Augmentation de la cohérence des communautés détectées dans des graphes dynamiques par une approche multi-agent

Mémoire de Master 2 Recherche
Intelligence Artificielle, Intelligence collective, Interaction(IAICI)
à l'Université Paul Sabatier - Toulouse III

VERSTAEVEL Nicolas

Laboratoire d'accueil : Institut de Recherche en Informatique de Toulouse (IRIT)
Responsables de stage : Amblard Frédéric, Cazabet Rémy
Equipe d'accueil : Systèmes Multi-Agents Coopératifs (SMAC)

Mots-clés : Réseaux sociaux, détection de communautés, SMA

Résumé : L'analyse de réseaux sociaux est un outil qui s'impose dans de nombreuses sciences. Un de ces outils spécifiques à l'analyse de réseaux sociaux est la détection de communautés. De nombreux algorithmes de détection de communautés ont été développés mais beaucoup ont une approche statique, c'est à dire ne considèrent pas que l'ordre d'apparition à une importance. De plus, ils posent le problème de la robustesse, car ces différents algorithmes proposent des résultats très différents.
L'objectif de ce travail est de présenter une méthodologie permettant d'améliorer la cohérence des communautés détectées en inscrivant dans un graphe des informations sur sa dynamique grâce à une pondération des liens basés sur la topologie et la dynamique du réseau, et en renforçant, à l'aide de cette pondération, ses structures communautaires.
Les résultats de la méthodologie développés tendent à confirmer son effet sur la cohérence des communautés et montrent que non seulement, la dynamique est bien une source d'informations, mais qu'elle peut aussi servir à créer de la connaissance en permettant la création de liens transitifs, et favorisant deux nouveaux types de communautés, l'une, locale, permettant de s'intéresser à la détection de communautés d'acteurs contemporains, l'autre, étendue, s'intéressant à la détection de communautés persistantes dans le temps.

Augmentation de la cohérence des communautés détectés dans des graphes dynamiques par une approche multi-agent

VERSTAEVEL Nicolas

14 juin 2012

Résumé

L'analyse de réseaux sociaux est un outil qui s'impose dans de nombreuses sciences. Un de ces outils spécifiques à l'analyse de réseaux sociaux est la détection de communautés. De nombreux algorithmes de détection de communautés ont été développés mais beaucoup ont une approche statique, c'est à dire ne considèrent pas que l'ordre d'apparition à une importance. De plus, ils posent le problème de la robustesse, car ces différents algorithmes proposent des résultats très différents. L'objectif de ce travail est de présenter une méthodologie permettant d'améliorer la cohérence des communautés détectées en inscrivant dans un graphe des informations sur sa dynamique grâce à une pondération des liens basés sur la topologie et la dynamique du réseau, et en renforçant, à l'aide de cette pondération, ses structures communautaires. Les résultats de la méthodologie développés tendent à confirmer son effet sur la cohérence des communautés et montrent que non seulement, la dynamique est bien une source d'informations, mais qu'elle peut aussi servir à créer de la connaissance en permettant la création de liens transitifs, et favorisant deux nouveaux types de communautés, l'une, locale, permettant de s'intéresser à la détection de communautés d'acteurs contemporains, l'autre, étendue, s'intéressant à la détection de communautés persistantes dans le temps.

Abstract

Social network analysis(SNA) is a tool used in many sciences. One such tool specific to SNA is community detection. Many community detection algorithms have been developed but most of them have a static approach, ie do not consider the order of links appearance has important. In addition, they raise the problem of robustness, because these algorithms offer very different results. The objective of this work is to present a methodology to improve the coherence of detected communities by writing information on its dynamics in a graph, by weighting links based on the topology and dynamics of the network and strengthening, with these weights, its community structures. The results of the methodology developed tend to confirm its effects on the consistency of the community and show that not only the dynamics is a source of information, but it can also be used to create knowledge by allowing the creation of transitive links, and encouraging new kinds of communities, one local, to take an interest in community detection of contemporary actors, the other extended, with an interest in the detection of persistent communities over time.

Table des matières

Remerciements

Je remercie particulièrement Frédéric Amblard, tout d'abord d'avoir été mon encadrant pour la réalisation de ce projet et de m'avoir accordé sa confiance, mais aussi pour sa disponibilité, ses conseils, sa pédagogie et ses relectures, qui ont, pour beaucoup, contribué à la qualité de mes travaux.

Je remercie tout aussi chaleureusement Remy Cazabet, qui a consacré beaucoup de temps à m'encadrer, pour son sens de la pédagogie et ses conseils toujours avisés, pour son sens critique qui pousse à l'excellence, mais aussi pour la sympathie dont il a su faire preuve.

Je remercie tout les occupants du bureau 310 pour leur convivialité, leur bonne humeur, leur soutien et conseils tout au long de ce stage.

Je remercie également tout les membres de l'équipe SMAC pour leur sympathie et leur accueil chaleureux et plus particulièrement la responsable Marie-Pierre Gleize, pour le soutien et l'encadrement qu'elle a su nous apporter avant et pendant ce stage.

Enfin, je remercie celles et ceux, proches, famille, et inconnus(es) qui m'auront permis de m'épanouir dans mon travail.

A tous, sincèrement.

Introduction

Les thématiques de recherche de l'équipe $SMAC$ (Systèmes Multi-Agent Coopératifs) au sein de l'$IRIT$ portent sur la conception de systèmes informatiques et l'analyse d'organisations sociales robustes et pérennes évoluant de façon autonome pour s'adapter aux évolutions de leur environnement. Elle est le résultat d'un processus auto-organisationnel de chercheurs convergeant de plusieurs horizons : intelligence artificielle distribuée, systèmes distribués, simulations sociales, optimisation par recherche locale. Aujourd'hui confirmée par les faits, la problématique scientifique de l'équipe SMAC s'inscrit dans une évolution de l'étude des systèmes naturels et artificiels sous trois aspects, leur diversité, leur complexité et leur dynamique.

C'est sur ce dernier aspect, la dynamique, qu'une partie de l'équipe travaille sur l'analyse de réseaux sociaux pour permettre l'étude et la compréhension des processus d'interaction entre agents et la détection de structures communautaires. De nombreux travaux [Amblard et al.(2011)] [Sueur et al.(2011)] [Quattrociocchi et al.(2010)] [Santoro et al.(2011)] ont déjà donnés lieu à des publications dans ce domaine et l'étude des réseaux sociaux a su s'imposer comme une science utile et nécessaire en sociologie, biologie ou encore en informatique.

Les algorithmes de détection de communautés posent le problème de la robustesse. En effet, les différents algorithmes offrent des résultats très différents [Cazabet et al.(2012)]. Ce problème vient soit de la nature des approches, soit de leurs aspects non déterministes, qui font que pour un même algorithme, deux détections de communautés fourniront des résultats pouvant être très différents. De plus, beaucoup travaillent sur des approches statiques, c'est à dire ne considèrent pas que l'ordre d'apparition des liens à une importance. Pour améliorer ces deux critères, deux solutions s'offrent à

nous.

La première est d'améliorer les algorithmes de détection de communautés. Cette solution nécessite une modification de l'implémentation de ceux ci, mais aussi parfois, de leur fonctionnement. Il faut donc repenser de nouveaux algorithmes adaptés à l'analyse de réseaux dynamiques [Cazabet et al.(2011)]. Cette solution nécessite donc une forte ingénierie logicielle et présuppose d'abandonner ou d'adapter les outils précédemment développés.

La seconde est de faire un pré-traitement des graphes pour augmenter la cohésion des communautés en prenant en compte leur dynamique. En modifiant la structure du graphe, sans pour autant l'altérer, c'est à dire en renforçant les structures communautaires déjà existantes, on pourrait alors faciliter le processus de détection de communautés en augmentant la définition de celles-ci. Cette approche permettrait alors de continuer d'utiliser les outils développés pour une analyse statique sur un graphe modifié par sa dynamique.

L'objectif de ce stage est de concevoir une méthode permettant la prise en compte de la dynamique des réseaux sociaux et grâce à cette dernière, d'extraire des informations afin de renforcer les structures communautaires au sein du réseau. La méthodologie devra donc permettre d'inscrire dans un graphe des informations sur sa dynamique afin de permettre une meilleur détection des communautés par les algorithmes classiques et mettre en évidence que la dynamique est une source d'information et qu'elle doit donc être prise en compte dans une analyse de réseau social.

Chapitre 1

État de l'art

1.1 Les évolutions d'un monde numérique : vers une science des données

C'est maintenant une chose sûre, internet a révolutionné la manière dont nous communiquons. De cette facilité naît une profusion d'informations utiles, issues de ces interactions. L'engouement pour les réseaux sociaux tels que Facebook ou Twitter est révélateur de ce changement. La figure 1.1 nous montre que le temps passé sur les réseaux sociaux aux USA entre 2007 et 2011 a considérablement augmenté, mais ce résultat n'est pas propre aux USA puisque l'ont retrouve le même phénomène sur toutes les régions du monde. Pour preuve, selon un sondage IFOP du 15 novembre 2011, en France, 74,3% de la population des 18 ans et plus sont quotidiennement connectés à internet et 77% des internautes se déclarent membre d'au moins un réseau social. Chaque jour, ce sont des millions de personnes qui communiquent, échangent et interagissent offrant un amas de données si conséquent qu'il nous faut en permanence développer de nouveaux outils pour en permettre l'exploitation.

L'étude de ces données complexes n'est pas l'apanage d'une unique science mais le point de convergence de nombreux domaines tels que la physique, la biologie, la géographie ou encore l'informatique. Pour chacune de ces sciences, la nécessité d'analyser et traiter des quantités importantes de données en un temps limité se révèle devenir une véritable nécessité. Aussi l'étude de données complexes est source de synthèse et d'émergence d'outils d'analyse

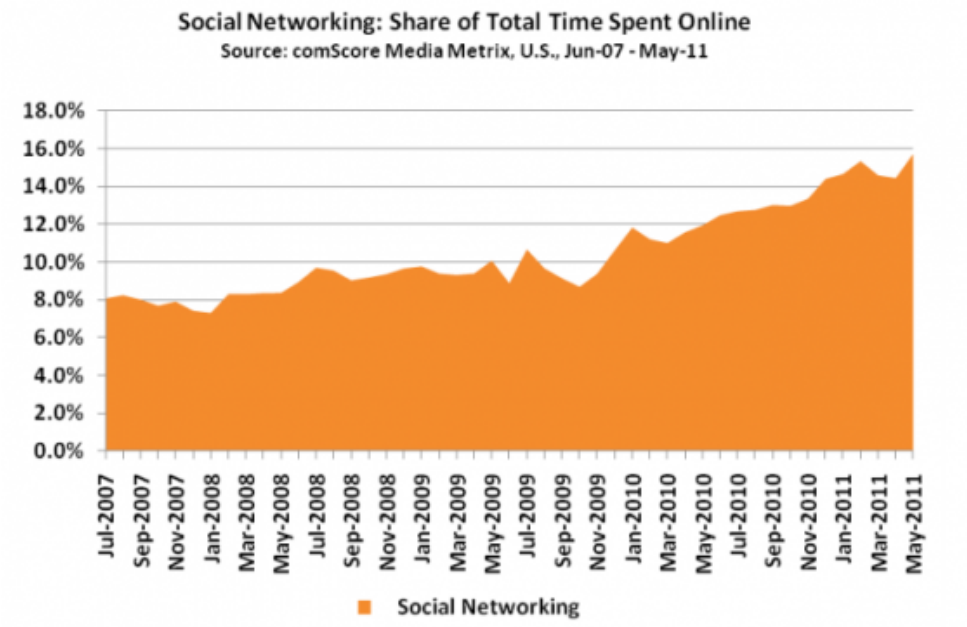

FIGURE 1.1 – Augmentation du temps passé sur les réseaux sociaux aux USA depuis 2007

issus de ces différents domaines.

Un de ces outils, l'analyse de réseaux sociaux, est un exemple de cette convergence des sciences.

1.2 Les réseaux sociaux

L'analyse de réseaux sociaux est étudiée depuis les années 30. Originellement développée pour la sociologie, cette science s'est diversifiée pour s'imposer comme une science utile et nécessaire dans de nombreux domaines tels que la biologie, la géographie ou encore les technologies de l'information. Là où les approches traditionnelles de la sociologie s'intéressaient aux caractéristiques individuelles des individus, l'analyse des réseaux sociaux se propose d'étudier les relations entre acteurs au sein du réseau.

L'analyse de réseaux sociaux se prête à une variété de réseaux allant du réseau d'amitié [Zachary(1977)](Figure 1.2), des réseaux de co-apparition [Knuth(1993)] ou encore d'interactions entre animaux [Lusseau et al.(2003)].

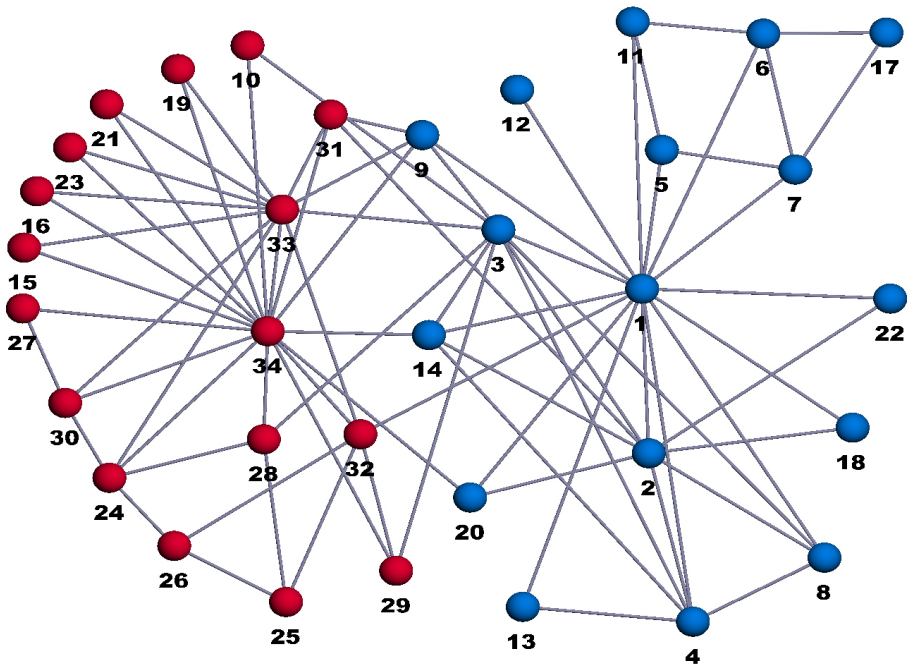

FIGURE 1.2 – Un exemple de réseau social : Réseau d'amitié du Zachary Karate Club

Là où un ensemble d'acteurs interagissent, une modélisation de ces interactions sous la forme d'un réseau social est possible, ce qui en fait un outil pluridisciplinaire.

L'analyse de réseaux sociaux s'appuie sur les acquis de la théorie des graphes [Scott(1988)] pour formaliser le réseau social comme un ensemble de nœuds et de liens où chaque nœud modélise un acteur et chaque lien une relation entre deux acteurs. Une valeur peut être affectée à un lien et représentera alors la force de celui-ci. Elle peut servir à représenter l'importance d'une relation, que ce soit en comptant simplement le nombre d'occurrences de cette relation, ou en prenant en compte d'autres processus de pondération (qualité de l'interaction, système d'évaluation, ...). A ce titre, l'analyse de réseaux sociaux dispose d'outils mathématiques issus directement de la théorie des graphes mais aussi d'outils et de techniques qui lui sont propres. On retrouvera donc dans l'analyse de réseaux sociaux des terminologies de la théorie des graphes telles que le degré, la force, ou le poids d'un lien. Elle est une science à part entière et hérite du formalisme de la théorie des graphes.

1.3 Systèmes multi-agent et simulation sociale

La simulation sociale tend à reproduire ou analyser le comportement d'un ensemble d'individus. Les individus évoluent dans un environnement spécifié et chaque individus disposent d'un ensemble de caractéristiques qu'il peut, ou non, avoir en commun avec les autres individus. Il répond à un ensemble de règles, qu'elles soient fixées par le système (contraintes d'environnement), ou sociales (contraintes comportementales). L'approche multi-agent semble alors la plus adaptée à ce type de pratique. Un système multi-agent(SMA) est un système composé d'un ensemble d'agents, situés dans un certain environnement, et interagissant selon certaines relations. L'approche multi-agent permet alors de définir un ensemble d'agents, plus ou moins autonomes, ayant la capacité d'interagir ensemble et permet donc de simuler l'évolution du système.

Depuis quelques temps, de nombreux modèles multi-agent sont développés pour modéliser le vivant, notamment pour l'étude de comportements sociétaires [Thomas et al.(2002)].

Plus récemment, ce type d'approche a été utilisé dans le cadre des réseaux sociaux pour de la détection de communautés [Cazabet et al.(2011)] où le réseau social est considéré comme l'environnement, chaque nœud étant un acteur, et les liens, la marque de leur interaction. Par exemple, dans [Cazabet et al.(2011)], les acteurs participent à la décision d'intégrer un nouvel individu aux communautés dont ils font partis en tenant compte de leurs perceptions, ici leur voisinage. Ces approches montrent des résultats intéressants, permettent la simulation du comportement d'un grand nombre d'acteurs et offrent l'avantage d'un système ouvert, ou un individu peut facilement être ajouté ou retiré sans nécessiter la refonte de l'analyse ou de la méthodologie.

1.4 Analyse de communautés

L'analyse de réseaux sociaux permet une approche sous deux aspects, l'une individus-centrée, l'autre s'intéressant aux structures sociétaires.

L'approche individus-centrée propose d'étudier la manière dont les indi-

vidus établissent des relations. On s'intéresse alors à comprendre et à modéliser, à l'aide de lois de distribution ou encore d'approches probabilistes, le comportement des acteurs. A l'aide de ces lois de distribution et propriétés, on peut alors simuler le comportement de groupes d'acteurs. Cette analyse repose sur des outils mathématiques et statistiques et s'appuie fortement sur la parenté avec la théorie des graphes pour formaliser les règles d'interactions qui prennent place dans un réseau social.

Ce type d'approche offre des résultats intéressants mais pose des problèmes sur des systèmes où des stratégies locales entrent en jeu. Elle présuppose qu'une majorité des acteurs répondent aux même règles d'interactions. Cet axiome fort ne semble pas adapté pour étudier des systèmes réels où chaque acteur établi sa propre stratégie d'interaction. Par exemple, il est possible qu'un acteur au sein d'un système sélectionne les personnes avec qui il interagit selon un critère de décision qui lui est propre (affection, renommée, ...) et par conséquent, qu'il ne suive pas une loi mathématique. Aussi l'approche individus-centrée pose des problèmes d'interprétation et de plausibilité.

La seconde approche tend elle à s'intéresser aux structures sociétaires, c'est à dire, à la manière dont des groupes d'individus vont interagir et à la détection de ces groupes. Ce procédé se nomme analyse de communautés. L'existence de zones plus densément connectées constitue une caractéristique que l'on retrouve dans de nombreux cas. L'analyse de communautés tend à rendre capable la détection de ces zones, nommées communautés. Elle utilise les propriétés topologiques du réseau, c'est à dire sa structure, pour déterminer si une structure communautaire existe.

La détection de communautés a des applications diverses, souvent dépendantes de la nature du réseau. Il n'existe pas de définition formelle et unanime de ce qu'est une communauté et le terme "communauté" n'a de sens que celui que l'on lui prête. Il faut souvent contextualiser le domaine d'analyse afin d'en comprendre le sens.

Dans le cadre de l'analyse de réseaux sociaux, une communauté peut se voir comme un groupe d'individus qui ont tendance à plus agir ensemble qu'avec les autres. Par exemple, une communauté dans un graphe d'amitié désignera un groupe d'amis ou dans un graphe de collaboration, un groupe de

travail. La manière dont on détermine si un groupe est une communauté est fortement dépendante de l'algorithme utilisé, ce qui pose donc le problème de la robustesse. En effet, de par leurs approches, différents algorithmes vont fournir des résultats très différents [Cazabet et al.(2012)]. Un même algorithme peut d'ailleurs lui même, de par son approche non déterministe, à chaque exécution, fournir des résultats différents. Ces différences de résultats rendent difficile la comparaison des communautés détectées et il faut analyser les résultats des différents algorithmes pour déterminer ceux qui semblent les plus probants. Cette sensibilité à l'angle d'approche et au déterminisme explique aussi la profusion d'algorithmes qui ont été développés.

Cependant, plus une structure communautaire est cohésive, c'est à dire bien distincte dans le réseau, plus elle sera bien détectée par les différents algorithmes. On pourra alors retrouver des communautés similaires détectées par chaque algorithmes. Ainsi, augmenter la cohésion des structures communautaires apparait naturellement comme une approche pertinente pour augmenter la qualité des communautés détectées. Cependant, l'augmentation de cette cohésion doit se faire dans le respect de la topologie et ne doit pas altérer la structure du réseau et détecter des communautés qui n'auront pas de sens.

Les enjeux de la détection de communauté sont majeurs, puisqu'ils permettent de favoriser la compréhension des processus d'interaction entre acteurs et trouvent des intérêts en économie, sociologie, biologie ou informatique. Elle sera par exemple en géographie utilisée pour identifier des zones densément peuplées ou en biologie pour étudier le comportement sociétaire de groupes d'animaux. Elle devient même, à l'heure de l'avènement des réseaux sociaux, un enjeu majeur du marketing et de l'économie et est, en ce sens, un domaine porteur dans la recherche socio-économique.

Parmi tous les algorithmes de détection de communautés, on distingue deux familles : ceux qui ont une approche statique et ceux qui ont une approche dynamique.

1.4.1 Algorithmes statiques

L'approche statique considère le réseau social comme un tout et ne s'intéresse pas aux processus qui ont amenés à sa création. Elle considère que l'ordre dans lequel les liens apparaissent ne revêt pas d'importance dans la compréhension du réseau. C'est cette approche qui a donné la plus grande profusion d'algorithmes tels que [Blondel et al.(2008)], [Lancichinetti et al.(2011)], ou [Derényi et al.(2005)], avec, pour ces deux dernières méthodes, la particularité de gérer le recouvrement, c'est à dire, qu'un nœud appartienne à deux communautés distinctes. L'approche statique s'intéresse surtout à l'état actuel d'un réseau et aux structures communautaires qui tiennent actuellement place.

1.4.2 Algorithmes dynamiques

L'approche dynamique quant à elle considère l'ordre d'apparition des liens comme une source d'information facilitant le processus de classification en communautés. Sous cette approche, l'état courant d'un réseau est dépendant de l'évolution des interactions et détecter une communauté sans prendre compte de l'ordre dans lequel les liens sont apparus est du non sens. Par exemple, un lien peut avoir été fortement actif au début du réseau et avoir cesser son activité bien avant la fin de sa construction. Cette cessation d'activité fait qu'il est difficilement classable dans des communautés où les liens ont été plus fortement actifs en fin de période. Aussi, la prise en compte de ce lien dans des communautés où tout les liens sont récents peut poser problème. On citera notamment les travaux de [Cazabet et al.(2011)] et de leur algorithme ILCD qui prend en compte la dynamique pour construire des communautés.

1.5 De l'intérêt de la dynamique

La plus part des outils développés pour l'analyse de réseaux sociaux adoptent une approche statique. Ces outils ont maintenant montré leur intérêt dans l'analyse de réseaux. Pourtant, ils ne semblent plus adaptés à l'étude

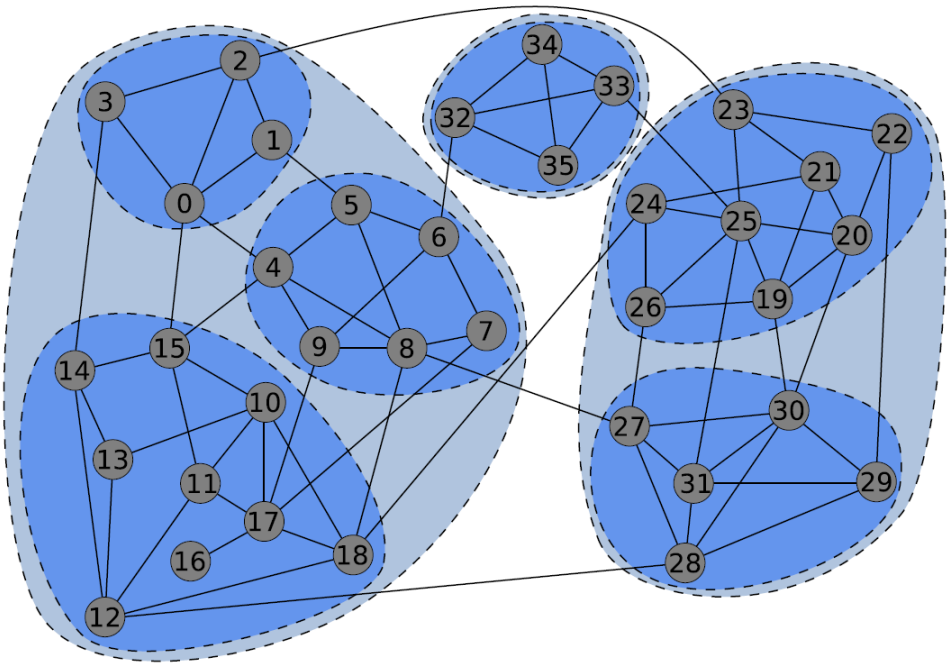

FIGURE 1.3 – Exemple de structures de communautés dans un graphe : deux partitions en communauté correspondant à deux échelles différentes sont représentées.

des données réelles, où la dynamique apparait être devenue trop importante pour être négligée.

La prise en compte de la dynamique naît d'abord d'un besoin d'étudier de nouveaux types de données. Là où il fallait précédemment réaliser de nombreux sondages, collecter et synthétiser les données, l'outil numérique nous permet d'acquérir quantités d'informations et ce, rapidement. De nos jours, il est courant de travailler sur des données collectées sur plus de 20 ans. Sur des périodes aussi longues, il est logique que les dynamiques d'interaction n'aient pas toujours été les mêmes. Les stratégies d'interaction ayant eu cours au début de la période ne sont pas les mêmes que celles qui ont pris cours à la fin de la période. Par exemple, les dynamiques de publication d'un centre de recherche ont fortement évoluées. Cette évolution vient de la taille du centre, qui tend à augmenter, mais aussi de l'évolution des technologies qui rendent la publication plus facile et d'enjeux politiques. Aussi, les communautés qui

existaient dans les années 80 ne sont pas les mêmes que celles des années 2000. De nouveaux acteurs sont entrés dans le système, d'autre en sont sortis et de nouvelles dynamiques de publication se sont mises en place. De ce besoin de prendre en compte les différentes étapes dans l'évolution d'un réseau naît l'utilité de la dynamique, mais il n'est pas le seul.

L'objet principal qui réside dans la dynamique, et donc l'objet de ce mémoire, est que l'étude de cette dynamique peut être une nouvelle source d'information complémentaire et sa prise en compte pourrait alors apporter une nouvelle vision sur les données collectées. Son objectif n'est pas de remplacer une analyse statique mais bel et bien d'en être le complément. Ainsi, étudier les dynamiques d'un réseau c'est aussi améliorer la compréhension des résultats d'une approche statique.

Dans tous les cas, l'approche de la dynamique n'en est qu'à ses débuts et révèle d'ors et déjà bien des promesses.

Chapitre 2

Méthode

2.1 Enjeux

Nous avons vu qu'il existait une profusion d'outils et de méthodes portant sur l'analyse statique des réseaux mais que peu d'outils permettaient d'en aborder la dynamique. Nous avons vu que les résultats différaient d'un algorithme à l'autre. Il est donc intéressant de pouvoir améliorer la robustesse de ces algorithmes en proposant une méthode de pré-traitement du graphe pour renforcer sa cohésion. Récemment, des méthodes [Farkas et al.(2007)] [Radicchi et al.(2011)] ont mis en avant l'utilité, lorsque l'information est disponible, de pondérer les relations d'un réseau social. En effet, pondérer les relations, c'est à dire, mettre une valeur sur la force de cette relation, c'est permettre de les comparer, et donc, de décider laquelle est la plus importante pour un acteur. Cette prise en compte du poids permet aux algorithmes de détection de communautés d'augmenter la qualité et la robustesse de leurs résultats.

L'objectif de ce mémoire est de proposer une approche permettant de faire porter des informations sur la dynamique d'un réseau sur un graphe qui serait ensuite exploitable par les méthodes d'analyse statique. L'idée derrière cette approche est de réussir à extraire de la dynamique une information sur la force d'une interaction entre deux acteurs et ainsi mettre en avant l'axiome énonçant que la dynamique est source de nouvelles connaissances.

La méthodologie que nous présentons ici se propose d'extraire de la dy-

namique une information sur la force des liens, afin de créer un réseau où les relations sont pondérées, et d'utiliser cette information pour améliorer les résultats des algorithmes de détection de communautés. Le second objectif est de mettre en avant la source d'information que peut être la dynamique d'un réseau social.

2.2 Pondérer les relations grâce à la dynamique

Fournir une information sur la qualité d'une interaction, être capable de la pondérer, c'est se rendre capable de différencier une interaction forte d'une interaction faible. Faire cette différenciation c'est permettre, lorsque la nécessité se présente de réaliser des césures sur les liens les plus faibles et donc de conserver et favoriser les liens forts. [Farkas et al.(2007)] ont montré que pondérer les relations permettait d'augmenter la qualité des communautés découvertes. L'idée ici est d'utiliser la dynamique du réseau comme seule information pour quantifier les interactions entre acteurs et créer un nouveau réseau social pondéré.

Cette section se propose donc d'expliciter la méthodologie que nous avons développée pour permettre une telle pondération. Elle définit l'ensemble des outils et techniques issus de l'étude de la dynamique du réseau.

2.2.1 Intervalles, période et graphe d'intervalle

Étudier les dynamiques d'un réseau c'est étudier son existence et ses évolutions au cours du temps. Mais avant de s'intéresser à son étude, il faut d'abord porter attention à la forme que prennent les données à étudier.

La nature des données influence fortement l'analyse du réseau social. Une des caractéristiques importantes est la fréquence avec laquelle ces données nous sont fournies, c'est à dire à quel rythme les données nous sont fournies.

On définit i comme l'*intervalle* des données, c'est à dire, l'intervalle sur lequel porte un jeu de données. i_n désigne alors le i^{eme} intervalle. Ici tous les i_n sont de même longueur et i_{n+1} est successeur de i_n (Cette condition de même longueur n'est cependant pas nécessaire, mais ici, ce choix a été fait pour faciliter l'implémentation).

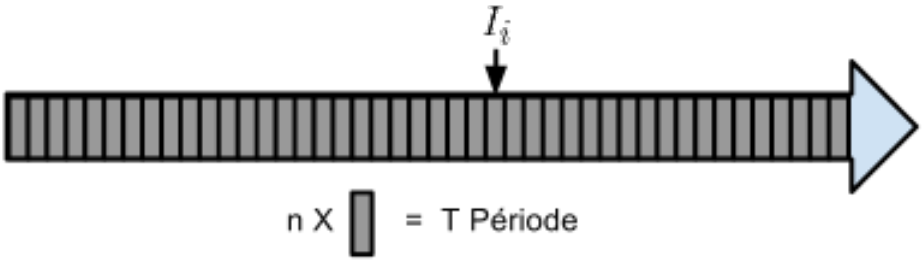

FIGURE 2.1 – Découpage de la période d'étude en intervalles
Ici, la flèche représente l'ensemble des données. Elles sont découpées en n blocs d'intervalles de longueur identique. Le ième bloc représente la ième acquisition de données. Les n blocs définissent une période d'étude, nommée T.

Soit T la *période* d'étude telle que $T = \sum_{n=0}^{N} i_n$ où N désigne le nombre d'intervalles. Ainsi, T représente l'intervalle global de notre étude. i_n peut s'exprimer en minutes, heures, années, ou toute unité temporelle.

Par exemple, si l'on traite des données collectées annuellement sur 10 ans, on aura $\|i_0\| = \|i_1\| = ... = \|i9\| = 1$ et $T = \sum_{n=0}^{9} 1 = 10$. T et i doivent être de même unité ici, l'an.

Pour chaque i_n on définit $G(i_n)$ le réseau social issu de la i^{eme} capture de données. $G(i_n)$ représente donc les interactions ayant eu cours au i^{me} intervalle. Soit $\triangleleft$ désigne l'antériorité, on a $G(0) \triangleleft G(1) \triangleleft ... \triangleleft G(n-1)$. On nomme $G(i_n)$ le *graphe d'intervalle* i_n.

A un jeu de données correspond donc N graphes d'intervalle, composés des interactions ayant pris place au cours de ce bloc de données. Étudier indépendamment chaque graphe d'intervalle permet l'étude des interactions ayant pris place localement (au sens temporel).

Il est à noter que l'aspect dynamique de la méthode n'est pas à mélanger avec une approche déterministe. Ici aucune règle n'est extraite ou utilisée pour passer du graphe $G(i_n)$ à $G(i_{n+1})$ et l'analyse de l'aspect dynamique du réseau ne s'intéresse pas à connaître l'existence ou l'approximation de telles règles mais tend à utiliser l'apparition et la disparition de liens pour rendre compte de ses dynamiques.

2.2.2 Valeur d'intérêt, définition et application

On désire pondérer les relations de notre graphe d'intervalle en se basant uniquement sur la topologie du réseau. Pour un i_n, on connait pour un acteur A le nombre de personnes avec lequel A a interagi. En se basant sur ce nombre, c'est à dire, sur son nombre de voisins, on veut être capable de quantifier chacune des relations que A a établit. Soit A et B deux acteurs tels que $A, B \in G(i_n)$. On définit $I(A, B)$ comme l'intérêt de A pour B. Sa méthode de calcul est dérivée de "l'acte de collaboration" défini dans [Farkas et al.(2007)]. Chaque acteur possède un intérêt de 1 qu'il répartit entre les différents acteurs avec qui il est impliqué. Faute d'information sur la qualité des interactions, nous considérons ici qu'un acteur partage son intérêt équitablement entre ses voisins. Cela revient à considérer qu'un acteur va équitablement partager ses compétences avec les personnes avec qui il collabore. Ce processus n'a pas pour objectif de modéliser une réalité dans la manière dont l'agent considère ses collaborateurs. En effet, puisque l'on désire montrer que l'usage seul de la dynamique permet l'extraction d'informations quantitatives, aucun processus cognitif n'intervient ici. Cependant, comme expliqué au chapitre 6 dans les perspectives, la méthode de calcul de l'intérêt pourra être modifiée pour faire intervenir un processus cognitif au sein de l'agent.

Définition On appelle ***Valeur d'intérêt*** ou ***intérêt*** la valeur $I_n(A, B)$ et $I_n(A, B) = 1/NbVoisins(A)$ où $NbVoisins(A)$ retourne le nombre de voisins de A dans $G(i_n)$. On a $I_nA, B \neq I_nB, A$ et $\exists e(A, B) \Rightarrow \exists e(B, A)$.

2.2.3 Approche Multi-agent

L'objectif est de créer un graphe d'intérêt modélisant pour chaque nœud le sentiment de proximité/d'importance d'une relation avec son voisin, basé sur la valeur d'intérêt énoncée précédemment. On désire obtenir un réseau orienté où chaque lien existant dans le graphe $G(i_n)$ entre deux agents A et B existe dans le nouveau réseau sous deux liens orientés de A vers B et de B vers A. Le lien $A \rightarrow B$ représente alors l'intérêt de A pour B. L'idée est donc de créer dynamiquement un graphe orienté, pondéré et agrégé $G(T)$ à

partir des graphes d'intervalle $G(i_n)$ tel que $G(T) = \bigcup_{n=0..N} G(i_n)$.

On introduit alors le système multi-agent suivant :

Agents Chaque nœud est considéré comme un agent.

Environnement L'environnement est un graphe dirigé où pour chaque lien,$\exists e(A, B) \Rightarrow \exists e(B, A)$. Chaque nœud perçoit son voisinage à rang 1. Il possède une mémoire temporelle lui permettant de connaître les liens actifs sur un intervalle temporel $\delta t < T$. Il a donc une connaissance du passé restreinte.

Dynamiques Le système permet l'ajout de nœuds et de liens. L'ajout de nœuds et de liens sont indépendants du comportement des agents. Le système est régit par une horloge globale permettant la gestion des cycles. Passer de i_n à i_{n+1} incrémente l'horloge de 1. Entre chaque tour d'horloge, le système charge les nouvelles interactions. C'est à dire, si l'on prend comme horloge les années de 1990 à 2000 avec un pas de 1 an, le système chargera les données de 1990, puis les suivantes, et ce, année par année. Cette dynamique d'horloge dépend de la fréquence des données, c'est à dire de l'intervalle entre deux jeux de données.

Actions L'agent peut pondérer un lien en fonction de ses perceptions et de sa mémoire. Un agent A qui collabore avec un agent B fixera le poids de sa relation à $I(A, B)$. L'agent peut demander à un de ses voisins la liste de ses voisins.

Pour chaque $G(i_n)$, les agents pondèrent les relations sortantes grâce à la valeur d'intérêt selon l'algorithme 1. On obtient alors un graphe $G(I)$ où chaque lien est dirigé et pondéré.

L'image 2.4 montre un exemple de pondération d'un lien entre deux agents A et B. Deux nouveaux liens dirigés sont construits et pondérés selon le voisinage de A et B.

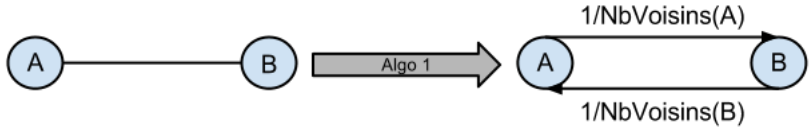

FIGURE 2.2 – Principe de pondération des liens

Algorithm 1 Pondération des liens sortants d'un nœud par un agent

1: **for all** Node $A \in G(i_n)$ **do**
2: **for all** Edge $e \in OUT(A)$ **do**
3: $W_{G(i_n)}(e) \leftarrow 1 \div NbVoisins(A)$ ▷ Ici, $NbVoisins(A)$ renvoie les voisins du nœud A, c'est à dire l'ensemble des liens actifs dans $G(i_n)$
4: **end for**
5: **end for**

Définition On appelle ***Graphe sur l'intervalle n d'intérêt ou*** $G(I_n)$ le graphe obtenu par cette méthode.

2.2.4 Graphe agrégé

Puisque l'on s'intéresse à l'étude de la dynamique sur T, il nous faut extraire des $G^2(i_n)$ un graphe reflétant l'intégralité des interactions ayant eu cours sur T.

Soit $G(T)$ la concaténation (au sens topologique) des $G(i_n)$, c'est à dire l'ajout des nœuds et liens de $G(i_n)$ dans $G(T)$. $G(T)$ est appelé *graphe agrégé*. Chaque $G(i_n)$ est chronologiquement ajouté à $G(T)$. Le graphe agrégé est créé selon l'algorithme 2. Il est donc le résultat d'une succession de *fusions* des graphes d'intervalle et donc, résultat d'une succession d'interactions. Il modélise l'intégralité des relations ayant eu cours sur la période T. Le poids d'un lien, définit par la fonction $W(e)$ où e est un lien, est définit comme $W_{G(T)}(A, B) = \sum_{n=0}^{N} I_{G(i_n)}(A, B), \forall i_n \in T$ où A, B sont deux nœuds et n le nombre de graphes d'intervalles. Ainsi, un lien de $G(T)$ est la somme des intérêts pour chaque $G(i_n)$.

Définition On appelle ***Graphe d'intérêt*** le graphe obtenu par l'agrégation des graphes d'intervalle d'intérêt.

Algorithm 2 Fusion

1: **for all** Node $n \in G(i_n)$ **do**
2: **if** $n \notin G(T)$ **then**
3: $G(T).addNode(n)$
4: **end if**
5: **end for**
6: **for all** Edge $e \in G(i_n)$ **do**
7: **if** $e \notin G(T)$ **then**
8: $G(T).addDirectedEdge(s,t)$
9: $W_{G(T)}(s,t) \leftarrow I_{G(i_n)}(s,t)$
10: **else**
11: $W_{G(T)}(s,t) \leftarrow W_{G(T)}(s,t) + I_{G(i_n)}(s,t)$
12: **end if**
13: **end for**

2.2.5 Vers de nouveaux outils

Grâce au processus de pondération des relations, on est amené à reconsidérer certaines valeurs dont le sens et l'intérêt dans le cadre de cette méthodologie sont amenés à évoluer. C'est le cas de la *force* d'un nœud. La force d'un nœud est définie comme la somme de ses liens, c'est à dire, la somme du poids des liens qui entrent ou partent du nœud. On distingue ici deux types de forces d'un nœud : la force entrante définie comme la somme du poids des liens entrants, et la force sortante, définie comme la somme du poids des liens sortants.

Ici, la force nous fournit un pseudo-indicateur de l'intérêt généré pour ou par une personne. La force entrante représente le niveau d'intérêt que l'agent génère pour les autres acteurs, alors que la force sortante nous fournit le niveau d'intérêt pour les autres acteurs. On distingue donc ici la réputation de l'acteur par sa force entrante, de son implication pour le système par la force sortante.

Ainsi, un acteur avec une forte force entrante, mais une faible force sortante, est dissociable d'un acteur avec une faible force entrante mais une forte force sortante. Le premier indiquera alors un acteur sollicité dans le réseau,

lorsque le second indiquera un acteur sollicitant.

Les applications et l'interprétation de ces valeurs ne prend alors sens que dans le cadre de l'analyse et dépendent a la fois du contexte et de la structure du graphe mais peuvent devenir de nouveaux outils pour sa compréhension.

2.2.6 Analyse de la méthode de pondération

Nous avons exposé une méthode avec approche multi-agent permettant la construction dynamique d'un graphe d'intérêt $G(T)$, quantifiant pour chaque acteur les relations qu'ils entretiennent. Grâce à ce graphe d'intérêt, on peut alors réaliser une analyse à l'aide des outils classiques de l'analyse de réseaux et notamment réaliser une détection de communautés sur graphe dirigé et pondéré.

L'avantage de cette méthode est son aspect uniquement topologique. Elle met donc en évidence que la simple prise en compte de la modification de la topologie du réseau social est source d'informations pour quantifier les interactions.

Les résultats de [Farkas et al.(2007)] et les résultats expérimentaux (voir chapitre 3 et 4), tendent à montrer que la prise en compte du poids des liens dans le processus de détection de communauté améliore la qualité des communautés détectées.

Cependant, d'autres utilités peuvent être trouvées à ce graphe d'intérêt. Nous allons maintenant nous intéresser au phénomène d'"intérêt transitif" et à la mise en relation par transitivité d'acteurs.

2.3 Intérêt transitif

2.3.1 Pourquoi la transitivité : prédiction de liens, liens manquants, résilience

Un des domaines de l'analyse de réseaux est tout ce qui a trait à la prédiction de liens, la détection des liens manquants et l'étude de la résilience du réseau. L'utilité de ces études n'est plus à démontrer. Être capable d'anticiper l'évolution du réseau, les problèmes de données manquantes, ou encore,

renforcer son intégrité structurelle face à la disparition d'acteurs ou de liens est d'une utilité certaine. Sans être spécifique à ce domaine, la méthodologie et les outils traités dans cette section peuvent être utiles.

2.3.2 Détection de liens transitifs et filtrage de liens faibles

Un cas d'exemple : le cas des stagiaires Soit un acteur A qui accueille la même année deux stagiaires B et C sur deux thématiques similaires mais indépendantes. Le travail de A et B donne lieu à la rédaction d'un mémoire. Si on considère le graphe de collaboration, il existe donc une relation entre A et B ; de même pour le travail de A avec C. Considérons donc le graphe obtenu par ces trois acteurs. On a une relation entre A et B et entre A et C mais qu'en est-il de la relation entre B et C. Malgré l'absence de publications entre eux, n'existe-t-il pas un intérêt entre ces deux personnes, et par extension, un sentiment d'appartenir à la même communauté ? N'ont-il pas, au travers de leur responsable, communiqué et interagis ?

En créant le graphe d'intérêt, B et C développent un intérêt de 1 pour A et A développe un intérêt de 0.5 pour B et C. Le graphe d'intérêt met donc en évidence l'existence d'un chemin entre B et C. L'idée est donc de modéliser cette transitivité en créant un lien entre B et C de poids $1 * 0.5$ et réciproquement pour C vers B. En ajoutant ce lien, on créé une clique entre A,B,C renforçant structurellement l'idée d'une communauté entre A, B et C. De plus, on fournit une indication sur la "force" de ce "sentiment" grâce à la valeur d'intérêt rendant possible de déterminer la proximité d'un nœud par rapport à une communauté.

Enfin, si A vient à disparaître, l'ajout de ce lien transitif permettra à B et C de rester en interaction, renforçant ainsi la résilience du réseau. En effet, puisqu'il existait auparavant une relation implicite entre B et C, la matérialiser en rajoutant un lien physique entre B et C permet de renforcer la solidité du réseau et, si A venait à disparaitre, laisse la possibilité à B et à C de conserver leur valeur d'intérêt.

On propose alors de formuler cette transitivité :

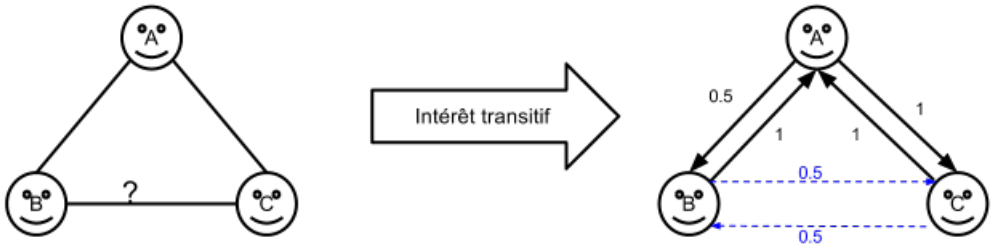

FIGURE 2.3 – Intérêt transitif

Définition de l'intérêt transitif Soit $s, t \in Nodes(G(T))$, si $\exists p \in Nodes(G(T))$ tel que $p \in Out(S) \bigcap In(t)$ alors $\exists e \in Edges(G(T))$ tel que e est un lien de s vers p et que son poids $W_{G(T)}(s,t) = W_{G(T)}(s,p) * W_{G(T)}(p,t)$. De manière étendue, on a, $\forall s, t \in Nodes(G(T))$, $\forall p \in Out(S) \bigcap In(t)$, alors $\exists e \in Edges(G(T))$ tel que e est un lien de s vers p et que son poids $W_{G(T)}(s,t) = \sum_{\forall p} W_{G(T)}(s,p) * W_{G(T)}(p,t)$.

L'application de ce principe permet l'ajout de tous les liens de transitivité mais tous ne sont pas forts (c'est à dire, tous ces poids n'ont pas une valeur "élevée" par rapport à un seuil fixé). On propose donc de ne retenir que les liens forts afin de ne garder que les liens (transitifs ou non) les plus significatifs pour ne pas obtenir un graphe complet.

Définition On appelle ***Filtrage*** le procédé qui, pour un *seuil* fixé, tend à supprimer tous les liens faibles (inférieurs stricts au seuil) , conformément à l'algorithme 3

Le choix du *seuil* est dépendant du graphe et doit donc être choisi en conséquence. Une approche simple peut être de le choisir autour de la moyenne ou de l'écart type des valeurs d'intérêt.

2.3.3 Le problème des "Hubs"

Il existe au sein des réseaux sociaux une fraction de nœuds dont le degré est significativement plus grand que la moyenne. Ces nœuds "hyper-

Algorithm 3 Filtrage

```
1: for all Edge e ∈ G do
2:     if W(e) < seuil then
3:         G.remove(e)
4:     end if
5: end for
6: for all Node n ∈ G do
7:     if Degree(n) = 0 then
8:         G.remove(n)
9:     end if
10: end for
```

connectés" sont appelés "Hubs". Soit A un "hub" et B un nœud connecté à A. De par sa nature, A est connecté avec beaucoup de nœuds, donc par intérêt transitif, B sera connecté avec beaucoup de nœuds. L'hyper-connectivité des "Hubs" pose le problème des liens *significatifs*. Connecter B avec tous les voisins de A n'est évidemment pas une solution. Il faudrait être capable d'identifier quels voisins de A sont suffisamment *significatifs* pour connecter B. De même, il faudrait être capable de déterminer pour B si sa relation avec A est plus *importante* que celle avec ses autres voisins.

Nous avons vu précédemment que le filtrage permet de ne retenir que les liens forts, mais ce filtrage ne permet pas d'exclure les "Hubs" du processus de transitivité. Il faut donc être capable de déterminer localement si un lien est significatif pour l'acteur par rapport à ses autres liens. Ainsi, un "Hub" aura tendance à avoir beaucoup de liens, mais peu seront significatifs (ils auront tous des valeurs faibles), contrairement à un non "Hub". Pour cela, nous introduisons le concept énoncé par [Glattfelder et Battiston(2009)], les fractions de contrôle.

2.3.4 Fraction de contrôle et liens transitifs

Glattfelder dans [Glattfelder et Battiston(2009)] définit la fraction de contrôle H_{ij} d'un lien par $H_{ij} = W_{ij}^2 / \sum_{l=1}^{k_j^{in}} W_{ij}^2$.

Cette valeur, qui correspond à une normalisation des poids, a l'avantage

Algorithm 4 Transitivité d'interêt avec *seuil* fixé

```
1: for all Node A ∈ G(i_n) do
2:     for all Edge e where e.source == A do
3:         if P_A(e) > seuil then
4:             for all Edge e2 where e2.source == e.target do
5:                 if P_{e.target}(e2) > seuil then
6:                     CreerLienTransitif(A, e2.target)
7:                 end if
8:             end for
9:         end if
10:    end for
11: end for
```

de favoriser les liens forts par rapport aux liens faibles. En s'intéressant ensuite au *pourcentage de contrôle* P que représente un lien, c'est à dire, le pourcentage que cette valeur représente par rapport à la somme des liens, on peut donc déterminer si, localement, un lien est significatif.

En fixant alors une valeur de seuil de pourcentage, on se dote alors de la capacité de localement définir si un lien est important pour un agent. On introduit alors ce procédé pour corriger la méthode d'intérêt transitif et obtenir l'algorithme 4.

2.3.5 Cohérence locale et cohérence étendue

L'intérêt transitif nous permet de mettre en relation des acteurs qui, pour des raisons diverses, n'ont pas interagis mais s'influencent mutuellement. On désire mettre ce principe en application pour augmenter la *cohérence temporelle* des communautés détectées. Pour ce faire, on se doit de distinguer deux niveaux de cohérences temporelles, et par extension, deux types de communautés cohérentes temporellement.

Cohérence locale Le premier type de cohérence que l'on souhaite appliquer est la *cohérence locale*. On appelle ***Cohérence Locale*** la mise en relation d'acteurs contemporains, c'est à dire d'acteurs actifs sur un même

intervalle mais qui n'ont pas interagis ensemble.

Ici, il s'agit d'identifier des acteurs qui ont eu la possibilité d'interagir (puisque que contemporains) mais ne l'ont pas fait. Augmenter la cohérence locale d'une communauté, c'est regrouper des acteurs actifs sur la même période. On cherche ici à détecté des communautés d'acteurs contemporains, c'est à dire des communautés où les membres on tous été actifs à la même période. Considérons par exemple deux acteurs de cinéma A et B. A est un acteur des années 50, décédé en 80 et B est un acteur des années 90 toujours vivant. Avec la cohérence locale, on désire rapprocher B d'une communauté d'acteurs des années 90 et A d'une communauté des années 50. Les communautés obtenues reflèteraient alors une réalité temporelle et non plus seulement topologique. Les membre d'une communauté on tous été actifs à la même période. Elle permet d'étudier l'existence de groupes communautaires contemporains, au sens d'actifs en même temps et le terme communauté prend alors le sens de groupe d'acteurs ayant mutuellement plus agis ensemble qu'avec le reste de la communauté sur une période donnée.

On est amené à reconsidérer les méthodes $In()$ et $Out()$. Dans le cadre de la cohérence locale, à l'instant i, $In(node)$ et $Out(node)$ désignent les voisins (respectivement entrants et sortants) actifs sur l'intervalle $[i_j; i_n]$ où $j < n$. Les valeurs n et j nous permettent de définir une fenêtre temporelle de cohérence, dont la taille définit si deux acteurs sont contemporains.

Cohérence étendue Le second type de cohérence est dite *étendue* et se propose de regrouper des acteurs non contemporains, c'est à dire, non actifs durant la même période. Il s'agit ici d'identifier des acteurs qui n'ont pas eu la possibilité d'interagir mais qui pour autant ont par transitivité travaillé avec les mêmes personnes. En reprenant l'exemple précédent, où A et B sont acteurs de deux périodes différentes, on désire étudier l'existence de communautés persistantes. Est-ce que la communauté formée par les acteurs des années 50 a continuée d'exister pour intégrer celle des années 90 ou, à contrario, s'agit-il de deux communautés complètement distinctes. Aussi, en mettant en relation des acteurs qui n'ont pas été contemporains, on permet à des communautés de perdurer dans le temps. L'idée est donc de rendre plus persistantes les communautés et mettant en relation des acteurs non

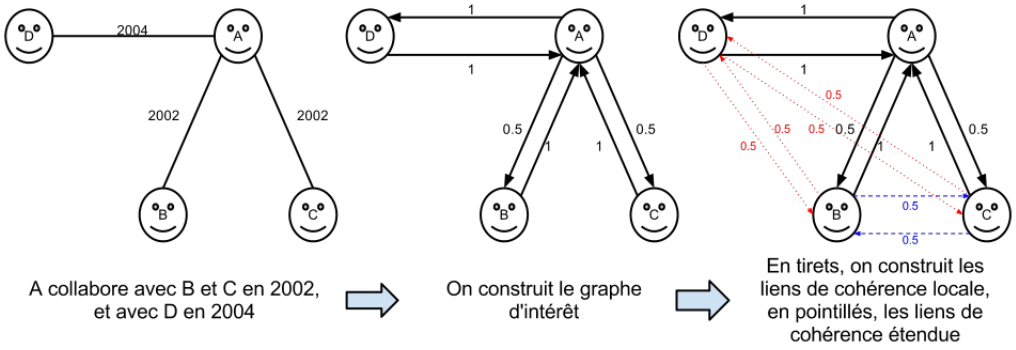

FIGURE 2.4 – Exemple de cohérence locale et étendue

contemporains, mais partageant un intérêt fort.

Dans le cadre de la cohérence locale, à l'instant i, $In(node)$ et $Out(node)$ désignent les voisins (respectivement entrants et sortants) actifs sur l'intervalle $[i_0; i_j]$ où i_0 désigne G_{i_0}.

2.4 Synthèse de la méthode

Nous proposons ici une méthode basée sur la dynamique du réseau, c'est à dire, ses évolutions topologiques, pour pondérer les relations entre acteurs. Grâce à la valeur d'intérêt obtenue, nous pouvons mettre en relation, par transitivité d'intérêt, des acteurs qui n'avaient pas établi de relation. Cette transitivité nous permet alors, soit de favoriser les connections entre acteurs contemporains, c'est à dire, actifs sur la même période, soit entre acteurs non contemporains. Ainsi elle nous rend capable d'étudier deux comportements communautaires distincts. Le premier, avec la cohérence locale, étudie les communautés d'acteurs actifs à une même période. Le second, avec la cohérence étendue, étudie les communautés persistantes dans le temps.

L'intérêt est qu'elle met en avant l'utilité de la dynamique comme source d'information sur la qualité des liens mais aussi qu'elle permet de favoriser la cohérence temporelle. [Wang et al.(2011)] ont montré que l'ajout de liens est préférable à leur suppression, car la suppression d'un lien à un effet plus perturbant pour le réseau et, par conséquent, il est plus intéressant d'en

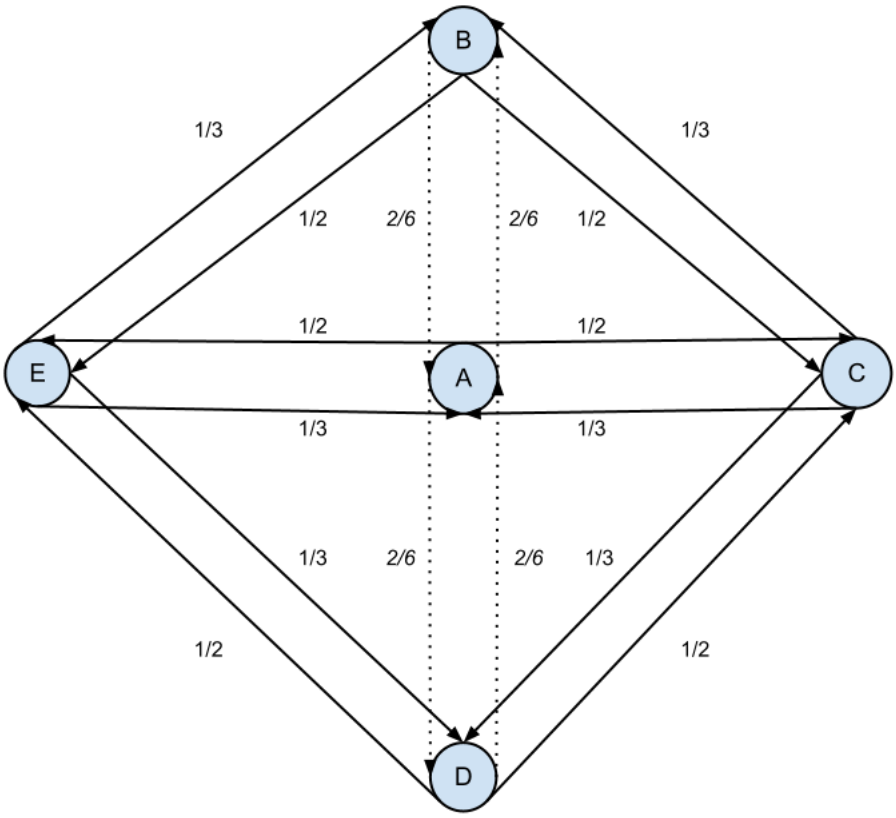

FIGURE 2.5 – Un exemple de pondération avec intérêt transitif

ajouter, confirmant ainsi que notre volonté d'ajouter des liens transitifs ne peut qu'améliorer les résultats de nos détections de communautés.

On se propose donc de mettre cette méthode en pratique afin d'étudier son influence sur la détection de communautés.

Chapitre 3

Données

3.1 IMDB

L'*Internet Movie DataBase* [IMDB] est une base de donnée en ligne sur le cinéma mondial et la télévision. Elle regroupe un grand nombre d'informations concernant films, acteurs, réalisateurs, et toute personne intervenant dans la création d'un film, d'un téléfilm ou d'une série télévisée. A l'aide de cette base de donnée, on peut extraire un graphe de co-apparition où les nœuds sont des acteurs et deux acteurs A et B sont en interaction si et seulement si ils ont joué au moins une fois dans un même film. Deux versions de ce graphe ont été générées :

Version 1 : Ont été retenus uniquement les films comptabilisant plus de 1000 votes par les visiteurs, réalisés après 1980, où les acteurs apparaissent au moins 10 fois et les films présentent au moins 2 acteurs, ce qui représente 1019 nœuds et 19377 liens.

Version 2 : Ont été retenus uniquement les films comptabilisant plus de 100000 votes et réalisés après 1980, soit 14194 nœuds pour 1070266 liens.

3.2 Collaborations au sein de l'IRIT

L'IRIT (Institut de Recherche en Informatique de Toulouse) gère une base de données recensant toutes ses publications depuis les années 1970 soit 3726 publications en Décembre 2011. On extrait de cette base le graphe de collaboration où un auteur est représenté par un nœuds et un lien entre deux acteurs existent si et seulement si ces deux acteurs ont publiés ensemble. Le graphe ainsi obtenu se compose de 2571 auteurs pour 18990 collaborations.

Chapitre 4

Application de la méthode

4.1 Procédé d'expérimentation

L'implémentation et la paramétrisation de la méthode ont été réalisées sur les deux versions du graphe de co-apparition IMDB.

L'étude des effets et avantages de la méthode est réalisée sur le graphe de co-publication de l'IRIT. Afin de concentrer l'étude sur les personnels internes à l'IRIT, on filtre le graphe en ne retenant que les auteurs ayant au moins deux publications. Afin de ne conserver que les liens les plus forts, on fixe empiriquement le poids minimal d'un lien à 0.3 et la valeur minimale du pourcentage de contrôle à 10%. On s'intéressera dans un premier temps à la cohérence locale, puis, sous les mêmes conditions, à la cohérence étendue.

On se propose d'étudier 3 versions de ce graphe :

Version 1 Graphe de co-publication non pondéré sans intérêt transitif : 2571 auteurs et 18990 collaborations.

Version 2 Graphe de co-publication pondéré sans intérêt transitif : 2272 auteurs et 9779 collaborations.

Version 3 Graphe de co-publication pondéré avec intérêt transitif (cohérence locale) : 2272 auteurs et 10504 collaborations.

Dans la version 2 et 3, le filtrage des liens a diminuer le nombre de liens et

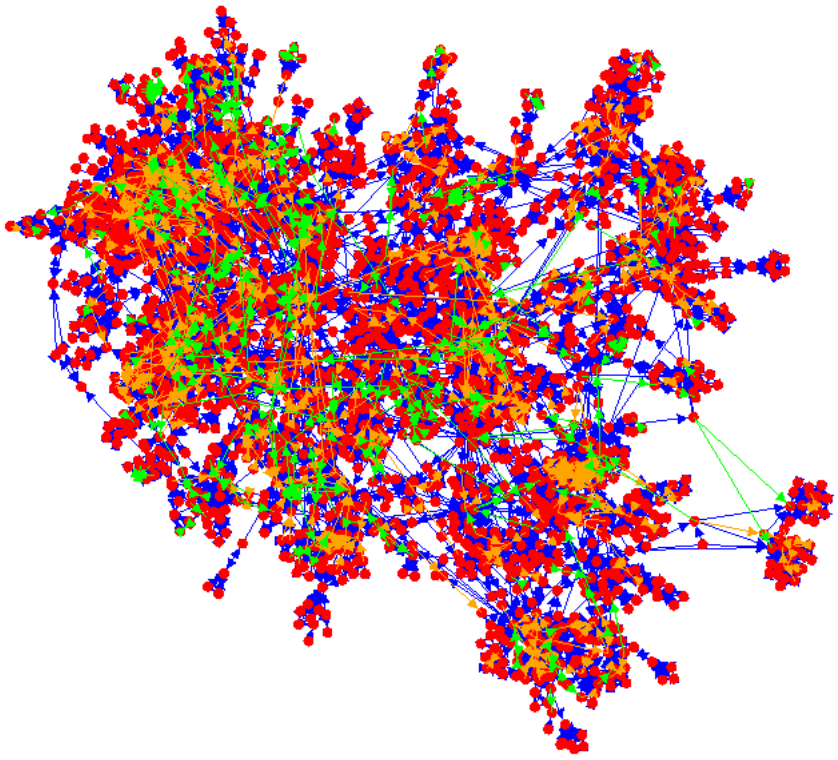

FIGURE 4.1 – Une vue de la Version 3 du graphe de co-publication de l'IRIT

d'auteurs en supprimant tout les liens faibles et les auteurs devenus orphelins, c'est à dire sans relation, suite à ce filtrage.

Pour chacune de ces trois versions, on réalise plusieurs détections de communautés à l'aide de l'algorithme OSLOM [Lancichinetti et al.(2011)].

4.2 Statistiques de prédiction de liens

Avant de ne retenir que les liens forts, la méthode détecte 72456 liens d'intérêts transitifs dont 1171 se réaliseront, c'est à dire que les auteurs impliqués dans ces liens transitifs publieront ensemble, soit 1.62%.

Après filtrage, la méthode ne conserve que 1020 dont 660 se réaliseront, soit 64.71%. On a donc plus de 98% des liens transitifs qui s'avèrent n'être que des liens faibles. Les 1020 liens transitifs représentent moins de 10% des

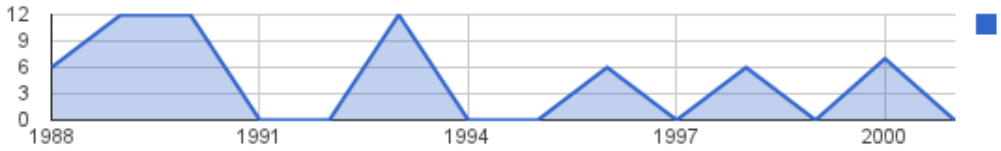

FIGURE 4.2 – Un exemple de profil d'activation d'une communauté

10504 collaborations.

4.3 Profils d'activation

Pour étudier l'impact de la méthode sur la cohérence des communautés, il nous faut définir un nouvel outil d'analyse de communautés, le profil d'activation.

Pour une communauté donnée, le profil d'activation est la courbe obtenue en relevant, à chaque intervalle, le nombre de liens internes à la communauté actifs dans l'intervalle (le nombre de liens entre deux acteurs de la communauté), le nombre de liens externes (le nombre de liens entre un acteur de la communauté et un acteur externe). L'image 4.2 représente un exemple de profil d'activation.

4.4 Résultats de l'analyse de communauté

Les résultats d'analyse de communauté présentés ci-dessous résultent de l'analyse de communautés par OSLOM[Lancichinetti et al.(2011)]. Ils sont représentatifs des résultats obtenus lors des différentes détections.

4.4.1 Version 1

- 241 communautés
- 2079 nœuds classés
- Nombre moyen de communautés par nœud : 1.06
- Taille moyenne des communautés : 9.15
- Degré entrant moyen des nœuds non classés : 1.20

– Degré sortant moyen des nœuds non classés : 1.38

4.4.2 Version 2

– 97 communautés
– 1956 nœuds classés
– Nombre moyen de communautés par nœud : 1.03
– Taille moyenne des communautés : 20.8
– Degré entrant moyen des nœuds non classés : 1.25
– Degré sortant moyen des nœuds non classés : 1.52

4.4.3 Version 3

– 117 communautés
– 2049 nœuds classés
– Nombre moyen de communautés par nœud : 1.06
– Taille moyenne des communautés : 18.65
– Degré entrant moyen des nœuds non classés : 1.34
– Degré sortant moyen des nœuds non classés : 2.08

4.5 Interprétation

4.5.1 Première analyse

On constate clairement que la prise en compte du poids a un effet sur l'analyse de communautés. L'algorithme détecte moins de communautés, mais leur taille est plus grande. Proportionnellement, la version 2 classe plus de nœuds. On peut donc en déduire que la prise en compte du poids dans la détection de communautés améliore les qualités d'intégration de cette dernière. On obtient donc moins de communautés, mais plus grandes.

Entre la version 2 et la version 3, on constate une augmentation du nombre de nœuds classés, du nombre de communautés et une diminution de la taille de ces dernières. Ces résultats laissent à penser que l'ajout des liens transitifs a amélioré la spécialisation détectée précédemment. Intuitivement, on peut interpréter l'augmentation du nombre de communautés et la baisse de leur

FIGURE 4.3 – Exemple de communauté cohérente localement

taille par la création de nouvelles communautés dont la découverte a été rendue possible par la création de liens transitifs.

4.5.2 Analyse de la cohérence locale

Après analyse et comparaison des différentes communautés obtenues, on peut mettre en avant trois résultats :

- Sur les communautés déjà bien définies (où il existe peu de liens externes à la communauté), l'application de la méthode n'a qu'un faible effet sur la cohérence. Cela provient du fait que la communauté ayant peu de connections vers l'extérieur, elle offre peu de chance à l'intérêt transitif. Aussi, le peu de liens transitifs créés ne fait qu'augmenter la taille de la communauté.
- Sur les communautés mal définies, l'application de la méthode provoque deux effets : soit la cohérence de la communauté est augmentée et on constate sur le profil d'activation l'apparition ou l'amplification d'un pic de localité, soit l'explosion de la communauté au profit d'autres.
- Enfin, on constate l'apparition de nouvelles communautés cohérentes dont le profil d'activation met en exergue un pic d'activité mettant en avant leur caractère temporel.

FIGURE 4.4 – Exemple de communauté à cohérence étendue

L'image 4.3 fournit un exemple de communauté à cohérence locale bien définie.

4.5.3 Analyse de la cohérence étendue

Sous les mêmes conditions d'expérimentation, on constate pour la cohérence étendue :

- Le même effet sur les communautés déjà bien définies.
- Sur les communautés mal définies, soit le lissage de la courbe d'activation au profit d'une activation plus "plate", soit l'explosion de la communauté.
- Enfin, l'apparition de nouvelles communautés cohérentes arborant un profil d'activation lisse.

L'image 4.4 fournit un exemple de communauté à cohérence étendue bien définie.

4.6 Synthèse

Nous avons appliqué la méthode au graphe de collaboration de l'IRIT sur lequel nous avons ensuite réalisé une analyse de communauté. Avec les résultats de l'analyse, nous avons dressé le profil d'activation des communautés obtenues afin d'étudier les différences entre les trois versions. Nous avons déterminé que la méthode produisait trois types d'effets : faible sur les

communautés bien définies, améliorant sur les communautés mal définies, et créatif pour les nouvelles communautés.

L'application de la méthode confirme un effet sur la cohérence, qu'elle soit locale ou étendue. Elle confirme l'idée que la dynamique des réseaux sociaux est source d'informations.

Cependant, bien que significatif, l'effet n'est pas aussi fort qu'escomptai et soulève la question de la sensibilité à l'intérêt transitif d'un graphe. Le graphe de collaboration de l'IRIT se prête-t-il mieux à la transitivité que d'autres graphes ? Existe t-il des structures plus sensibles à la transitivité ? La méthode reste donc à appliquer à divers domaines pour en tester les limites et les forces.

Chapitre 5

Domaines d'applications et perspectives d'évolutions

5.1 Domaines d'applications

La méthodologie présentée dans ce mémoire est applicable à tout système où un groupe plus ou moins important d'agents interagissent au cour du temps et où une étude des structures communautaires est intéressante.

Ainsi, on peut appliquer cette méthodologie pour étudier l'évolution de structures communautaires au cours du temps en s'intéressant soit à la découverte de communautés regroupant des acteurs contemporains, soit à la persistance de communautés dans le temps. Elle trouve donc des applications dans des domaines telles que l'économie, la sociologie, ou encore, la biologie.

Outre les aspects détection de communautés, la méthodologie présenté ici est applicable là où la construction d'un graphe orienté et pondéré est nécessaire. On peut par exemple, grâce à la méthode de pondération, réaliser une étude sur la circulation d'une information au sein d'un groupe, en considérant le poids des liens (après une simple normalisation) comme la probabilité d'interaction entre deux agents. Cette approche trouve donc des applications en marketing, épidémiologie ou encore dans l'étude de la circulation d'informations au sein d'un système multi-agent.

Enfin, la détection de liens transitifs permet la mise en relation d'acteurs

qui n'ont pas interagis mais partagent un intérêt commun. Cette mise en relation peut trouver son application dans des systèmes d'aide à la décision pour proposer, par exemple, à deux chercheurs partageant un même intérêt pour une troisième personne, de travailler ensemble. Aussi, elle trouve des applications dans tout domaine souhaitant favoriser les relations entre acteurs en proposant de nouveaux contacts. Cette approche trouve notamment des application en sociologie, et en aide à la décision.

5.2 Perspectives d'évolutions

5.2.1 Améliorer la rationalité des acteurs

Une des principales perspectives d'évolution de la méthode se situe au niveau de la rationalité des agents lors du processus de pondération. Dans la méthode présentée, les agents pondèrent leur relation uniquement en fonction de la topologie de leur voisinage. Ici, aucun processus cognitif n'intervient ni ne prend compte de facteurs propres à l'agent.

Une des pistes d'évolution serait d'augmenter la rationalité de l'agent en lui intégrant un processus de décision basé sur d'autres facteurs que la seule topologie. On peut par exemple utiliser un processus de notation de la qualité de l'interaction pour l'agent, intégrer des processus émotionnels ou autres facteurs internes à l'agent pour construire non pas un intérêt topologique seul, mais un réel intérêt de l'agent dans sa collaboration avec ses voisins.

En augmentant ainsi la rationalité des acteurs, on pourrait alors plus facilement discriminer les interactions entre acteurs, détecter des situations de non intérêt commun (un agent A considère sa relation avec B comme importante alors que B la considère faible) et probablement augmenter la qualité des résultats de la détection de communauté en détectant non plus des communautés topologiques, mais des communautés d'intérêt réel. On peut notamment se rapprocher de l'approche faite dans le cadre du projet SocLab (http ://soclab.univ-tlse1.fr) et de sa représentation de la satisfaction des agents et de l'influence des interactions sur cette dernière.

Un autre axe d'évolution serait le développement d'une méthode de détection de communauté basée sur notre méthode. Intégrer directement le

processus de pondération et d'intérêt à une méthode de détection de communauté pourrait, à terme, améliorer la qualité des résultats et fournir un outil complet d'analyse de communautés.

5.2.2 Améliorer le filtrage des liens

Dans la méthode présentée, nous filtrons les liens avec un seuil empirique. Bien qu'efficace, cette méthode ne permet pas de préserver la distribution des poids et peut alors sembler abusive. Des méthodes récemment développées[Radicchi et al.(2011)] proposent un filtrage des poids permettant la conservation de la loi de distribution et de la topologie du réseau. L'implémentation d'une telle technique adaptée à notre méthodologie pourrait alors présenter un intérêt certain dans cette méthode et pourrait alors en augmenter les résultats.

5.2.3 Vers un nouvel algorithme de détection de communauté

Notre méthode permet de classifier les liens par force. En se basant sur cette valeur, nous pourrions alors développer un nouvel algorithme de détection de communauté.

En fixant n pivots de communautés basés sur les n liens non connexes les plus forts, c'est à dire, n noyaux autours desquels construire nos n communautés, on pourrait alors construire n communautés où chaque communauté regroupe les nœuds les plus proches, en terme de chemins pondérés. Cette approche présente l'avantage de regrouper des acteurs autours de relations fortes et pourrait obtenir des résultats intéressants. Une autre méthode consisterait à créer un dendogramme en regroupant deux par deux les liens de poids les plus élevés. Reste alors à sélectionner le niveau de découpage approprié pour isoler les communautés.

Chapitre 6

Conclusion

Nous avons présenté dans ce mémoire une méthode originale proposant, à l'aide d'une approche multi-agent, d'extraire de l'information de la dynamique des réseaux sociaux. Elle nous permet d'obtenir un graphe orienté et pondéré sur lequel l'information de la dynamique a été reportée.

Grâce à cette information, on a montré que l'on pouvait améliorer les résultats des détections de communautés en créant deux types de cohérence, locale et étendue. De ces deux types de cohérence découle deux types de communautés. L'une, la locale, tend à s'intéresser à des groupes d'acteurs actifs sur les mêmes périodes, et donc à étudier si des communautés existent de manière ponctuelle. Cette approche trouve par exemple un intérêt pour détecter si des coalitions ont pris place à certaines années. L'autre, étendue, tend à s'intéresser à la persistance de communautés dans le temps.

L'application de la méthode à l'analyse du graphe de collaboration de l'IRIT nous a permis de mettre en avant l'effet que produit notre méthode sur l'analyse de communautés. On constate alors que, de manière générale, la méthode améliore la cohérence des communautés détectées mais est peu efficace sur des communautés bien définies.

Les domaines d'application de la méthode sont divers et variés et ses perspectives d'évolutions de la méthode promettent d'apporter un gain de qualité et ouvre la voie à une nouvelle approche de la dynamique des réseaux.

L'apport de ce mémoire est donc multiple. Il met en avant que la dynamique est une source certaine d'informations et ouvre la voie vers la détection

de nouveaux types de communautés. Le travail présenté ici reste donc à poursuivre et à améliorer, mais il est un pas en avant dans la prise en compte de la dynamique dans l'analyse de réseaux sociaux.

Bibliographie

[Amblard et al.(2011)] Amblard, F., Casteigts, A., Flocchini, P., Quattrociocchi, W. and Santoro, N., *On the temporal analysis of scientific network evolution*, in Proceedings the Third International Conference on Computational Aspects of Social Networks (CASoN 2011), Salamanca, Spain, CD-ROM, ISBN : 978-1-4577-1131-2

[Blondel et al.(2008)] Vincent D. Blondel,Jean-Loup Guillaume, Renaud Lambiotte,Etienne Lefebvre *Fast unfolding of communities in large networks* arXiv (2008) : 0803.0476.

[Cazabet et al.(2010)] Remy Cazabet, Frédéric Amblard, H. Hanachi, *Detection of overlapping communities in dynamical social networks*, Symposium on Social Intelligence and Networking, Minneapolis, Minnesota, USA, 20/08/2010-22/08/2010, Alex Pentland, Justin Zhan (Eds.), **IEEE Computer Society - Conference Publishing Services**, p. 309-314, 2010

[Cazabet et al.(2011)] Remy Cazabet, Frédéric Amblard, *Simulate to Detect : a Multi-agent System for Community Detection*, IEEE/WIC/ACM International Conference on Intelligent Agent Technology (IAT 2011), Lyon, France, 22/08/2011-27/08/2011, **IEEE Computer Society - Conference Publishing Services**, p. 402-407, 2011.

[Cazabet et al.(2012)] Cazabet Rémy, Leguistin Maud, Amblard Frédéric, *Automated community detection on social networks : useful ? Efficient ? asking the users*, Workshop on Web-Intelligence & Communities (WI&C 2012) (Lyon - France)

[Derényi et al.(2005)] I. Derényi, G. Palla, T. Vicsek *Clique percolation in random networks* Phys. Rev. Lett. 94, 160202 (2005)

[Farkas et al.(2007)] I. J. Farkas, D. Ábel, G. Palla, T. Vicsek *Weighted network modules* New J. Phys. 9, 180 (2007)

[Glattfelder et Battiston(2009)] J.B. Glattfelder and S. Battiston *Backbone of complex networks of corporations : The flow of control* Physical Review E 80 (2009)

[IMDB] Information courtesy of The Internet Movie Database (http ://www.imdb.com). Used with permission.

[Knuth(1993)] D. E. Knuth, *The Stanford GraphBase : A Platform for Combinatorial Computing*, Addison-Wesley, Reading, MA (1993).

[Lancichinetti et al.(2011)] Lancichinetti A, Radicchi F, Ramasco JJ, Fortunato S (2011) *Finding Statistically Significant Communities in Networks.* PLoS ONE 6(4) : e18961. doi :10.1371/journal.pone.0018961

[Lusseau et al.(2003)] D. Lusseau, K. Schneider, O. J. Boisseau, P. Haase, E. Slooten, and S. M. Dawson, *Behavioral Ecology and Sociobiology* **54**, 396-405 (2003)

[Palla et al.(2005)] Palla et. al., *Clique Percolation Method (CPM)* Nature 435, 814-818 (2005)

[Quattrociocchi et al.(2010)] Quattrociocchi W., Amblard F., Galeota E., *Selection in scientific networks*, Social Network Analysis and Mining journal (2010)

[Radicchi et al.(2011)] F. Radicchi, J. J. Ramasco, S. Fortunato *Information filtering in weighted complex networks* Physical Review E 83, 046101 (2011)

[Santoro et al.(2011)] Santoro, N., Quattrociocchi, W., Flocchini, P., Casteigts, A. and Amblard, F. *Time-Varying Graphs and Social Network Analysis : Temporal Indicators and Metrics*, in Social Network and Multi-Agent Systems Symposium (SNAMAS) @ Artificial Intelligence and Simulation of Behaviour Convention (AISB), York, UK, p.33-38, 2011

[Scott(1988)] John Scott *Social Network Analysis* Sociology February 1988 vol. 22 no. 1 109-127

[Sueur et al.(2011)] Sueur C., Jacobs A., Amblard F., Petit O. and King A.J., *How can social network analysis improve the study of primate behavior ?*, American Journal of Primatology, vol.71, p.1-17, 2011

[Thomas et al.(2002)] Vincent Thomas, Christine Bourjot, Vincent Chevrier, Didier Desor *MAS and RATS : Multi-agent simulation of social differentiation in rats' groups* International Workshop on Self-Organization and Evolution of Social Behaviour 10 p (2002)

[Zachary(1977)] W. W. Zachary, *An information flow model for conflict and fission in small groups* Journal of Anthropological Research **33**, 452-473 (1977).

[Wang et al.(2011)] Dan Wang, Xiaolin Shi, Daniel A. McFarland, Jure Leskovec. *Measurement error in network data : a re-classification.* Social Networks. Forthcoming, July, 34 (3)(2011).

List of Algorithms

Table des figures