur la distribution des degrés de nos réseaux, nous avons estimé une loi de puissance sur la

distribution des degrés d'un réseau choisi au hasard parmi les 100 simulations, pour chacun des scénarios, à l'aide du logiciel pour iPad *DataAnalysis* de la compagnie *Data Evaluation Systems*. La figure 4.7 illustre les trois courbes estimées.



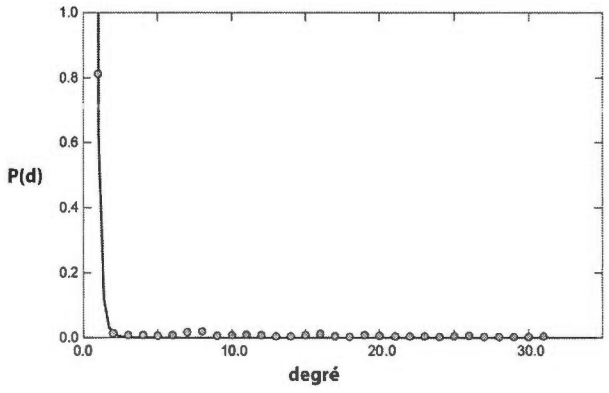
$P(d) = 0.8818d^{-7.8806}$

scénario 1b



$P(d) = 0.8725d^{-7.15}$

scénario 2b



$P(d) = 0.8120d^{-5.8042}$

scénario 3b

Figure 4.7 Estimation d'une loi de puissance sur la distribution des degrés d'un réseau pour chaque simulation avec socialisation.

Premièrement, on observe que les pentes initiales sont extrêmement abruptes. Ceci s'explique par le fait que dans nos réseaux, la majorité des nœuds, les nœuds simples, ont tous un degré égal à 1 et que seule la petite proportion des nœuds pivots possède des degrés variés pouvant aller jusqu'à 50 (comme on l'a fixé dans les paramètres communs des scénarios). Donc, ou bien l'on a une très forte probabilité d'un degré égal à 1 (pour les nœuds simples), ou bien l'on a une très faible probabilité pour tous les autres degrés (pour les nœuds pivots), et aucune possibilité entre les deux. Dans cette optique, on note aussi que la pente initiale dans le scénario 3b (données faiblement structurées) est un peu plus douce que dans les autres scénarios.

En effet dans le cas de données faiblement structurées, on a beaucoup plus de nœuds pivots que dans les autres cas (voir fig. 4.3). Ainsi, la probabilité d'un degré différent de 1 est un peu plus forte que dans les autres scénarios (la répartition des degrés est un peu moins radicale), mais à peine.

Les pentes à ce point abruptes sont dues au coefficient de loi de puissance α très élevé : 7.88 pour le scénario 1b, 7.15 pour le scénario 2b, et 5.80 pour le scénario 3b. Ces coefficients sont en effet plus grands que ce qu'on observe en général, dans les réseaux sociaux réels, qui se situent, le plus souvent entre 2 et 3. Par exemple, le tableau 4.15 présente les coefficients de loi de puissance pour divers réseaux sociaux réels.

Tableau 4.15
Coefficients de loi de puissance observés dans divers réseaux sociaux réels

Réseau	α	Référence
Réseaux d'acteurs de films	2.3	Barabási et Albert, 1999
Réseaux de co-auteurs sur SPIRES	1.2	Newman, 2001b
Réseau de co-auteurs en mathématique	2.5	Barabási et al., 2002
Réseau de co-auteurs en neurosciences	2.1	Barabási et al., 2002
Réseau de contacts sexuels	3.4	Liljeros et al., 2001

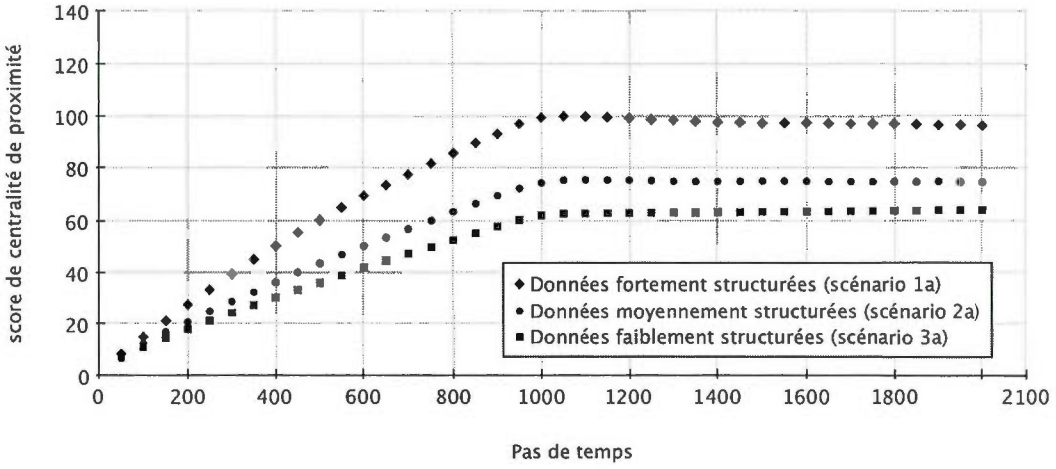
Pour modéliser plus finement le genre de loi de puissance qu'on observe typiquement dans les réseaux sociaux réels, on pourrait contrôler le coefficient de loi de puissance, au cours de l'évolution du réseau, pour qu'il se rapproche des valeurs observées dans le monde réel. En fait, si l'on intégrait un mécanisme de fermeture des triangles discuté au paragraphe 4.6.2.3, on pourrait le faire en préférant légèrement, selon une probabilité q, l'attachement d'un nœud simple à un autre noeud simple de degré élevé (attachement préférentiel). Ceci aurait probablement pour effet d'atténuer la forte population de nœuds de degré 1 en permettant de contrôler la distribution des degrés sur les nœuds simples à l'aide de la probabilité q. Ceci n'est qu'une hypothèse, cependant, qui n'est pas traitée dans le cadre de cette thèse.

À la section suivante, nous examinons la centralité de proximité au sein des réseaux générés, en tant que mesure du capital social. Nous voulons déterminer si nos mécanismes de socialisation favorisent ce type de capital social.

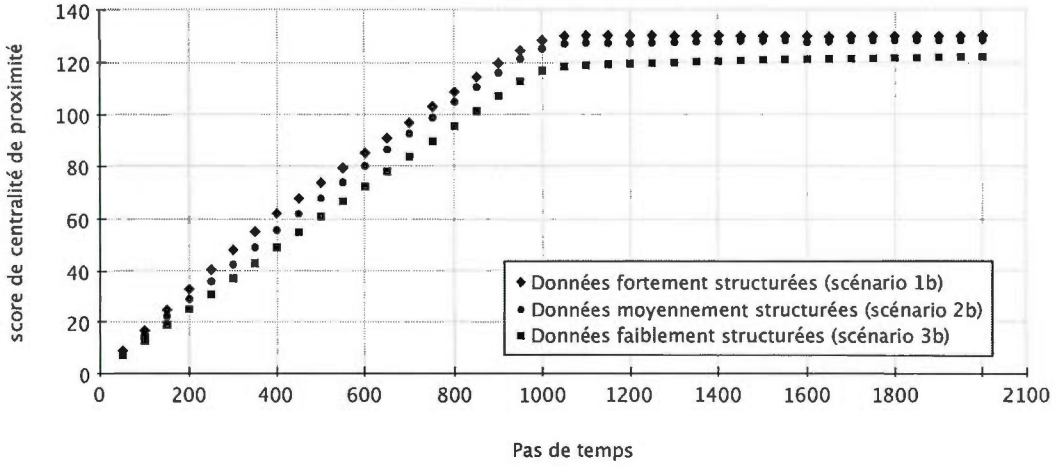
4.6.2.5 Mesure de la centralité de proximité

Tout d'abord, à tous les intervalles de 50 pas de temps, nous avons évalué la moyenne des scores de centralité de proximité normalisés entre 0 et 1 (voir par. 2.3.1.7) de tous les acteurs présents dans le réseau au moment de la mesure. Ensuite, puisqu'une proximité x, dans un réseau de 500 acteurs, est beaucoup plus significative qu'une proximité x dans un réseau de 10 acteurs, nous avons pondéré les valeurs obtenues par la taille du réseau en divisant le score moyen pour chaque pas de temps, par l'inverse du nombre d'acteurs dans le réseau. La figure 4.8 illustre l'évolution des scores moyens de centralité de proximité (pondérés) en fonction du temps.

On note que dans tous les scénarios, la centralité pondérée augmente au cours de la période de croissance du réseau (jusqu'au pas de temps 1000) pour atteindre un plateau qui se stabilise ensuite jusqu'à la fin de la simulation. La différence majeure entre les scénarios avec socialisation (fig. 4.8 (b)) et ceux sans socialisation (fig. 4.8 (a)) est qu'au cours de la période de croissance du réseau, le taux d'augmentation de la centralité est plus élevé (la pente est plus abrupte) dans les scénarios avec socialisation, ce qui produit une valeur de stabilisation plus grande que dans les scénarios sans socialisation.



(a) sans socialisation



(b) avec socialisation

Figure 4.8 Évolution de la centralité de proximité en fonction du temps.

Le tableau 4.16 présente les valeurs calculées au pas de temps 2000. On note que dans les scénarios avec socialisation, la différence des plateaux atteints selon que les données sont plus ou moins structurées est moins prononcée que dans les scénarios sans socialisation. Ceci s'explique encore par le phénomène de renforcement mutuel entre la socialisation et la navigabilité : la socialisation favorise la navigabilité qui favorise à son tour la socialisation. Ainsi, les mécanismes de socialisation sont capables de tirer vraiment parti de la structure des données, si faible soit-elle, et réussissent mieux à former des communautés (à rapprocher les acteurs les uns des autres), même lorsque les données sont peu structurées.

Tableau 4.16
Centralité de proximité pondérée au pas de temps 2000

Scénario	Centralité de proximité			
	moyenne	écart-type	minimum	maximum
1a	96.3065	7.2163	76.9852	114.1711
1b	**130.5078**	**3.6324**	121.1227	138.6701
2a	74.5960	6.3284	61.0882	93.6229
2b	**128.8465**	**3.2015**	118.1779	137.8232
3a	63.8803	4.7388	51.3121	79.7450
3b	**122.3929**	**2.6813**	117.1149	128.6446

Les scénarios 1, 2 et 3 représentent respectivement les scénarios avec données fortement, moyennement et faiblement structurées. Les scénarios (a) représentent les simulations sans socialisation et les scénarios (b), avec socialisation.

Nous pouvons donc dire que les mécanismes de socialisation favorisent la formation de capital social, en terme de centralité de proximité.

Notons qu'en étudiant la centralité de proximité moyenne, on considère celle-ci au niveau du réseau plutôt qu'au niveau des acteurs. Si l'on regarde, cependant, l'écart-type associé à chacune des moyennes calculées dans le tableau 4.16, on remarque que ceux-ci sont à peu près 2 fois moins grands dans les scénarios avec socialisation (1b, 2b, 3b) que dans les scénarios sans socialisation (1a, 2a, 3a). Ceci indique que la répartition des scores de centralités dans les réseaux avec socialisation se concentre plus fortement autour de la moyenne : les acteurs ont une plus forte tendance à occuper des positions équivalentes (en terme de centralité de proximité) dans les réseaux avec socialisation. En ce sens, au niveau du capital social individuel, la plupart des acteurs sont égaux et ne possèdent donc pas d'avantages notables les uns sur les autres. L'avantage se situe au niveau du réseau global : tous les acteurs, en général, sont proches les uns des autres, ils peuvent donc communiquer facilement les uns avec les autres et être rapidement au fait de l'information qui circule dans le réseau. Si le réseau représentait celui des employés d'une entreprise, par exemple, la proximité de tous les acteurs constitue un avantage pour cette entreprise. Par contre, lorsque seulement un petit nombre d'acteurs se trouve en position de pouvoir, en terme de centralité de proximité, ceux-ci peuvent alors mieux contrôler, et donc bloquer ou filtrer le flux

d'informations dans le réseau, ce qui n'est pas désirable pour le bon fonctionnement de l'entreprise (au niveau du réseau global).

Donc, non seulement les mécanismes de socialisation avantagent la centralité de proximité au niveau du réseau, ils favorisent aussi l'égalité entre les acteurs.

4.7 Conclusion

D'un point de vue informatique, nous avons vérifié notre hypothèse qui stipulait que les mécanismes de socialisation favorisaient les rencontres ainsi que la formation de communautés d'intérêts, en rapprochant les individus de profils similaires. En effet, nous avons montré que l'homophilie globale et la proportion des rencontres effectuées sont nettement supérieures dans les scénarios avec socialisation.

Nous avons aussi montré que les propriétés structurales imposées produisent bel et bien des structures qui avantagent la navigabilité. En effet, nous avons constaté que la densité des réseaux générés par notre modèle demeure très faible au cours de l'évolution des réseaux (près du minimum pour conserver un réseau connexe). De plus, nous avons montré que la distance moyenne entre les acteurs ainsi que le diamètre de nos réseaux demeurent très modestes par rapport à la taille du réseau et qu'ils diminuent encore plus lorsqu'il y a socialisation. Enfin, nous avons observé une différence remarquable entre le nombre de nœuds pivots présents dans les réseaux générés par les scénarios avec et sans socialisation : les scénarios avec socialisation favorisent nettement la diminution du nombre de pivots. Comme nos parcours de connexion/socialisation ne s'effectuent que sur les nœuds pivots, moins il y en a, plus nos parcours risquent d'être courts. En somme, les propriétés structurales que nous avons choisi d'imposer pour la formation de nos réseaux, combinées avec nos mécanismes de socialisation, favorisent véritablement la navigabilité au sein des réseaux générés.

À plusieurs reprises, on a observé une dynamique de renforcement entre la navigabilité et la socialisation : la navigabilité favorise les rencontres (et recommandations) qui ont pour effet de rapprocher les individus entre eux et ce rapprochement favorise à son tour la navigabilité.

Au niveau de la modélisation sociocognitive, nous avons vérifié notre hypothèse principale qui stipulait que les mécanismes de socialisation pouvaient expliquer, dans une certaine mesure, la présence de propriétés structurales typiques qu'on observe dans les réseaux sociaux réels. Nous avons montré que nos réseaux possèdent plusieurs de ces caractéristiques, dont l'effet des petits mondes, l'émergence de communautés structurales significatives et une distribution des degrés suivant une loi de puissance. Comme nous l'avons mentionné, nous n'avons pas pu démontrer la présence de *clustering* dans nos réseaux, puisque nous avons choisi de réduire au possible ce phénomène, pour des considérations d'ordre fonctionnel. Bien que le modèle proposé ne soit pas une représentation fine de l'évolution des réseaux sociaux parce que nous avons privilégié l'utilisabilité informatique sur la modélisation sociocognitive, nous pensons toutefois qu'il est raisonnablement vraisemblable.

Finalement, nous avions fait l'hypothèse que la socialisation favorisait la création de capital social et nous avons montré que c'était bien le cas, en regard de la centralité de proximité.

CHAPITRE V

CONCLUSION ET PERSPECTIVES

Nos contacts sociaux sont des atouts précieux lorsqu'il s'agit de trouver de l'information utile et c'est par le biais de la socialisation que nous entretenons et renouvelons nos réseaux de contacts personnels.

Le développement rapide des technologies, qui a favorisé l'ouverture ainsi que la décentralisation des sociétés traditionnelle et, plus récemment, la prolifération de divers médias sociaux, permet aujourd'hui d'avoir accès à une grande variété d'individus, provenant de divers milieux sociaux. Cependant, que ce soit dans nos réseaux sociaux réels ou virtuels, cette grande accessibilité à de nombreux individus ne facilite pas la tâche de localisation de ceux qui sont vraiment intéressants. Il faut du temps et un certain talent pour trouver les "bonnes connexions". C'est dans cette optique que nous avons proposé, dans le cadre de cette thèse, un modèle de socialisation automatique qui favorise la rencontre des individus étant intéressants et utiles les uns pour les autres.

Les règles d'évolution du modèle réalisé s'inspirent des mécanismes de socialisation qu'on observe dans nos réseaux sociaux habituels, et donc, gèrent un réseau complètement décentralisé. Bien que nous ayons conçu notre modélisation dans une optique d'implémentation fonctionnelle, nous avons aussi montré, d'un point de vue cognitif, que les mécanismes de socialisation formalisés pouvaient expliquer, dans une certaine mesure, la formation de nos réseaux sociaux usuels, au niveau structural.

5.1 Contributions

Les travaux réalisés dans cette thèse proposent donc un modèle d'évolution de réseaux, cohérent et fonctionnel, basé sur les mécanismes de socialisation. D'une part, notre modèle

reproduit et explique, en partie, certaines propriétés structurales typiques de nos réseaux sociaux usuels et d'autre part, il est informatiquement utilisable pour l'implémentation d'applications distribuées, basées sur la formation de communautés d'intérêts au sein d'un réseau social virtuel.

Au niveau de la modélisation sociocognitive, nous proposons notre modèle de socialisation comme modèle d'évolution des réseaux sociaux réels :

- Les règles d'évolution de notre modèle se basent sur l'observation des processus de socialisation qui ont lieu dans nos réseaux sociaux habituels (réels ou virtuels).
- Les réseaux générés par notre modèle montrent plusieurs des caractéristiques structurales qu'on remarque, de manière récurrente, dans nos réseaux sociaux usuels : l'effet des petits mondes, l'émergence de communautés structurales bien définies et une distribution des degrés suivant une loi de puissance. Nous avons montré, cependant, que nos réseaux ne montrent pas la propriété de transitivité, pour des raisons fonctionnelles.

Notre modèle explique aussi, dans une certaine mesure, l'utilité de la socialisation : les motivations sous-jacentes des individus qui socialisent peuvent se définir en termes de capital social. Notre modèle montre que la socialisation favorise effectivement la création d'un certain capital social qui a trait à la proximité des acteurs. Comme nous l'avons vu, cependant, nous ne pouvons pas discuter du capital social en termes de centralité de degré et d'intermédiarité, étant donné que les positions des acteurs les plus centraux et les plus intermédiaires (les nœuds pivots) sont assignées automatiquement par nos algorithmes et sont des positions strictement fonctionnelles qui n'offrent aucun avantage "social" particulier à ces acteurs pivots.

D'un point de vue fonctionnel, nous proposons des algorithmes d'évolution du réseau qui :

- *gèrent efficacement un réseau social dynamique en contexte décentralisé.* En n'utilisant que l'information locale, nos algorithmes arrivent à former, maintenir et renouveler les communautés d'intérêts au sein du réseau, et ce, malgré sa constante évolution.
- *implémentent des mécanismes de socialisation originaux* (les rencontres aléatoires et les recommandations) *qui favorisent clairement les regroupements d'individus* de profils

similaires en communautés d'intérêt. Plus précisément, ils avantagent l'homophilie globale et la rencontre des individus ayant des intérêts communs, au sein du réseau.

- *maintiennent des propriétés structurales qui favorisent la navigabilité.* Nos réseaux demeurent constamment, au cours de leur évolution, de densité très faible et de petit diamètre.
- *produisent des parcours efficaces dans le réseau.* Le type de regroupement structural que nous avons choisi pour former les communautés au sein de notre réseau permet de restreindre les recherches dans le graphe (les parcours de connexion/socialisation) en visitant seulement une très petite proportion des nœuds du réseau, les nœuds pivots. De plus, on a vu que le regroupement des individus par l'intermédiaire de nos mécanismes de socialisation favorisait la diminution du nombre de nœuds pivots dans le réseau. La navigation efficace due aux propriétés structurales imposées se voit donc renforcée par les mécanismes de socialisation et vice versa.
- *forment des communautés structurales non équivoques.* Une fois regroupés, les membres d'une même communauté peuvent avoir accès les uns aux autres, de manière optimale, pour favoriser les échanges.

Dans l'éventualité d'une implémentation de notre modèle, comme contribution secondaire, nous proposons un modèle générique qui permet de fournir une méthode de comparaison de profils sur mesure pouvant tirer parti du contexte d'utilisation. De plus, notre modèle prévoit un mécanisme automatique de mise à jour des relations lorsque le profil d'un individu change radicalement au cours du temps et qu'il ne se trouve plus dans une communauté qui lui ressemble. Ainsi, si l'on fournit une méthode de comparaison de profil capable de mettre à jour automatiquement le profil des utilisateurs, la mise à jour du profil et des relations peut se faire automatiquement.

Comme nous l'avons vu dans la littérature, on peut profiter des propriétés structurales déjà présentes dans nos réseaux pour améliorer leur navigabilité. On a profité de la présence de pages Web particulières qui sont des *authorities* ou des *hubs* pour améliorer la recherche d'informations dans le Web, on a étudié les caractéristiques des petits mondes et des réseaux en loi de puissance pour tenter de tirer parti des courts chemins dans les uns et des super connecteurs dans les autres. Au contraire, nous avons explicitement choisi des propriétés

structurales performantes pour notre modèle de réseaux, au lieu d'en dépendre et de nous y adapter.

Notre algorithme de parcours dans les graphes est un simple algorithme glouton, bien connu dans la littérature. L'originalité de notre approche réside dans le fait que nos parcours sont efficaces d'une part, grâce aux propriétés structurales imposées, mais d'autre part, parce qu'ils sont renforcés par les mécanismes de socialisations proposés. L'efficacité des parcours émerge d'une dynamique de renforcement entre la navigabilité et la socialisation. Les parcours de socialisation individuels, qu'on a voulu conserver simples et courts (pour une bonne performance en temps réel), ne sont pas toujours fructueux cependant, globalement ils le deviennent grâce à la restructuration du réseau par les mécanismes de fusion de communautés : une seule rencontre ou recommandation profite à plusieurs individus à la fois et l'on a vu que globalement, collectivement, cela fonctionne. L'effet global est en quelque sorte plus grand que la somme des effets individuels. On parle alors de cognition collective située socialement.

Ce projet de recherche se situe donc à l'intersection des études structurales des réseaux sociaux et des travaux effectués dans le domaine des systèmes sociaux et collaboratifs : nous associons les individus similaires par comparaison de profils d'intérêts (systèmes collaboratifs) et l'information obtenue est *aussitôt* enregistrée (de manière efficace) dans la structure du réseau sous forme de communautés structurales non équivoques (études structurales). C'est de cette manière que notre modèle arrive à former, maintenir et renouveler les communautés d'intérêt au sein d'un réseau complètement *décentralisé* et en constante évolution.

En somme, nous avons réalisé tous nos objectifs. Toutefois, il reste beaucoup à faire, tant sur le plan de la modélisation sociocognitive que du point de vue informatique. Nous exposons, à la section suivante, les travaux que nous prévoyons réaliser en continuité de ceux présentés dans le cadre de cette thèse.

5.2 Travaux futurs

5.2.1 Au niveau sociocognitif

5.2.1.1 Raffinement du modèle

Comme nous l'avons vu au chapitre précédent, nous pourrions raffiner notre modélisation en incorporant des règles de formation de *clustering* et certains ajustements au niveau du coefficient de loi de puissance de la distribution des degrés. Dans cette optique, nous voulons laisser de côté l'aspect fonctionnel du modèle en nous concentrant uniquement sur sa vraisemblance sociale. En effet, si nous n'avions pas minimisé la formation de liens dans nos algorithmes, pour favoriser une densité faible désirable d'un point de vue fonctionnel, il y aurait beaucoup plus de *clustering* au niveau des nœuds pivots. Nous voulons donc étudier les réseaux générés par notre modèle en retirant les mécanismes d'optimisation.

Nous voulons aussi parfaire notre modèle en tentant de voir, par exemple, s'il n'y aurait pas d'autres processus que celui de la fermeture des triangles qui sont responsables du fort taux de *clustering* dans nos réseaux sociaux usuels. Nous aimerions comprendre, aussi, pourquoi, dans nos réseaux sociaux, le coefficient de loi de puissance de la distribution des degrés se situe le plus souvent autour de 2 et 3. Toutes ces questions demandent une étude plus poussée des mécanismes de formation des réseaux sociaux.

En ce qui a trait à la formation de relations, le modèle proposé se base uniquement sur la similarité des intérêts entre les individus, l'objectif principal étant de créer et maintenir des communautés d'intérêts au sein du réseau. Or, les raisons qui motivent les individus à créer des liens dans les réseaux sociaux sont certainement plus diversifiées. Par exemple, la langue, l'âge, le bagage culturel, le niveau de formation générale sont certainement des facteurs qui influencent la création de liens entre les individus. Mais de plus, au-delà de la simple comparaison d'attributs pour déterminer la similarité entre les individus (homophilie), nous voudrions considérer une forme de réalisme psychologique plus actif, qui tient compte des motivations personnelles poussant les individus à se rencontrer (ou non). Par exemple, on peut vouloir se lier préférablement à des individus en qui l'on a confiance ou que l'on trouve

tout simplement sympathiques. On peut stratégiquement souhaiter, dans l'optique d'acquérir du capital social, rencontrer une personne parce qu'elle a du prestige ou du pouvoir, ou bien parce qu'on sait qu'elle-même connaît des individus prestigieux. On pourrait aussi considérer la variabilité du niveau de sociabilité intrinsèque des individus ; pour diverses raisons, certains individus sont plus habiles que d'autres lorsqu'il s'agit de socialiser.

Dans cette optique, nous voudrions rendre notre modèle plus réaliste en explorant davantage la littérature en psychologie sociale pour découvrir les mécanismes cognitifs qui motivent (ou découragent) les individus à vouloir former des relations. Nous pourrions alors complexifier la modélisation de nos agents. D'une part, pour le calcul de la similarité, nous pourrions considérer d'autres dimensions que celle des intérêts comme l'âge, la culture, la formation, etc. D'autre part, nous pourrions incorporer au modèle ces mécanismes cognitifs plus actifs qui viendraient aussi influencer la formation de liens entre les individus.

5.2.2 Au niveau informatique

5.2.2.1 Incorporation de centres d'intérêt diversifiés

Comme nous l'avons déjà mentionné au chapitre 3, notre modèle de socialisation ne permet pas la représentation de profils d'intérêts diversifiés. En réalité, la plupart des individus ont des centres d'intérêt divers. Nous prévoyons donc perfectionner notre modèle pour qu'il tienne compte de cet aspect en gérant plusieurs profils d'intérêts par utilisateur.

5.2.2.2 Implémentation d'un système distribué de socialisation automatique

En tirant parti des technologies d'Internet permettant de créer facilement des réseaux sociaux virtuels par l'intermédiaire d'applications distribuées qui connectent les utilisateurs entre eux, nous voulons concevoir un système de socialisation automatique simple qui implémente notre modèle. Nous expliquons ici comment fonctionnerait une telle application dans un réseau pair-à-pair (P2P).

Une architecture pair-à-pair consiste à créer des liens directs entre les ordinateurs (ou autres appareils mobiles) des utilisateurs du logiciel sans avoir à passer par un serveur central. Chaque utilisateur connecté n'a qu'une vue locale de son environnement dans le réseau, c'est-à-dire, dans notre cas, qu'il ne connaît que les membres de sa communauté. Contrairement aux architectures classiques client/serveur, les services et les données sont distribués parmi les utilisateurs où chaque utilisateur est à la fois un client/consommateur, et un serveur/fournisseur. Ce type de réseautage est en accord avec notre modèle de socialisation, car il ne suppose aucune coordination centrale et demande la participation volontaire de chaque pair.

Lors de la première utilisation, le système demanderait à l'utilisateur de fournir un nom d'utilisateur et des mots-clés qui représentent ses intérêts pour composer son profil. L'utilisateur pourrait évidemment modifier ce profil par la suite, s'il le désire. Ensuite, l'utilisateur se connecterait au réseau par l'intermédiaire du logiciel (installé sur son ordinateur ou son appareil mobile). Le système s'occuperait alors de socialiser pour lui et de trouver des communautés d'intérêt qui lui conviennent au sein du réseau, composé de tous les utilisateurs connectés. L'acte de socialisation pourrait être planifié pour s'exécuter à tous les intervalles de temps spécifiés ou bien s'effectuer à la demande de l'utilisateur.

La figure 5.1 illustre ce à quoi pourrait ressembler une telle application dans sa plus simple expression. L'interface afficherait une liste de tous les membres de la communauté d'appartenance de l'utilisateur et lorsque celui-ci sélectionnerait un nom dans la liste, le profil de ce membre s'afficherait juste à côté. Pour permettre les échanges, on aurait un espace de clavardage pour communiquer avec l'utilisateur sélectionné. On devrait aussi pouvoir envoyer des messages à toute sa communauté d'appartenance.

Un système de ce genre pourrait être utile lors de congrès, par exemple. Les organisateurs du congrès n'auraient qu'à fournir l'application et tout le reste serait accompli par les utilisateurs et le logiciel. Lorsqu'un utilisateur se connecterait au réseau du congrès, l'application socialiserait pour lui et lui fournirait une liste d'utilisateurs avec qui il a des affinités. Ceux-ci pourraient alors bavarder ensemble sur des sujets d'intérêts, par l'intermédiaire de l'espace de

clavardage. Aussi, étant donné le contexte, les membres d'une communauté pourraient même se donner rendez-vous sur le site du congrès, pour une rencontre plus personnelle.

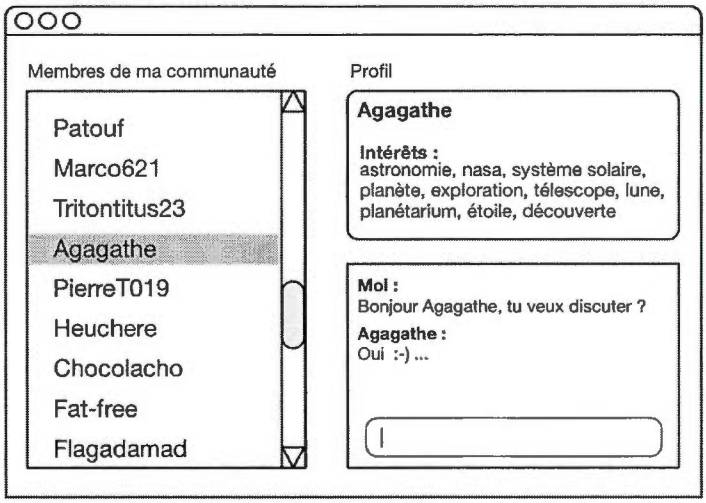


Figure 5.1 Interface simple pour une application de socialisation automatique.

On pourrait aussi implémenter ce genre d'application pour créer un réseau de chercheurs dont le profil représenterait ses intérêts de recherche. L'utilisateur qui se connecterait au réseau obtiendrait une liste d'autres utilisateurs qui ont des intérêts de recherche connexes. De plus, en supposant aussi que le profil des utilisateurs comprenne des informations sur la localité des individus ou l'université d'attache, par exemple, cette application pourrait être utile, entre autres, lorsqu'un nouveau professeur (ou un nouvel étudiant) arrive dans une nouvelle université (d'un autre pays) et qu'il veut rapidement localiser des chercheurs locaux qui partagent ses intérêts. Il pourrait ainsi amorcer de nouvelles collaborations.

Les applications sont nombreuses, mais l'avantage majeur d'une architecture pair-à-pair (décentralisée) est que celle-ci ne nécessite qu'une application pour se connecter au réseau, et c'est tout. Nul besoin de se procurer des serveurs pour stocker les données et d'avoir recours à des individus pour maintenir ces serveurs. Il n'y a pas de frais centralisés pour la bande passante ou l'hébergement d'un site Web, par exemple. Les frais et les ressources nécessaires sont partagés par tous les utilisateurs du réseau : chaque utilisateur fournit son espace de stockage et assume le coût de sa propre connexion Internet.

Le fait de concevoir une telle application nous permettra d'aborder les problèmes liés aux réseaux pair-à-pair comme la quantité de bande passante disponible au niveau des utilisateurs. En effet, peut-être que nous devrons ajuster nos algorithmes pour qu'ils tiennent compte des ressources disponibles d'un utilisateur avant de le choisir comme nœud pivot. Nous devrons aussi prévoir un certain degré de redondance ainsi que des mécanismes de reprises pour conserver l'intégrité du réseau (dans le cas d'attaques ou de malfonctionnement). Nous pourrons évaluer si le taux de connexions et de déconnexions des utilisateurs affecte la dynamique du réseau, etc.

De plus, en validant notre application auprès d'utilisateurs réels et en recueillant leurs commentaires, nous pourrons améliorer notre système et mieux ajuster certains paramètres comme la limite d'équivalence, par exemple, qui détermine le niveau de similarité nécessaire pour que deux individus soient considérés comme identiques lors des parcours de socialisation dans le réseau.

5.2.2.3 Ajout d'un mécanisme d'échange d'informations automatique

Lorsque nous aurons réglé les problèmes liés à la diversité des profils et aux réseaux pair-à-pair, nous prévoyons concevoir un système de socialisation avec partage automatique de l'information.

Pour le bon fonctionnement d'une telle application, les utilisateurs du système doivent accepter de partager certaines ressources (documents textes, vidéo ou audio, références bibliographiques, URL, etc.) qu'ils doivent aussi étiqueter de mots-clés décrivant leur contenu. En supposant raisonnablement que les ressources qu'un utilisateur décide de partager représentent ses intérêts, comme nous l'avons expliqué au chapitre 3, l'agrégation des mots-clés de toutes les ressources partagées peut servir à représenter le profil de l'utilisateur sous forme de vecteur. Nous pouvons ensuite comparer les profils dans un espace vectoriel commun.

Comme nos mécanismes de socialisation ont pour effet de regrouper les individus ayant des intérêts communs (basés, ici, sur les ressources partagées), on peut supposer que les

documents partagés par les individus qui se trouvent au sein d'une même communauté sont intéressants les uns pour les autres. Nous voulons donc implémenter un mécanisme de partage automatique des ressources au sein des communautés.

Pour éviter une distribution massive de l'information, nous devrons concevoir des méthodes capables de déterminer quelles ressources sont pertinentes pour quels individus dans la communauté. Pour décider si un individu i serait potentiellement intéressé à recevoir une ressource x, nous pourrions examiner la collection de ressources de i et voir, dans quelle mesure, celle-ci contient des ressources dont les mots-clés correspondent aux mots-clés associés à la ressource x.

Plus précisément, nous pourrions tenir compte du contexte de chaque mot-clé (les autres mots-clés qui l'accompagnent) lorsque nous comparons les ressources ceci, dans le but d'effectuer une certaine désambiguïsation. Par exemple, la figure 5.2 montre quatre ressources différentes, chacune d'entre elles étant associée à quatre mots-clés. Le mot "aube" est parmi les mots-clés de chacune des ressources, mais en regardant les autres mots-clés de chaque ressource, on comprend qu'il ne s'agit pas de la même chose.

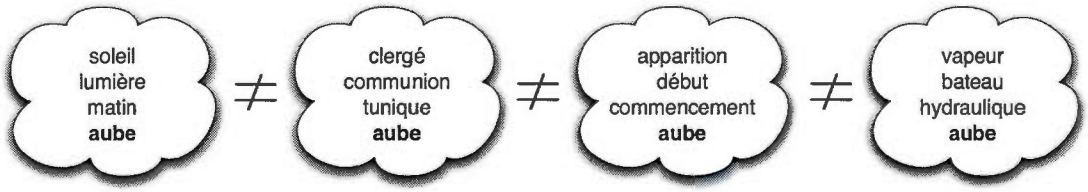


Figure 5.2 Désambiguïsation de termes par contextualisation.

Par ailleurs, comme nous l'avons brièvement mentionné au chapitre 1, l'échange de mots-clés permettrait d'enrichir le vocabulaire des uns par le vocabulaire des autres et pourrait de cette manière minimiser (dans une certaine mesure) le problème relié à la synonymie de certains termes. Par exemple, chaque fois qu'un utilisateur recevrait une ressource qu'il déciderait de conserver, l'application lui demanderait de lui assigner des mots-clés qui viendraient s'ajouter aux mots-clés déjà assignés à cette ressource par d'autres utilisateurs avant lui. Ainsi, chaque fois qu'un utilisateur ajouterait des mots-clés, ceux-ci seraient propagés à tous les membres

de sa communauté qui possèdent cette même ressource dans leur collection partagée (et viendraient s'ajouter automatiquement à la liste des mots-clés de cette ressource).

Dans cette optique, supposons, par exemple, qu'un utilisateur *a* ait assigné le mot-clé "science-fiction" à la ressource *x* et qu'un utilisateur *b* lui ait assigné le mot-clé "sci-fi". En ne considérant que ces deux mots-clés, notre système n'a aucun moyen de savoir que les deux termes sont des synonymes et il ne détectera pas leur similarité. Cependant, en agrégeant les mots-clés par propagation, la ressource *x* possède désormais les deux termes et l'on pourra ultérieurement détecter une ressemblance entre cette ressource *x* et une autre ressource qui contient soit le mot "science-fiction", soit le mot "sci-fi".

L'extraction des profils, basée sur les ressources partagées, permet aussi la mise à jour implicite (automatique) des profils d'intérêt. En effet, chaque fois que l'utilisateur modifie sa collection de ressources partagées (ajout ou suppression de ressources ou modification de la liste des mots-clés associée aux ressources), son profil s'ajuste implicitement. Ensuite, le mécanisme de mise à jour des relations, qui peut vérifier, à intervalle régulier, si le profil a changé radicalement, s'exécutera automatiquement si c'est le cas.

De plus, avec des utilisateurs réels, nous pourrions aussi implémenter un système d'évaluation des ressources partagées par les membres d'une communauté (comme dans les systèmes de recommandation collaboratifs). Nous pourrions ensuite ordonnancer les ressources selon leur appréciation globale avant de les présenter aux membres susceptibles d'être intéressées par ces ressources.

Toutes ces considérations ne sont que quelques exemples des nombreuses pistes de recherche possibles. En effet, il est assez intuitif, mais pas certain qu'un tel mécanisme d'échange d'informations entre les membres d'une même communauté soit vraiment pertinent pour les utilisateurs. C'est ce que nous voudrions vérifier en implémentant cette application et en recueillant les commentaires et évaluations des participants.

BIBLIOGRAPHIE

Adamic L. A. (1999). The small world web. In *Proceedings of the third European conference on research and advanced technology for digital libraries*, Springer, Berlin, p. 443–452.

Adamic L. A., et Huberman, B. A. (2000). Power-law distribution of the world wide web. *Science,* vol. 287, p. 2115.

Adamic L. A., Lukose, R. M., Puniyani, A. R., et Huberman, B. A. (2001). Search in power-law networks. *Phys. Rev. E.*, vol. 64, p. 046135.

Adamic L. A., Büyükkökten O., et Adar E. (2003). A social network caught in the Web. *First Monday*, vol.8, no 6.

Adomavicius G., et Tuzhilin A. (2005). Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Trans. Knowledge and Data Eng.*, vol. 17, no 6, p. 734-49.

Aiello W., Chung F., et Lu L. (2000). A random graph model for massive graphs, In *Proceedings of the 32nd Annual ACM Symposium on Theory of Computing*, p.171-180, Association of Computing Machinery, NewYork.

Aiello W., Chung F., et Lu L. (2002). Random evolution of massive graphs. In *Handbook of Massive Data Sets*, J. Abello, P. M. Pardalos, et M. G. C. Resende (eds.), Springer, p. 97-122.

Albert R. H., et Barabási A.-L. (2000). Topology of evolving networks : local events and universality. *Physical Review Letters*, vol. 85, p. 5234.

Albert R. H., Jeong H. et Barabási A.-L. (1999). Diameter of the World-Wide Web. *Nature*, vol. 401, p. 130-131.

Albert R. H., et Barabási A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.* vol. 74, p. 47-97.

Amaral L. A. N., Scala A., Barthélémy M. et Stanley H. E. (2000). Classes of small-world networks. In *Proceedings of National Academy of Sciences USA*, vol. 97, p. 11149-11152.

Barabási A.-L., Albert R., et Jeong H. (1999). Mean-field theory for scale-free random networks. *Physica A*, vol. 272, p. 173-187.

Barabási A.-L., et Albert R. (1999) Emergence of scaling in random networks. *Science*, vol. 286, no 5439, p. 509-512.

Barabási A.-L., Jeong H., Ravasz E., Néda Z., Schuberts A., et Vicsek T. (2002). Evolution of the social network of scientific collaborations. *Physica A : Statistical Mechanics and its Applications*, vol. 311, no 3-4, p. 590-614.

Bender E. A., et Canfield E. R. (1978). The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory A*, vol. 24, p. 296-307.

Bianconi G., et Barabási A.-L. (2001). Competition and multiscaling in evolving networks. *Europhys. Lett.*, vol. 54, p. 436-442.

Borgatti S. P., Jones C., et Everett M. G. (1998). Network Measures of Social Capital. *Connections*, vol. 21, no 2, p. 27-36.

Brin S., et Page L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australie, p. 107-117.

Broder A., Kumar R., Maghoul F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., et Wiener J. (2000). Graph structure in the Web. *Computer Networks*, vol. 33, p. 309-320.

Brown J. S., Collins A., et Duguid P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, vol. 18, no 1, p. 32-42.

Burt R. S. (1992). *Structural Holes*. Cambridge : Cambridge University Press.

Castells M. (2001). *The internet galaxy,* Oxford : Oxford University Press.

Cattuto C. (2006). Semiotic dynamics in online social communities. *The European Physical Journal C-Particles and Fields*, vol. 46, p. 33-37.

Chen Q., Chang H., Govindan R., Jamin S., Shenker S. J., et Willinger W. (2002). The origin of power laws in Internet topologies revisited. In *Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies*, vol 2, p. 608-617.

Chung F., et Lu L. (2002). The average distances in random graphs with given expected degrees. In *Proc. Natl. Acad. Sci. USA,* vol. 99, p. 15879-15882.

Crandall D., Cosley D., Huttenlocher D., Kleinberg J., et Suri S. (2008). Feedback effects between similarity and social influence in online communities. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '08)*. ACM, New York, NY, USA, p. 160-168.

Davidsen J., Ebel H., et Bornholdt S. (2002). Emergence of a small world from local interactions: Modeling acquaintance networks. *Phys. Rev. Lett.*, vol. 88, p. 128701.

Dellschaft K., et Staab S. (2008). An epistemic dynamic model for tagging systems. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia (HT '08)*. ACM, New York, NY, USA, p. 71-80.

Diederich J., et Iofciu T. (2006). Finding Communities of Practice from User Profiles Based On Folksonomies. In *Proc. of the 1st International Workshop on Building Technology Enhanced Learning solutions for Communities of Practice (TEL-CoPs '06)*, Crete, Greece.

Dodds P. S., Muhamad R., et Watts D. J. (2003). An experimental study of search in global social networks, *Science*, vol. 301, no 5634, p. 827–829.

Dorogovtsev S. N., et Mendes J. F. F. (2000a). Exactly solvable analogy of small-world networks. *Europhysics Letters (EPL)*, vol. 50, no 1, p. 1-7.

Dorogovtsev S. N., et Mendes J. F. F. (2000b), Evolution of reference networks with aging. *Physical Review E*, vol. 62, p. 1842.

Dorogovtsev S. N., et Mendes J. F. F. (2001). Effect of the accelerating growth of communications networks on their structure. *Phys. Rev. E*, vol. 63, no 2, p. 025101.

Dourish P., et Chalmers M. (1994). Running out of space: models of information navigation. In *Proceedings of HCI'94*. Glasgow.

Drucker P. F. (1970). Technology, Management and Society. Harper and Row, New York.

Ebel H., Mielsch L.-I., et Bornholdt S. (2002). Scale-free topology of e-mail networks. *Phys. Rev. E*, vol. 66, p. 035103.

Erdös P., et Rényi A. (1960). On the evolution of random graphs. *Publications of the Mathematics Institute of Hungarian Academy of Science*, vol. 5, p.17-61.

Faloutsos M., Faloutsos P., et Faloutsos C. (1999). On power-law relationships of the Internet topology. *SIGCOMM Comput. Commun. Rev.*, vol. 29, no 4, p. 251-262.

Farzan R. et Brusilovsky P. (2005). Social navigation support through annotation-based group modeling. In *Proc. of 10th International User Modeling Conference. Lecture Notes in Artificial Intelligence*, Ardissono L., Brna P., Mitrovic A. (eds.), Springer Verlag, vol. 3538, p. 463-472.

Fell D. A., et Wagner A. (2000). The small world of metabolism, *Nature Biotechnology*, vol. 18, p. 1121-1122.

Ferrara E., et Fiumara G. (2011). Topological Features of Online Social Networks. *Communications on Applied and Industrial Mathematics*, North America, vol. 2, no 2, p. 1-20.

Ferrer-i-Cancho R. et Solé R. V. (2001). The small world of human language. In *Proceedings of The Royal Society of London B*, vol. 268, no 1482, p. 2261-2265.

Foray D. (2000). *L'économie de la connaissance*. Paris, Repères, La Découverte.

Freyne J., Farzan R., Brusilovsky P., Smyth B. et Coyle M. (2007). Collecting community wisdom: integrating social search and social navigation, In *Proceedings of the 12th international conference on Intelligent user interfaces*, Honolulu, Hawaii, USA.

Friedkin N. E. (1990). Social Networks in Structural Equation Models. *Social Psychology Quarterly*, vol. 53, no 4, p. 316–328.

Giles C. L., Bollacker K. D., et Lawrence S. (1998). CiteSeer : an automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries* (Pittsburgh, Pennsylvanie, US, juin 23 - 26, 1998). I. Witten, R. Akscyn, and F. M. Shipman (eds). DL '98. ACM. New York (NY).

Girvan M., et Newman M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences (PNAS)*, vol. 99, no 12, p. 7821-7826.

Golder S. A., et Huberman B. A. (2006). The Structure of Collaborative Tagging Systems. *Journal of Information Science*, vol. 32, no 2, p. 198-208.

Granovetter M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, vol. 78, no 6, p. 1360-1380.

Guare J. (1990). *Six Degrees of Separation : a play*. Vintage, New York.

Harnad S. (2005). Distributed Processes, Distributed Cognizers and Collaborative Cognition. *Pragmatics & Cognition*, vol. 13, no 3, p. 501-514. (Special Issue on Cognitive Technology: Distributed Cognition).

Heymann P., Koutrika G., et Garcia-Molina H. (2008). Can Social Bookmarking Improve Web Search?. In *WSDM '08: Proceedings of the intl. conf. on Web search and web data mining*, Palo Alto, California, US, New York (NY). ACM, p. 195-206.

Holme P., et Kim B. J. (2002). Growing scale-free networks with tunable clustering. *Phys. Rev. E*, vol. 65, p. 026107.

Huberman B. A. (2001). *The Laws of the Web*, MIT Press, Cambridge, 115 p.

Jespersen S., et Blumen A. (2000) Small-world networks : Links with long-tailed distributions, *Phys. Rev. E*, vol 62, p.6270-6274.

Jin E. M., Girvan M., et Newman M. E. J. (2001). The structure of growing social networks. *Phys. Rev. E*, vol. 64, p. 046132.

Kelly D., et Teevan J. (2003). Implicit Feedback for Inferring User Preference: A Bibliography. *SIGIR Forum*, vol. 37, no 2, p. 18-28.

Kleinberg J. M. (1999). Authoritative sources in a hyperlinked environment. *J. ACM*, vol. 46, no 5, p. 604-632.

Kleinberg J. M. (2000). Navigation in a small world. *Nature*, vol. 406, no 845.

Kleinberg J. M. (2000b). The small-world phenomenon: an algorithm perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing* (STOC '00). ACM, New York, NY, USA, p. 163-170.

Kleinberg J. M. (2001). Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems (NIPS),* vol. 14.
(récupéré sur le site : http://books.nips.cc/nips14.html)

Kleinberg J. M., et Lawrence S. (2001). The Structure of the Web. *Science*, vol. 294, p. 1849-1850.

Klemm K., et Eguiluz V. M. (2002). Highly clustered scale-free networks. *Phys. Rev. E*, vol. 65, p. 036123.

Kirsch S., Gnasa M., et Cremers A. (2006). Beyond the Web: retrieval in social information spaces. In *Proc. 28th European Conference on Information Retrieval.*

Konstan J. A., Miller B. N., Maltz D., Herlocker J. L., Gordon L. R. et Riedl J. (1997). GroupLens: Applying Collaborative Filtering to Usenet News. *Communications of the ACM*, vol. 40, no 3, p. 77-87.

Korzeny F.(1978). A theory of electronic propinquity : Mediated communication in organizations. *Communication Research*, vol. 5, p. 3-23.

Kumar R., Raghavan P., Rajagopalan S., et Tomkins A. (1999). Trawling the Web for Emerging Cyber-Communities. *Computer Networks*, vol. 31, p. 1481–1493.

Kumar R., Novak J., et Tomkins A. (2006). Structure and evolution of online social networks. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '06)*. ACM, New York, NY, USA, p. 611-617.

Lazega E. (1998). *Réseaux sociaux et structures relationnelles*. Paris, Que sais-je? No 3399, PUF.

Lazer D., et Friedman A. (2005). The hare and the tortoise : The network structure of exploration and exploitation. *ACM International Conference Proceeding Series*, vol. 89, p. 253-254.

Leskovec J., Kleinberg J. M., et Faloutsos C. (2007). Graph evolution: Densification and shrinking diameters. *ACM Trans. Knowl. Discov. Data*, vol 1, no 1, art. 2.

Lewis K., Kaufman J., Gonzalez M., Wimmer A., and Christakis N., Tastes, ties, and time: A new social network dataset using Facebook.com, *Social Networks*, vol. 30, no 4, p. 330-342.

Liljeros F., Edling C. R., Amaral L. A. N., Stanley H. E., et Aberg Y. (2001). The web of human sexual contacts. *Nature* (London), vol. 411, p. 907-908.

Lin N. (1999). Building a network theory of social capital. *Connections*, vol. 22, no 1, p. 28-51.

Linden G., Smith B., and York J. (2003) Amazon.com Recommendations: Item-to-Item Collaborative Filtering. In *Internet Computing IEEE,* (Jan.-Fev. 2003), vol. 4, no 1, p. 76-80.

Lord M. (2007). NetSim : un logiciel de modélisation et de simulation de réseaux d'information. Maîtrise en Informatique, UQAM, Montréal.

Lun L., Alderson D., Doyle J. C., et Willinger W. (2006). Towards a Theory of Scale-Free Graphs: Definitions, Properties, and Implications. *Internet Mathematics*, vol. 2, no 4, p. 431-523.

Manning C. D., Raghavan P., et Schütze H. (2008). *Introduction to Information Retrieval*, Cambridge University Press, 496 p.

Mason W. A., Jones A., et Goldstone R. L. (2008). Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, vol. 137, no 3 p. 422-433.

McPherson M., Smith-Lovin L., et Cook J. M. (2001) Birds of a Feather : Homophily in Social Networks, *Annual Review of Sociology*, vol. 27, no 1, p.415-444.

Memmi D., et Nérot O. (2003). Building virtual communities for information retrieval, in *Groupware: Design, Implementation and Use*, Favela et Decouchant (eds), Springer, Berlin.

Memmi D. (2009). Sociology of virtual communities and social software design, In S. Murugesan (ed), *Handbook of Research on Web 2.0, 3.0, and X.0*, Information Science Reference, IGI Global.

Menczer F., et Belew R. K. (2000). Adaptive retrieval agents: Internalizing local context and scaling up to the Web. *Machine Learning*, vol. 39 no 2-3, p. 203-242.

Michlmayr E., et Cayzer S. (2007). Learning User Profiles from Tagging Data and Leveraging them for Personal(ized) Information Access. In *Proceedings of the Workshop on Tagging and Metadata for Social Information Organization, 16th International World Wide Web Conference WWW2007.*

Mika P. (2005). Ontologies are us: A unified model of social networks and semantics. In Gil Y. et al. (éds). *The Semantic Web – ISWC 2005*, p 522-536. Berlin, Heidelberg. Springer.

Milgram S. (1967). The small world problem, *Psychology Today*, vol. 2, p. 61–67.

Millen D. R., et Feinberg J. (2006). Using Social Tagging to Improve Social Navigation. In *Proceedings of the Workshop on the Social Navigation and Community-Based Adaptation Technologies at AH 2006*. Dublin, Ireland, p. 532–541.

Mislove A., Marcon M., Gummadi K. P., Druschel P., et Bhattacharjee B. (2007). Measurement and Analysis of Online Social Networks. In *Proc. Internet Measurement Conference*, 2007.

Mislove A. (2009). Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems. PhD thesis, Rice University, 2009.

Molloy M., et Reed B. (1995). A critical point for random graphs with a given degree sequence. *Random Structures and Algorithms*, vol 6, p. 161-179.

Motter A. E., de Moura A. P., Lai Y.-C., et Dasgupta P. (2002). Topology of the conceptual network of language, *Phys. Rev. E*, vol. 65, p.065102.

Moukarzel C. F., et de Menezes M. A. (2002). Shortest paths on systems with power-law distributed long-range connections. *Phys. Rev. E*, vol. 65, p.056709 (2002).

Nazir A., Raza S., et Chuah C.-C. (2008). Unveiling facebook: a measurement study of social network based applications. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement (IMC '08)*. ACM, New York, NY, USA, p. 43-56.

Nekovee M., Moreno Y., Bianconi G. et Marsili, M. (2007). Theory of rumour spreading in complex social networks. *Physica A*, vol. 374, no 1, p. 457–470.

Newman M. E. J. (1999). *Small worlds : the structure of social networks*. Working Papers 99-12-080, Santa Fe Institute.

Newman M. E. J. (2001a), Clustering and preferential attachment in growing networks, *Physical Review E*, vol. 64, no 2, p.025102.

Newman M. E. J. (2001b). The Structure of Scientific Collaboration Networks. In *Proceedings of the National Academy of Sciences (PNAS)*, vol. 98, no 2, p. 404-409.

Newman M. E. J., Strogatz S. H., et Watts D. J. (2001) Random graphs with arbitrary degree distributions and their applications. Phys. Rev. E, vol. 64, no 2, p. 026118.

Newman M. E. J. (2003). The structure and function of complex networks, *SIAM Reviews*, vol. 45, no 2, p. 167-256.

Newman M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev. E*, vol. 69, no 6, p. 066133.

Newman M. E. J. (2004b). Finding and evaluating community structure in networks. *Phys. Rev. E*, vol. 69, no 2, p. 026113.

Nonaka I., et Takeuchi H. (1995). *The Knowledge-Creating Company : How Japanese Companies Create the Dynamics of Innovation*. New York. Oxford University Press.

Page L., Brin S., Motwani R., et Winograd T. (1998). The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University, 1998.

Papagelis A., Papagelis M., et Zaroliagis C. (2008). Enabling Social Navigation on the Web. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-AT)*, vol. 1, p. 162-168.

Porter M. F. (1980). An Algorithm for Suffix Stripping. *Program*, vol. 14, no 3, p. 130-137.

Potterat J. J., Phillips-Plummer L., Muth S. Q., Rothenberg R. B., Woodhouse D. E., MaldonadoLong T. S., Zimmerman H. P., et Muth, J. B. (2002). Risk network structure in the early epidemic phase of HIV transmission in Colorado Springs. *Sexually Transmitted Infections*, vol. 78, p. i159-i163 (2002).

Redner S. (1998). How popular is your paper? An empirical study of the citation distribution, *Eur. Phys. J. B*, vol. 4, p.131-134.

Resnick P., Lacovou N., Suchak M. et Bergstorm J. R. P. (1994). GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proc. of the ACM 1994 Conference on Computer Supported Cooperative Work (CSCW '94)*, p. 175–186.

Robbins P., et Aydede M. (eds.) (2009). *The Cambridge Handbook of Situated Cognition*. Cambridge. Cambridge University Press.

Salomon G. (1993). Introduction de l'éditeur In G. Salomon (Ed.), *Distributed cognitions*, p. xi-xii, New York : Cambridge University Press.

Saramaki J., et Kaski K. (2004). Scale-free networks generated by random walkers. *Physica A*, vol. 341, p.80.

Scott J. (2000). *Social Network Analysis: A Handbook*, Sage Publications, London, 2e ed.

Simmel G. (1989). *Philosophie des Geldes*. Frankfurt, Suhrkamp. [première publication: 1900].

Smyth B., Balfe E., Freyne J., Briggs P., Coyle M. et Boydell O. (2005). Exploiting Query Repetition and Regularity in an Adaptive Community-Based Web Search Engine. *User Modeling and User-Adapted Interaction*, vol. 14, no 5, p. 383-423.

Solé R. V., et Montoya J. M. (2001). Complexity and fragility in ecological networks, In *Proceedings of the Royal Society B*, London, vol. 268, no 1480, p. 2039–2045.

Solomonoff R., et Rapoport A. (1951). Connectivity of random nets. *Bulletin of Mathematical Biophysics*, vol. 13, p. 107-117.

Teevan J., Dumais S. T., et Horvitz E. (2005). Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference (SIGIR'05)*, Salvador, Brazil, August 2005.

Tönnies F. (1963). *Gemeinschaft und Gesellschaft*. Darmstad: Wissenschaftliche Buchgesellschaft. [première publication: 1887].

Traud A. L., Mucha P. J., et Porter M. A. (2012) Social structure of Facebook networks. *Physica A: Statistical Mechanics and its Applications*, vol. 391, no 16, p. 4165-4180.

Vásquez A. (2003). Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Physics Review E*, vol. 67, no 5, p. 056104.

Vázquez A., Boguñá M., Moreno Y., Pastor-Satorras R., et Vespignani A. (2003). Topology and correlations in structured scale-free networks. *Phys. Rev. E*, vol. 67, no 4, p. 046111.

Wallace P. (1999). *The Psychology of the Internet*. Cambridge, UK, Cambridge University Press.

Wassermann S., et Faust K. (1994). *Social Network Analysis: Methods and Applications*, Cambridge University Press, Cambridge.

Watts D., et Strogatz S. H. (1998). Collective Dynamics of small world networks. *Nature*, vol. 393, p. 440-442.

Watts D. J. (1999). Networks, dynamics, and the small world phenomenon. *Am. J. Sociol.*, vol. 105, p. 493-592.

Watts D. J. (1999a) *Small Worlds*, Princeton University Press, Princeton, NJ.

Watts D. J., Dodds P. S., et Newman M. E. J. (2002). Identity and search in social networks. *Science*, vol. 296, no 5571, p. 1302-1305.

Weber M. (1956). *Wirtschaft und Gesellschaft*. Tübingen: Mohr. [first publication: 1925].

Wellman B. (2001). Physical place and cyber-place: changing portals and the rise of networked individualism. *International Journal for Urban and Regional Research,* vol. 25, no 2, p. 227-252.

Wexelblat A., et Maes P. (1999). Footprints: History-rich tools for information foraging. In *ACM Conference on Human-Computer Interaction (CHI'99)*. Pittsburgh (PA), p. 270-277. New York (NY). ACM.

White J. G., Southgate E., Thompson J. N., et Brenner S. (1986). The structure of the nervous system of the nematode C. Elegans. *Phil. Trans. R. Soc. London*, vol. 314, p. 1-340.

Zanette D. H. (2002). Dynamics of rumor propagation on small-world networks. *Phys. Rev. E*, vol. 65, no 4, p. 041908.

Zubiaga, A. (2009). Enhancing Navigation on Wikipedia with Social Tags. In *Proceedings:104 of Wikimania 2009, the 5th International Conference of the Wikimedia Foundation*. Buenos Aires, Argentine.