Contributions à l'étude des réseaux sociaux : propagation, fouille, collecte de données

Erick Stattner

UNIVERSITÉ DES ANTILLES ET DE LA GUYANE

ÉCOLE DOCTORALE UAG
SCIENCES ET TECHNOLOGIES DE L'INFORMATION ET DE LA COMMUNICATION

T H È S E

pour obtenir le titre de

Docteur en Sciences

de l'Université des Antilles et de la Guyane
Mention : Informatique

Présentée et soutenue par

Erick Stattner

Contributions à l'étude des réseaux sociaux : propagation, fouille, collecte de données

Thèse dirigée par Martine Collard

préparée au Laboratoire LAMIA

soutenue le 10 Décembre 2012 en Guadeloupe

Jury :

Rapporteurs :	Christine Largeron, Professeur, Université Jean Monnet, Saint-Etienne
	Yves Duthen, Professeur, Université Paul Sabatier, Toulouse 1
Directeur :	Martine Collard, Professeur, Université des Antilles et la Guyane
Encadrant :	Nicolas Vidot, Maître de conférences, Université des Antilles et la Guyane
Président :	Jean Vaillant, Professeur, Université des Antilles et la Guyane
Examinateur :	Jacqueline Deloumeaux, Docteur en médecine, CHU de Guadeloupe

“ *Il pensait savoir ce qu'était la science. Rien de plus que la curiosité... tempérée par l'humilité, disciplinée par la patience.* ”

Robert-Charles Wilson

Remerciements

Ce travail n'aurait jamais été possible sans le soutien et l'appui d'un ensemble de personnes que je tiens à remercier ici chaleureusement.

Merci tout d'abord à l'ensemble des membres du jury pour m'avoir fait l'honneur d'accepter de juger ce travail de thèse.

En premier lieu, je tiens à remercier profondément mon directeur de recherche, Martine Collard, qui m'a fait confiance et m'a permis de développer un thème de recherche passionnant. Plus qu'un directeur, elle a su se montrer présente tout au long de ces années et me soutenir dans mes démarches. Je tiens ici à lui témoigner toute ma gratitude et ma reconnaissance pour sa générosité, le temps qu'elle m'a accordé et la patience dont elle a fait preuve. Ce travail n'aurait en effet jamais été possible sans son implication et son dévouement. Qu'elle trouve ici toutes les marques de mon profond respect et de mon admiration.

Je ne saurais également trop remercier mon encadrant, Nicolas Vidot. Sa grande disponibilité, ses conseils avisés et son optimisme à toute épreuve ont été de précieux atouts durant ces années. Mes nombreux échanges avec lui et ses questions toujours très pertinentes ont permis d'approfondir et d'enrichir ce travail de thèse.

Un grand merci également à Philippe Hunel, pour m'avoir toujours suivi et soutenu amicalement depuis mes débuts à l'Université. Il a su m'initier très tôt aux travaux de recherche et m'a montré la voie à suivre. Ces quelques lignes n'auraient certainement jamais été écrites sans ses encouragements et sa politique de dépassement de soi.

Un merci particulier à Harry et Jean-Raphael Gros-Desormeaux pour toutes ces discussions enrichissantes. Ils ont su partager avec moi leurs expériences, me permettant ainsi d'affronter les épreuves.

Évidemment, les mots ne sauraient décrire l'immense gratitude envers ceux, qui dans l'ombre, m'ont permis de façon indirecte, mais bien réelle, de mener à terme ce travail. Je pense tout d'abord à ma famille, Jacqueline, Claude et Laurent, que je remercie du fond du coeur pour leur indéfectible soutien.
Et ma compagne Stéphanie... qui a su m'épauler et me relever dans les moments de doutes et de découragements. Je la remercie d'ailleurs grandement d'avoir trouvé le temps de relire le manuscrit, malgré les délais très courts.

Enfin, merci à mes amis, Marc Panor, Jérémy Frominville, Laurent Charleroy, Jean-Gabriel Mazeroi, Carole Louis-Rose, Cedric Ramassamy, Nathalie Dessart, ... pour leur bonne humeur, leurs encouragements ainsi que tous ces moments de détente et d'évasion, ô combien nécessaires pour affronter les difficultés présentes tout au long de ce dur chemin qui mène à la thèse.

A tous, Merci...

Table des matières

Table des figures

Liste des algorithmes

Chapitre 1

Introduction

Contexte

Le concept de réseau offre un modèle de représentation pour une grande variété d'objets et de systèmes, aussi bien naturels que sociaux, dans lesquels un ensemble d'entités homogènes ou hétérogènes interagissent entre elles. Dérivé du mot latin *"rete"*, qui désignait à l'origine un tissu ou un filet, le terme de *"réseau"* est aujourd'hui employé couramment pour désigner divers types de structures relationnelles telles que des voies de communications, de circulations ou d'échanges, pouvant être réelles ou abstraites, entre des entités ou des groupes d'entités. Le concept mathématique de *graphe*, défini par un ensemble d'objets (ou noeuds), et un ensemble de liens (ou arêtes) entre les objets, est communément utilisé comme support formel pour modéliser la structure des réseaux dont la sémantique est très diverse. On parle par exemple de *réseaux de collaboration*, de *réseaux professionnels*, du *réseau Internet*, et plus récemment de *réseaux sociaux*, dont les liens ont une portée sociale et de *réseaux complexes*, dont l'évolution conduit à l'émergence de propriétés structurelles particulières. Pour faire référence aux noeuds (resp. aux arêtes) du graphe qui modélise un réseau, on utilise selon le contexte, le terme d'objet, d'entité, d'agent, d'individu, d'acteurs, etc. (resp. de lien, d'interaction, de relation, etc.).

Pourtant, si chacun a une idée plus ou moins précise de ce qu'est un réseau, de ce que sont les acteurs impliqués, et les types de relations qui les lient, nous ignorons encore souvent les implications qu'ont ces relations dans de nombreux phénomènes du monde qui nous entoure. Citons par exemple des processus tels que la diffusion d'une rumeur, la transmission d'une maladie, la réputation d'une enseigne ou d'un produit, ou même l'émergence de sujets d'intérêt commun à un groupe d'individus, dans lesquels les relations que maintiennent les individus entre eux et leur nature s'avèrent souvent être les principaux facteurs déterminant l'évolution du phénomène.
Dans le cas de la propagation d'une rumeur, il est aujourd'hui établi que les liens de confiance et d'influence que maintient un individu avec son entourage sont les principaux facteurs impliqués dans la probabilité qu'il a d'accepter et de transmettre l'information [Zanette 2002]. Un autre exemple intéressant est celui des phénomènes d'achat ; en effet, les travaux classiques, menés essentiellement en analyse statistique et en fouille de données, tentent généralement d'établir des modèles prédictifs uniquement sur la base d'attributs démographiques ou individuels. Or, les choix et les comportements sont majoritairement déterminés par les relations sociales des acteurs avec leur entourage et les communautés auxquelles ils appartiennent [Leskovec 2007]. Des études récentes ont même montré comment certains états physiologiques ou spirituels, ou comportements, tels que l'obésité [Christakis 2007], le bonheur [Fowler 2008] ou le tabagisme [Christakis 2008], sont fortement dépendants des liens sociaux entre les individus.

Récemment, dans une conversation publique [Barabasi 2009] entre Albert-Laszlo Barabasi, l'inventeur du concept de *réseau scale-free*, et le politologue James Fowler, Barabasi affirme que "*le fait que nous vivons à l'ère des réseaux est devenu un truisme*". En effet, les réseaux sont aujourd'hui partout autour de nous : l'Internet, les réseaux sociaux, les

réseaux d'amitié, les réseaux d'échange et de partage, les réseaux d'appels téléphoniques, etc. D'une certaine façon, chacun de nous est un noeud de multiples réseaux interconnectés, définis par des relations de différentes natures avec les membres de notre famille, nos amis, les individus avec lesquels nous travaillons, ceux que nous rencontrons lors de nos déplacements et ceux avec lesquels nous avons des activités.

L'étude des réseaux est devenue l'un des domaines phares du 21e siècle, renommée de manière consensuelle la "*Science des Réseaux*" par plusieurs auteurs reconnus du domaine [Barabasi 2002, Watts 2004, Borner 2007, Newman 2010]. Dans l'état actuel des connaissances, il est cependant assez difficile d'en donner une définition précise en raison de son caractère multidisciplinaire. Par exemple, Watts [Watts 2004] définit la science des réseaux comme "*la science du monde réel − du monde des gens, des relations d'amitié, des rumeurs, des maladies, des modes, des entreprises et des crises financières*". Cette définition est intéressante à deux points de vue, puisqu'elle met tout d'abord en avant le caractère fortement interdisciplinaire du domaine, mais également la dimension complexe des *problèmes abordés*. Le conseil national de la recherche des États-Unis définit la science des réseaux comme "*l'étude des représentations sous forme de réseaux de phénomènes physiques, biologiques et sociaux, conduisant à des modèles prédictifs*" [Press 2005].
D'un point de vue beaucoup plus général, la *science des réseaux* peut être considérée comme la discipline qui se donne pour but d'étudier les relations entre des objets (hommes, animaux, cellules, machines, etc.) et non pas les objets eux-mêmes. Autrement dit, le postulat initial en science des réseaux consiste à considérer les interactions entre entités comme les éléments les plus pertinents pour étudier et comprendre certains phénomènes du monde réel.

Les premiers travaux du domaine, principalement consacrés à l'analyse précise de la structure topologique des réseaux, ont naturellement exploité les outils issus de la théorie des graphes [West 2000]. Au cours de la dernière décennie, les recherches en science des réseaux, auxquelles ont contribué de nombreuses communautés scientifiques telles que la sociologie, la biologie, les mathématiques ou l'informatique, ont connu un intérêt croissant qui peut s'expliquer par deux facteurs majeurs.

Tout d'abord, les réseaux offrent un cadre simple et universel qui permet de décrire et d'étudier une grande variété de systèmes présents dans la nature, la société ou le monde technologique. Par exemple, une cellule est couramment décrite par un réseau de composants chimiques reliés par des réactions. Internet est, lui, un réseau de routeurs et d'ordinateurs liés par divers liens de communications. Les phénomènes de mode, les rumeurs ou les idées se répandent souvent à travers un réseau social dont les noeuds sont des êtres humains et les liens représentent divers types de relations sociales. De même, la toile est un immense réseau virtuel de pages reliées entre elles par des liens hypertextes.

Le deuxième facteur d'explication concerne les nouvelles technologies d'information et de communication (TIC) devenues omniprésentes dans nos vies, à l'origine des sites communautaires ou de la téléphonie mobile par exemple, qui donnent d'une part un support physique nécessaire à la vie du réseau, mais qui permettent également la collecte d'énormes quantités de données potentiellement utiles et pertinentes pour l'analyse des processus sociaux.

Dans ce mémoire, nous considérons l'ensemble des travaux qui tendent à une meilleure compréhension du rôle des réseaux dans les processus sociaux comme s'articulant autour de quatre grands axes de recherche, comme nous l'illustrons dans les Figures 1.1 et 1.2 :
(i) La modélisation des processus de création ou d'évolution des réseaux, ou des phénomènes qu'ils supportent, peut s'effectuer *a priori* à partir de la connaissance acquise dans le domaine, ou bien *a posteriori* en étudiant les données qui ont pu être relevées. Les com-

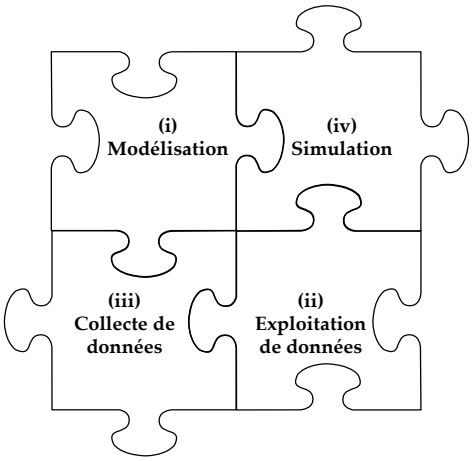


FIGURE 1.1 – Classification des travaux dans le domaine de la science des réseaux

portements d'influence offrent des cas typiques de problèmes étudiés à travers des modèles basés sur les réseaux [Cosley 2010]. Les phénomènes de propagation (maladie, rumeur, information, etc.), auxquels nous nous intéressons particulièrement dans ce mémoire, offrent également un exemple de problèmes couramment étudiés à travers des modèles de diffusion sur les réseaux [Christley 2005]. Des travaux fondateurs ont été consacrés à la modélisation des processus conduisant à la construction de réseaux ayant des caractéristiques structurelles similaires à celles observées sur les réseaux réels, par la mise en place de modèles visant à décrire la formation et l'évolution de leur structure [Toivonen 2009]. Citons par exemple Barabasi et Albert [Barabasi 1999] qui ont proposé un modèle basé sur la notion d'*attachement préférentiel*, permettant de générer des réseaux *scale-free*. Ces modèles peuvent être vus comme définis "*a priori*", car ils sont d'une certaine façon synthétiques et basés sur la formalisation d'une connaissance acquise dans le domaine médical, social ou économique par exemple. La validation ou la mise au point de ces modèles peut également s'effectuer sur des données réelles si le cas de figure étudié le permet. On peut y opposer des modèles obtenus "*a posteriori*" par **l'exploitation de données réelles ou simulées** issues de l'activité supportée par un réseau.

(ii) L'exploitation des données, deuxième axe de recherche dans ce domaine, recouvre l'ensemble des méthodes visant à analyser les activités supportées par les réseaux à partir de traces sauvegardées. Les méthodes traditionnelles exploitent uniquement les mesures issues de la théorie des graphes [Milgram 1967] pour caractériser les noeuds ou la structure du réseau. Par exemple, une étude célèbre menée par Elizabeth Bott [Bott 1957] sur un échantillon de familles Londoniennes a pu mettre en évidence la corrélation entre la densité de liens dans le réseau maintenu avec les amis et les membres d'une famille, et la répartition des tâches domestiques au sein des couples. Sur le même principe, d'autres mesures comme le degré ou le coefficient de clustering ont été utilisées pour explorer la structure des réseaux, identifier les individus centraux, ou classifier les noeuds selon divers critères portant sur leurs propriétés structurelles. Aujourd'hui les méthodes récentes, dites de *fouille de réseaux sociaux*, *fouille de données sociales*, ou *social mining*, visent à extraire de la connaissance sur le réseau en appliquant les concepts de la fouille de données classique. Les tâches les plus courantes sont la visualisation [Snasel 2008], la classification [Getoor 2005], la recherche de communautés [Fortunato 2009], la modélisation prédictive [Liben-Nowell 2007] ou la recherche de motifs fréquents [Kuramochi 2005]. Se pose ainsi, la question d'une part

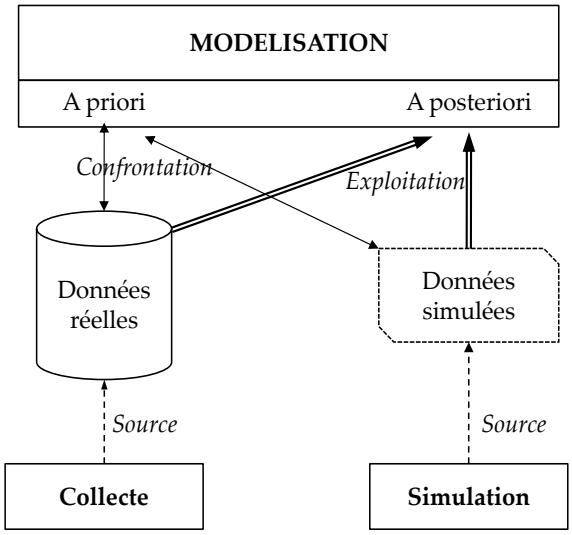


FIGURE 1.2 – Des données sociales aux modèles

de la **collecte** des données réelles et de leur fiabilité et d'autre part, dans les situations qui interdisent d'obtenir une trace de l'activité des processus, de la manière de les simuler.

(iii) Le troisième axe de recherche, **la collecte de données** pertinentes, est **tributaire** des infrastructures logicielles ou matérielles mises en place pour le recueil de données témoignant de l'activité sur un réseau. Typiquement, les sources d'informations fournies par les sites d'échange et de partage (blog, site communautaire, site d'achat, etc.) permettent de construire des réseaux sociaux impliquant les utilisateurs. Le fort développement que connaissent aujourd'hui les périphériques mobiles tels que les smartphones, les dispositifs GPS, ou les puces RFID, ouvre de nouvelles perspectives pour la collecte de données spatio-temporelles réelles sur les entités [Olguin 2008, Croft 2008a]. Le choix d'une technologie et la configuration des périphériques de collecte sont donc des éléments importants pour assurer la qualité des données à exploiter. Dans certaines situations, la collecte de données réelles est impossible. C'est par exemple le cas des phénomènes de diffusion. Dans le cas de la diffusion de maladies infectieuses à grande échelle, il est souvent extrêmement difficile de suivre l'évolution réelle du phénomène, parce que les liens ayant permis la transmission ne peuvent pas être connus, ou parce que que les individus infectés eux-même ne sont souvent pas totalement recensés. C'est également le cas de la diffusion des rumeurs dans la société, via par exemple un site communautaire ; bien que les liens sociaux entre les individus puissent être connus, l'extraction d'informations sur le phénomène pose souvent des questions juridiques qui n'autorisent pas leur exploitation. C'est ainsi que de nombreux travaux dans ce domaine ont recours à des **simulations**.

(iv) D'une façon générale, **la simulation**, quatrième axe et domaine de recherche complémentaire, permet de valider un modèle et d'en mesurer les performances. Elle peut également être mise en place pour évaluer l'efficacité d'une mesure, ou d'un algorithme dans différentes configurations. Elle offre une alternative dans le cas ou les données réelles sont difficiles à obtenir. Dans le cas des processus de diffusion par exemple, la simulation permet de comprendre comment se comporte le phénomène selon différentes configurations [Salathe 2010b]. Dans le cas de l'exploitation des données, les performances d'un algorithme peuvent être analysées avec plusieurs jeux de paramètres [Kuramochi 2001]. Enfin, en ce qui concerne la collecte des données, des outils de simulation peuvent être

conçus pour générer des jeux de données synthétiques basés par exemple sur des études statistiques [Barrett 2008].

Problématiques et Contributions

Dans le cadre de ce mémoire, nous nous limitons au cas des *réseaux sociaux*, dans le sens où les réseaux étudiés sont assimilables à un ensemble d'entités ayant un rôle social et des interactions possédant une sémantique sociale. Bien que la science des réseaux ait déjà permis d'aborder une grande variété de sujets, tels que l'étude des phénomènes de diffusion ou d'influence, la classification ou la prédiction de liens, un certain nombre de questions dans ce panorama très large, restent encore ouvertes.

(i) En ce qui concerne **la modélisation**, nous nous sommes intéressés au cas spécifique de **la diffusion dans les réseaux dynamiques**.
Pour les phénomènes de diffusion, des modèles mathématiques fondateurs (*SI*, *SIR*, *SIS*, etc.) ont été adaptés aux réseaux pour comprendre comment se propagent des maladies ou des rumeurs à travers un réseau social. Nous observons que peu de modèles sont capables de rendre compte de la dynamique du réseau. Or, les changements survenant sur un réseau peuvent être d'une part dus à des facteurs endogènes, tels que le voisinage d'un noeud ou la communauté à laquelle il appartient, mais également à des facteurs exogènes déterminant le comportement propre d'un noeud comme sa mobilité dans l'espace. Ainsi, l'un des sujets les plus importants en matière de modélisation réside dans la conception de modèles capables d'intégrer tous les aspects dynamiques des réseaux ainsi que leur inter-dépendance. Cependant, si l'approche réseau est aujourd'hui largement répandue dans les études menées sur les phénomènes de diffusion, la plupart des travaux s'intéressent uniquement à des réseaux statiques, alors que la majeure partie des réseaux du monde réel est dynamique. En effet, des liens et des noeuds peuvent apparaitre ou disparaitre dans le réseau au cours du temps.

Le premier sujet abordé dans ce mémoire est celui de **la diffusion dans des réseaux en évolution**, une problématique qui soulève des questions de modélisation et qui nous a conduits à mettre au point et à utiliser des outils de simulation.
Nous avons adopté une approche exploratoire en mesurant et comparant les effets de changements topologiques induits par différents facteurs. Nous avons montré que la dynamique du réseau peut être le résultat de facteurs endogènes et exogènes au réseau. Ainsi, l'une des premières contributions de ce travail a été de proposer le modèle *D2SNet* (**D**iffusion in **D**ynamic **S**ocial **Net**works), qui prend en compte les deux processus impliqués et leurs inter-dépendances : le processus de diffusion et le processus d'évolution du réseau lui-même [Stattner 2011c, Stattner 2012j]. Par la suite, nous avons utilisé ce modèle pour mesurer l'impact de mécanismes d'évolution simples et plus avancés.
Nous présentons dans la suite de ce mémoire une première étude menée sur **l'effet de mécanismes d'évolution élémentaires**, identifiés comme étant à la base de la formation des liens dans de nombreux réseaux du monde réel [Stattner 2011b, Stattner 2012d]. Cette première étude a permis de mettre en évidence l'effet de la dynamique du réseau en montrant que le processus de diffusion se comporte différemment selon le mécanisme considéré. Dans une deuxième étude [Stattner 2012e], nous considérons **l'environnement social** comme le principal moteur de la formation de liens au sein des réseaux. Cette deuxième approche est venue compléter nos résultats précédents en montrant que l'impact de la dynamique du réseau dépend de l'environnement social dans lequel évoluent les noeuds.
Enfin, une troisième étude s'intéresse au cas particulier où la dynamique du réseau est induite par **la mobilité des individus** [Collard 2012]. Pour cela, nous avons proposé le modèle de mobilité *ER* (*Eternal-Return*), qui tient compte des nouvelles connaissances ré-

cemment acquises à partir des traces réelles d'individus, et nous l'utilisons pour mesurer les effets de la mobilité sur le processus de diffusion. Cette étude à permis d'observer deux tendances : (1) Quand les agents sont en mouvement, un seuil de densité minimal doit être garanti pour qu'une diffusion soit possible. (2) Quand la diffusion est possible, le processus varie fortement selon la mobilité des individus.

Outre la mobilité spatiale, parmi les multiples facteurs susceptibles d'influencer le comportement social des agents, leurs propriétés intrinsèques jouent un rôle évident. La modélisation de la diffusion devrait ainsi prendre en compte les profils des individus et les réactions qu'ils induisent dans la transmission de l'information.

(ii) En ce qui concerne **l'exploitation des données**, nous nous sommes intéressés à **l'intégration des propriétés des noeuds dans l'extraction des motifs fréquents**. Avec l'émergence des réseaux sociaux, de nouvelles problématiques en termes de fouilles de données dites "*sociales*", ou "*fouille de réseaux sociaux*", sont apparues. La première concerne naturellement l'adaptation et la mise en place de nouveaux algorithmes d'extraction de connaissance à partir de données sous forme de réseaux. Bien que d'importants travaux aient déjà été menés sur l'adaptation des méthodes standards, un problème encore très peu abordé concerne la prise en compte des différentes sources d'informations sur les noeuds et les liens. En effet, de nombreux réseaux, et en particulier les réseaux sociaux, disposent d'information sur les propriétés des éléments du réseau et dans certains cas sur leur évolution. C'est par exemple le cas des réseaux d'amitié dans lesquels les individus sont décrits par un ensemble d'attributs démographiques (age, sexe, etc.), et personnels (intérêt pour un sujet, etc.). Les réseaux d'achats bipartis, liant les individus aux produits qu'ils achètent, et au sein desquels les clients et les produits sont définis par un ensemble de caractéristiques, offrent un autre exemple. Les méthodes d'analyse et de fouille de données actuelles prennent encore rarement en compte ces informations lors de la phase d'extraction de connaissance.

La majorité des méthodes de fouille de réseaux sociaux ne s'intéresse pour l'instant qu'à la structure topologique du réseau. Cependant, les propriétés d'un noeud apportent souvent des informations pertinentes sur le rôle ou l'influence que possède un noeud dans le réseau. Il est évident que la recherche de corrélations entre les propriétés d'un noeud et son rôle dans le réseau peut apporter des éléments d'information pour la modélisation des processus sociaux. Nous nous sommes particulièrement intéressés à la recherche de liens fréquents entre des groupes de noeuds vérifiant des propriétés communes. Plus précisément, nous recherchons des motifs de la forme "*les noeuds qui vérifient la propriété A sont fréquemment liés aux noeuds qui vérifient la propriété B*. Une connaissance de ce type est particulièrement intéressante pour la tâche de modélisation puisqu'elle apporte une information pertinente sur ce qui motive la formation du lien social entre les individus, et donc a fortiori sur les connexions pertinentes en mesure de porter la diffusion d'une maladie ou d'une rumeur. Cette problématique originale soulève des questions multiples en termes de formalisme, d'algorithmique et d'évaluation des performances.
Les deux principales contributions de ce travail ont été : (1) la définition d'un **motif fréquent** particulier, appelé "***lien conceptuel***", qui tient compte à la fois des attributs des noeuds et de la structure du réseau [Stattner 2012f, Stattner 2012g] et (2) **l'algorithme *MFCL-Min***, qui extrait l'ensemble de ces motifs d'un réseau social [Stattner 2012a, Stattner 2012i]. Nous avons montré que les "*liens conceptuels*" que nous proposons permettent d'une part d'extraire une connaissance pertinente sur le réseau et d'autre part fournissent une "*vue conceptuelle*" du réseau, qui résume l'ensemble des liens pertinents au sein de la structure [Stattner 2012c].

La pertinence de tels motifs est fortement liée à la quantité, la qualité et la fiabilité des

données manipulées. Bien qu'avec l'utilisation massive des sites d'échange et de partage, on soit tenté d'exploiter les données issues des sites communautaires, ceci soulève énormément de questions sur la fiabilité, la légalité, le partage ou la publication de telles données, ce qui nous a amenés naturellement à nous poser la question de la collecte de données.

(iii) En ce qui concerne **la collecte de données**, nous nous plaçons dans le cas pratique **du suivi d'interactions sociales**.
La question de la collecte de données est en effet un axe qui soulève divers types de problèmes. Sur un plan purement technique, la principale problématique concerne le type d'architecture matérielle ou logicielle à mettre en place, de manière à ce qu'elle soit adaptée au phénomène étudié. Une autre problématique découle directement des jeux de données obtenus à partir des informations en ligne. Certains travaux s'intéressent actuellement aux diverses questions liées à la sécurité, à la confidentialité lors de la collecte et à l'anonymisation pour la publication. L'approche la plus répandue actuellement pour la collecte de données sociales consiste en l'utilisation de dispositifs mobiles attachés aux entités étudiées. Cependant, il existe certaines situations dans lesquelles la mise en place de tels dispositifs n'est pas envisageable. Nous nous sommes particulièrement intéressés à ce dernier cas de figure dans le contexte de l'étude et de l'observation d'une espèce animale, pour laquelle la végétation très dense vient perturber le bon fonctionnement des périphériques de collecte. Le point de vue de la collecte de données sociales est ainsi abordé dans ce mémoire dans **le contexte d'environnements bruités et à l'aide de dispositifs fixes**.

Dans le cadre de cette étude, nous nous sommes plus particulièrement focalisés sur l'étude d'une espèce d'oiseaux protégée, pour laquelle ni la configuration des lieux, ni la règlementation ne permettent l'utilisation de colliers GPS. Nous avons proposé une architecture matérielle composée uniquement de capteurs fixes, capable d'identifier et d'enregistrer les interactions entre les oiseaux [Stattner 2010a, Stattner 2010b], puis nous avons analysé, par la simulation, l'efficacité de notre solution dans diverses configurations [Stattner 2011a].

Organisation du mémoire

La suite de ce mémoire est organisée en 6 chapitres.
Le Chapitre 2 est dédié à un état de l'art qui, dans un premier temps, énonce des rappels sur la théorie des graphes dans le but d'introduire les notions et le vocabulaire utilisés dans le mémoire, et dans un second temps, passe en revue les travaux effectués sur les questions de la propagation dans les réseaux, de la fouille de données sociales et des architectures de collecte de données sociales.
Le Chapitre 3 présente notre approche de **modélisation** des processus de diffusion sur un réseau dynamique. Nous y présentons dans un premier temps le modèle *D2SNet*, dont l'objectif est de représenter les processus de diffusion sur des réseaux en évolution, puis le modèle est utilisé pour mener deux études. L'outil graphique *DynSpread*, qui implémente l'approche *D2SNet* est également présenté dans ce chapitre.
Le Chapitre 4 est consacré à l'étude menée sur l'impact du facteur mobilité sur le processus de propagation. Nous discutons les principes de conception du modèle de mobilité *Eternal-Return* en montrant que les approches traditionnelles ne rendent pas totalement compte des observations faites récemment à l'aide de dispositifs mobiles ou dans le suivi de transactions bancaires. La pertinence de l'approche est démontrée sous plusieurs aspects et le modèle est utilisé pour étudier l'impact de la **dynamique du réseau**, induite par la mobilité des individus, sur le processus de propagation. Enfin, nous présentons *ER-Net*, l'outil graphique de simulation qui implémente le modèle.
Le Chapitre 5 présente nos travaux visant à intégrer les **propriétés des noeuds** lors la

recherche de **motifs fréquents lors de l'exploitation**. Nous commençons par nous intéresser aux méthodes qui visent à extraire des motifs au sein des réseaux et montrons leurs difficultés à répondre à certaines interrogations. Nous présentons ensuite formellement les concepts de *liens* et de *vues conceptuels* avant de détailler l'algorithme d'extraction *MFCL-Min* et d'en analyser la flexibilité et la complexité. Les performances de la solution sont analysées dans diverses configurations, en menant une série de tests à l'aide de l'outil graphique d'extraction *GT-FCLMin* qui implémente l'approche.
Le Chapitre 6 est consacré à l'architecture de capteurs sans-fils que nous avons proposée pour **la collecte d'interactions sociales**. Nous discutons le choix de l'architecture en présentant les avantages et les inconvénients de l'approche. Puis, à l'aide du simulateur *Lypus*, que nous avons conçu pour recréer un environnement virtuel sur lequel des oiseaux et un réseau de capteurs peuvent être disposés, nous vérifions que l'architecture proposée est en mesure de détecter différentes communautés en analysant le réseau de contacts collecté.
Le Chapitre 7 vient conclure ce mémoire et présenter les nouvelles perspectives de recherche ouvertes par nos travaux.

Publications

Les travaux présentés dans ce mémoire ont donné lieu à la publication de 20 articles dans des revues et des conférences nationales et internationales dont la liste est donnée ci-dessous :
Chapitre 3 :

- **Revues :**
  - E. Stattner, M. Collard, N. Vidot : D2SNet : Dynamics of diffusion and dynamic human behaviour in social networks. Computer in Human Behavior (CHB), Elseiver, 2012
  - E. Stattner, M. Collard, N. Vidot : Network-Based Modeling in Epidemiology : An Emphasis on Dynamics. International Journal of Information System Modeling and Design (IJISMD), IGI-Global, 2012
- **Conférences internationales :**
  - E. Stattner, M. Collard, N. Vidot : Sociability VS Network Dynamics : Impact of Two Aspects of Human Behavior on Diffusion Phenomena. Advances in Social Network Analysis and Mining (ASONAM), IEEE, 2012
  - E. Stattner, M. Collard, N. Vidot : Towards Merging Models of Information Spreading and Dynamic Phenomena in Social Networks. World Summit on the Knowledge Society (WSKS), Springer, 2011
  - E. Stattner, M. Collard, N. Vidot : Diffusion in Dynamic Social Networks : Application in Epidemiology. Database and Expert Systems Applications (DEXA), Springer, 2011
  - E. Stattner, N. Vidot : Social network analysis in epidemiology : Current trends and perspectives. Research Challenges in Information Science (RCIS), IEEE, 2011

Chapitre 4 :

- **Revues :**
  - M. Collard, P. Collard, E. Stattner : Human Mobility and Information Diffusion : The ETERNAL-RETURN multi-agent model. Journal of Artificial Societies and Social Simulation (JASSS), 2012 (SOUMIS)
- **Conférences internationales :**
  - M. Collard, P. Collard, E. Stattner : Mobility and information flow : percolation in a multi-agent model. Ambient Systems, Networks and Technologies (ANT),

Elseiver, 2012

Chapitre 5 :

- **Conférences internationales :**
  - E. Stattner, M. Collard : Social-Based Conceptual Links : Conceptual Analysis Applied to Social Networks. Advances in Social Network Analysis and Mining (ASONAM), IEEE, 2012
  - E. Stattner, M. Collard : MAX-FLMin : An Approach for Mining Maximal Frequent Links and Generating Semantical Structures from Social Networks. Database and Expert Systems Applications (DEXA), Springer, 2012
  - E. Stattner, M. Collard : Frequent Links : An Approach that Combines Attributes and Structure for Extracting Frequent Patterns in Social Networks. Advances in Databases and Information Systems (ADBIS), Springer, 2012
  - E. Stattner, M. Collard : How to extract frequent links with frequent itemsets in social networks ? Research Challenges in Information Science (RCIS), IEEE, 2012
  - E. Stattner, M. Collard : FLMin : An Approach for Mining Frequent Links in Social Networks. Networked Digital Technologies (NDT), Springer, 2012
- **Conférences nationales :**
  - E. Stattner, M. Collard : Vers une Analyse Conceptuelle des Réseaux Sociaux. Modèles et l'analyse des réseaux : Approches mathématiques et informatiques (MARAMI), 2012
  - E. Stattner, M. Collard : GT-FLMin : Un Outil Graphique pour l'Extraction de Liens Fréquents dans les Réseaux Sociaux. Extraction et gestion des connaissances (EGC), RNTI, 2012
  - E. Stattner, M. Collard : Extraction de Liens Fréquents dans les Réseaux Sociaux. Extraction et gestion des connaissances (EGC), RNTI, 2012

Chapitre 6 :

- **Conférences internationales :**
  - E. Stattner, N. Vidot, P. Hunel, M. Collard : Wireless sensor networks for Habitat Monitoring : A Counting Heuristic. Local Computer Networks (LCN), IEEE, 2012
  - E. Stattner, M. Collard, P. Hunel, N. Vidot : Wireless sensor networks for social network data collection. Local Computer Networks (LCN), IEEE, 2011
  - E. Stattner, P. Hunel, N. Vidot, M. Collard : Acoustic scheme to count bird songs with wireless sensor networks. World of Wireless, Mobile and Multimedia Networks (WOWMOM), IEEE, 2011
  - E. Stattner, M. Collard, P. Hunel, N. Vidot : Detecting movement patterns with wireless sensor networks : application to bird behavior. Advances in Mobile Multimedia (MoMM), ACM, 2010
  - E. Stattner, M. Collard, P. Hunel, N. Vidot : A Data Collection Framework for Tracking Collective Behaviour Patterns. Research Challenges in Information Science (RCIS), IEEE, 2010

Chapitre 2

Analyse, propagation, collecte : un état de l'art des approches "réseau"

Sommaire

Réseaux sociaux, réseaux de communication, réseaux de collaboration, réseau d'échange, réseaux de partage, réseaux routiers, réseaux électriques, réseaux de neurones, réseaux téléphoniques, etc. Il est surprenant de constater à quel point les réseaux font aujourd'hui partie inhérente de notre quotidien. Plus surprenant encore, certains réseaux, fort de leur succès, portent parfois le nom de marque comme Facebook, Twitter ou Google+, et même parfois carrément des noms propres comme Internet. Pourtant, si le concept de réseau semble bel et bien implanté durablement dans nos sociétés, l'intérêt de la recherche pour ces objets ne s'est manifesté que durant la dernière décennie, à travers le domaine de la *science des réseaux*.

Dans ce chapitre, nous nous intéressons aux travaux menés sur les réseaux pour la compréhension des phénomènes du monde réel. Outre la présentation formelle qui est faite des réseaux et de leurs caractéristiques, nous verrons que le terme de "*science des réseaux*"

recouvre un ensemble de concepts, de formalismes et d'algorithmes mis en place pour aborder des problèmes aussi variés que la classification, la prédiction, la recherche de motifs fréquents ou l'étude des phénomènes de propagation.

Ce premier chapitre a pour objectif de présenter l'essentiel des travaux menés sur les différents axes de recherche du domaine de la science des réseaux et de détailler les notions présentes tout au long de ce mémoire. Un état de l'art orienté sur les problématiques que nous abordons est ensuite présenté dans chaque chapitre.

La Section 2.1 est consacrée à l'origine et aux caractéristiques, propriétés et structure des réseaux. La Section 2.2 s'intéresse aux méthodes d'analyse et à leur évolution. La Section 2.3 fait un état de l'art des travaux menés sur l'étude de la propagation et montre comment l'approche réseau s'est aujourd'hui imposée dans l'étude de ce phénomène. La Section 2.4 expose les travaux menés sur la collecte de données sociales, et s'intéresse plus particulièrement aux méthodes de collecte automatiques basées sur le déploiement de capteurs. La Section 2.5 vient conclure cet état de l'art.

2.1 Fondements théoriques

Si la représentation mentale d'une structure de réseau est relativement intuitive, une description formelle de ces structures s'avère nécessaire quand des études approfondies veulent être menées. Ainsi, la référence pour une description formelle des réseaux est celle de la théorie des graphes [West 2000]. Cette première section a pour objectif de présenter les principales notions et les notations issues de cette théorie, que l'on retrouve le plus fréquemment dans la littérature sur les réseaux.

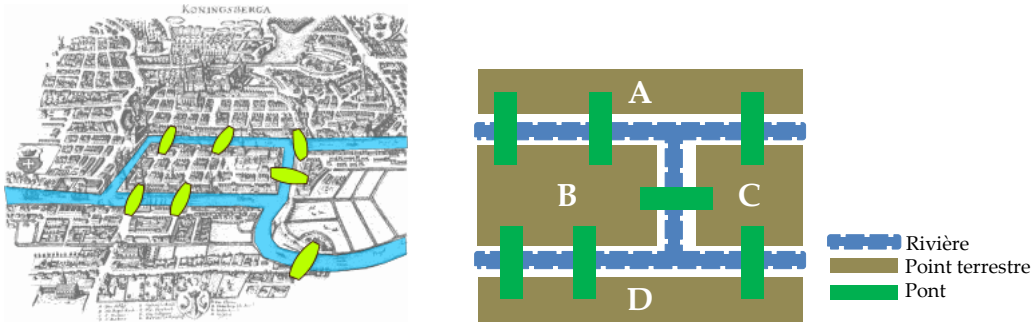
La Section 2.1.1 fait un rappel historique sur les origines de la modélisation réseau. La Section 2.1.2 décrit formellement les réseaux et leurs différentes caractéristiques. La Section 2.1.3 s'intéresse aux mesures locales et globales définies sur les réseaux. Dans la Section 2.1.4 nous présentons les principales structures topologiques observées sur les réseaux du monde réel et dans la Section 2.1.5 nous clarifions les notions de réseau social et de réseau complexe.

2.1.1 Origines

La première résolution d'un problème par un graphe a été proposée en 1741 dans un article du mathématicien suisse Leonhard Euler [Euler 1741], qui s'intéressait au problème des sept ponts de la ville de *Königsberg*[1].
Comme l'illustre la Figure 2.1, la ville de Königsberg, située en Prusse, est constituée de deux îles reliées par sept ponts. Le problème étudié par Euler consistait à trouver, à partir d'un point donné, une promenade permettant de traverser chaque pont une et une seule fois et permettant de revenir à ce point.

Pour apporter une réponse à ce problème, Leonhard Euler commença par dessiner un graphe dans lequel les terres accessibles étaient représentées par un noeud, et les ponts par des connexions entre ces noeuds. Le graphe ainsi créé contenait donc 4 noeuds (deux correspondant aux rives, et deux représentant les iles) et 7 liaisons représentant naturellement les sept ponts reliant les zones de terre (voir Figure 2.2).
Une fois une telle représentation obtenue, le problème initial se résume à : "*Partant d'un noeud donné, existe-t-il un chemin permettant de parcourir toutes les liaisons une seule fois avant de revenir au point de départ ?*".

1. Devenue aujourd'hui *Kaliningrad*



(a) (b)

FIGURE 2.1 – Les sept ponts de Königsberg
(a) Plan de la ville (source : Wikimedia) et (b) Plan simplifié

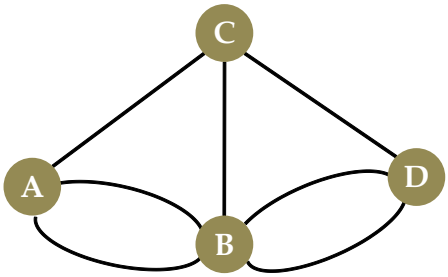


FIGURE 2.2 – Approche réseau des sept ponts de Königsberg

Un chemin passant par toutes les liaisons exactement une fois fut nommé *chemin eulérien*. Un chemin eulérien dans lequel le point de départ est le même que le point d'arrivée est appelé *circuit eulérien*. Par extension, un graphe admettant un circuit eulérien est dit *graphe eulérien*.

Euler formula ainsi l'hypothèse qu'un graphe n'est eulérien que si chaque sommet a un nombre pair de connexions. Cette hypothèse ne fut démontrée que 130 ans plus tard par le mathématicien Carl Hierholzer.
Dans le cas du problème des sept ponts de Königsberg, il devient évident que le nombre de connexions de chaque noeud étant toujours impair, il n'est pas possible depuis un point de terre visité en "*milieu*" de trajet de revenir directement à un point précédent sans réemprunter une liaison déjà utilisée. C'est donc la représentation sous forme de graphe du problème, qui permit à Euler d'affirmer qu'il n'existe pas de solution à ce problème.

Ce travail précurseur a ainsi jeté les bases de la théorie des graphes actuelle, qui est par la suite devenue l'un des axes de recherche les plus actifs dans le domaine des mathématiques discrètes. Impulsés par les travaux d'Euler, de nombreux autres problèmes ont ainsi été abordés par la théorie des graphes. Parmi les problèmes les plus courants, on retrouve par exemple le problème de l'attribution de ressources communes qui trouve classiquement des solutions avec les méthodes de *coloration de graphes* initiées par le mathématicien Francis Guthrie [May 1965]. Un autre problème largement abordé par la théorie des graphes est celui des flots dans les réseaux (liquide, circulation, transaction, etc.), dans lesquels les noeuds sont souvent soumis à des contraintes de capacité ou de production [Ostapenko 1991]. La recherche d'une distribution qui maximise ou minimise le flot dans le graphe est par exemple

un objectif courant de ces travaux.

La théorie des graphes a trouvé des applications dans des domaines divers tels que l'éthologie [Croft 2008a], la sociologie [Bott 1957], la biologie [Chavoya 2008], la géographie [Garrison 1960], la physique [Gutman 1972], les mathématiques [Erdos 2006] et l'informatique [Deo 2004]. Les travaux récents, qui tentent de regrouper les principes, les théories, les algorithmes et les mesures développés par ces différentes disciplines sont aujourd'hui identifiés comme faisant partie du domaine de la *Science des réseaux* [Borner 2007, Newman 2010].

2.1.2 Définitions et notations

Un réseau désigne généralement un ensemble d'entités (hommes, animaux, machines, cellules, etc.), que l'on nomme les *noeuds* du réseau, reliées entre elles par un ensemble de connexions appelées *liaisons* ou *liens*.

Un réseau est traditionnellement défini par un graphe $G = (V, E)$ dans lequel V est l'ensemble des noeuds et E l'ensemble des liaisons du réseau. E est un ensemble de couples de noeuds tel que $E \subseteq V \times V$.
Ainsi, soient v_i et v_j deux noeuds du réseau, $v_i, v_j \in V$, si $e = (v_i, v_j) \in E$, alors il existe une liaison entre le noeud v_i et le noeud v_j dans G. Les noeuds v_i et v_j sont dits *adjacents*, ou encore *connectés* ou *voisins*.
Le nombre total de noeuds dans le réseau est égal au cardinal de l'ensemble V et est noté $|V|$. Le nombre de noeuds $|V|$ est souvent utilisé pour désigner la *taille* du réseau.

Partant de cette définition de base, deux grandes familles de réseaux peuvent être distinguées :

- **Les réseaux non-orientés :**
  Dans un réseau non-orienté, l'ensemble des liens E est un ensemble de paires de noeuds (donc non ordonnées). Si les noeuds v_i et v_j sont connectés, il existe également un lien entre v_j et v_i (voir Figure 2.3(a)).
  Pour un réseau contenant N noeuds, c.-à-d. $|V| = N$, le nombre de liens maximal est de $\frac{N \times (N-1)}{2}$.
- **Les réseaux orientés :**
  Un réseau orienté est, lui, un réseau dont l'ensemble des liens E regroupe des couples de noeuds (donc ordonnés), appelés *liens orientés*. Dans un réseau orienté, la présence d'un lien $e_1 = (v_i, v_j)$ entre les noeuds v_i et v_j n'implique pas nécessairement l'existence d'un lien $e_2 = (v_j, v_i)$.
  Dans les représentations graphiques de tels réseaux, l'orientation du lien est généralement représentée par une flèche indiquant la direction du lien (voir Figure 2.3(b)).
  Pour un réseau contenant N noeuds, le nombre de liens maximal est de $N \times (N-1)$.

En plus de leur orientation, les réseaux peuvent posséder de nombreuses autres caractéristiques comme nous l'illustrons sur la Figure 2.3. Dans ce qui suit, nous présentons les caractéristiques les plus fréquemment décrites dans la littérature.

Réseaux unipartis

Un *réseau uniparti* est un réseau qui ne contient que des liens connectant des noeuds d'un même type. Des exemples classiques sont les réseaux sociaux liant des individus entre eux, le réseau Internet liant un ensemble de routeurs, ou le WEB qui connecte un ensemble de sites internet.
Un exemple de réseau uniparti peut être observé sur la Figure 2.3(a).

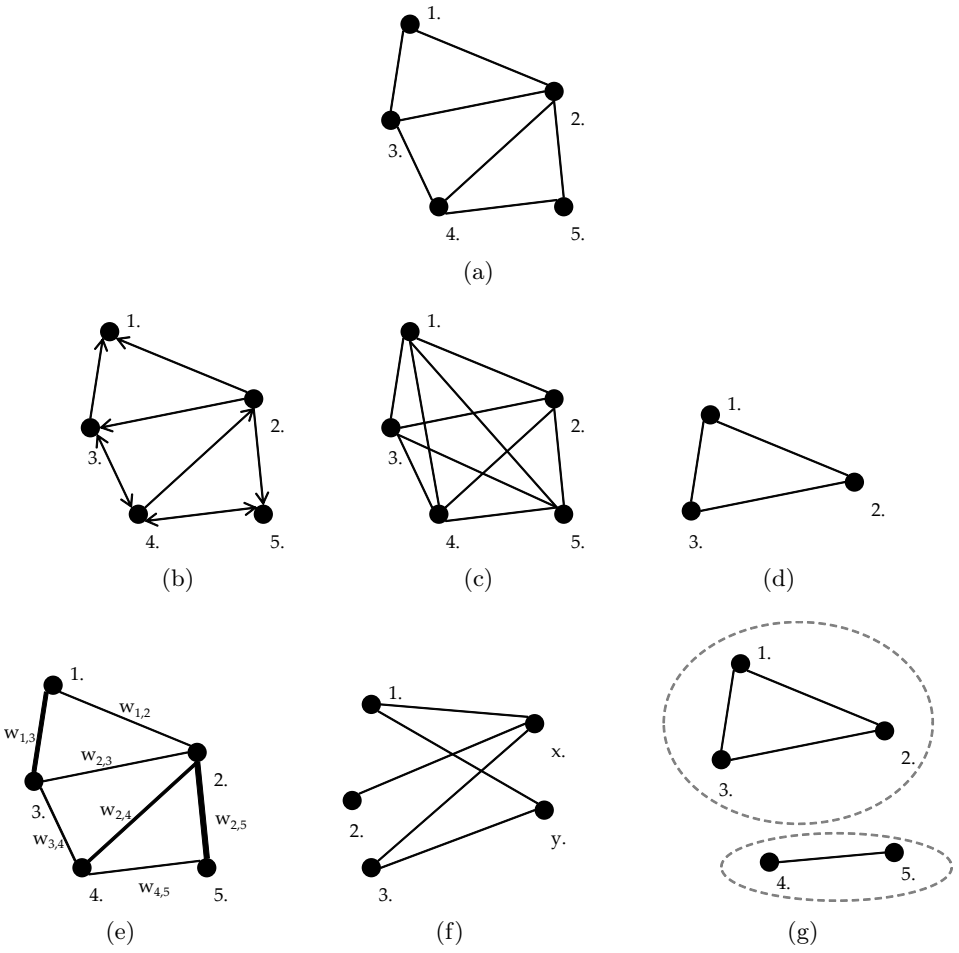


FIGURE 2.3 – Représentation graphique des différents types de réseaux
(a) Réseau non-orienté, (b) Réseau orienté, (c) Réseau complet, (d) Sous-réseau du réseau (a), (e) Réseau pondéré, (f) Réseau biparti, (g) Deux composantes connexes du réseau (a)

Réseaux complets

Un réseau uniparti dans lequel tous les noeuds sont liés entre eux, c'est-à-dire qui possède $\frac{N\times(N-1)}{2}$ liaisons dans le cas d'un réseau non-orienté ou $N \times (N-1)$ liaisons dans le cas d'un réseau orienté, est appelé *réseau complet*.
Un exemple de réseau complet peut être observé sur la Figure 2.3(c).

Sous-réseaux

Un réseau $G' = (V', E')$ est dit *sous-réseau* de $G = (V, E)$ si $V' \subseteq V$ et $E' \subseteq E$, on note alors $G' \subseteq G$.
Un exemple de sous-réseau du réseau de la Figure 2.3(a) peut être observé sur la Figure 2.3(d).

Réseaux pondérés

Un *réseau pondéré* est un réseau dans lequel chaque lien $e = (v_i, v_j)$ est caractérisé par un poids w_{v_i,v_j} qui correspond à une valeur numérique affectée au lien. Évidemment, dans un réseau non-orienté si le lien $e = (v_i, v_j)$ appartient à E, on a $w_{v_i v_j} = w_{v_j v_i}$.
Ce poids peut être soit calculé par des informations sur le graphe lui-même ou peut être obtenu à partir d'informations complémentaires. Par exemple, dans un réseau représentant les rencontres entre des individus, le poids peut être la fréquence de ces rencontres [Read 2008].
Nous montrons un exemple de réseau pondéré sur la Figure 2.3(e).

Réseaux bipartis

Certains réseaux, comme les réseaux d'achats entre des consommateurs et les produits qu'ils achètent font intervenir deux types de noeuds ; on parle alors de *réseaux bipartis*.
Plus formellement, un réseau est dit biparti s'il existe une partition de son ensemble de noeuds en deux sous-ensembles V_A et V_B telle que chaque lien du réseau ait une extrémité dans V_A et l'autre dans V_B.
Un réseau biparti est représenté par un graphe $G = (V_A, V_B, E)$ ou V_A et V_B représentent les deux ensembles indépendants et $E \subseteq V_A \times V_B$.
Un exemple de réseau biparti peut être observé sur la Figure 2.3(f). Les groupes de noeuds $\{1, 3, 4\}$ et $\{2, 5\}$ appartiennent respectivement aux ensembles V_A et V_B.

Composantes connexes

Une composante connexe C d'un réseau G est définie comme un sous-réseau connecté de G. Deux composantes connexes $C_1 = (V_1, E_1)$ et $C_2 = (V_2, E_2)$ de G sont dites *déconnectées* s'il n'existe aucun chemin reliant un noeud v_i de V_1 à un noeud v_j de V_2.
La Figure 2.3(g) illustre un exemple de deux composantes déconnectées obtenues à partir du réseau de la Figure 2.3(a).

2.1.3 Mesures

De nombreuses mesures ont été proposées pour décrire quantitativement la structure des réseaux. D'une façon générale, ces mesures caractérisent les réseaux à la fois d'un point de vue *local*, mais également d'un point de vue *global* [Boccaletti 2006, Borner 2007]. Les mesures locales s'intéressent uniquement aux propriétés des noeuds et des liens, alors que les mesures globales considèrent l'ensemble du réseau à travers des propriétés statistiques calculées sur l'ensemble de la structure.

Mesures locales

Il existe une multitude de mesures pour caractériser localement les réseaux. Ce type de mesures a pour objectif d'apporter des informations sur le voisinage d'un noeud ou de mettre en évidence certaines propriétés structurelles. Nous présentons les principales mesures utilisées localement, et pour chacune d'entre elles nous montrons son intérêt potentiel pour l'analyse des réseaux.

- **Le degré** d'un noeud v_i dans un réseau $G = (V, E)$, est le nombre de liaisons connectées à v_i. Il est noté k_{v_i}. Dans un réseau orienté, on distingue généralement le degré *entrant* k_{in} du degré *sortant* k_{out}.
  Le degré d'un noeud apporte essentiellement une information sur la connectivité

d'un noeud dans le réseau et permet de déterminer son rôle.

- **La centralité** W_{v_i} d'un noeud v_i est le degré k_{v_i} du noeud v_i, normalisé par le degré maximal potentiel, c.-à-d. $(|V|-1)$. Autrement dit, c'est le pourcentage de noeuds avec lesquels le noeud v_i est connecté.

$$W_{v_i} = \frac{k_{v_i}}{|V|-1} \tag{2.1}$$

  Cette mesure apporte également une information sur la connectivité du noeud et permet par exemple de déterminer les acteurs centraux, c'est-à-dire ceux qui sont les plus actifs et qui ont le plus de liens avec les autres noeuds.

- **La distance** d_{v_i,v_j} d'un noeud v_i à un noeud v_j est la taille du plus court chemin connectant le noeud v_i au noeud v_j. Autrement dit, c'est le plus petit nombre de liaisons nécessaires pour joindre ces deux noeuds. d_{v_i,v_j} est également appelée *distance géodésique*.
  Cette mesure fournit une indication locale entre deux noeuds et est utilisée dans le calcul d'autres mesures.

- **La distance moyenne** est une mesure individuelle de la distance qui sépare en moyenne un noeud v_i des autres $(|V|-1)$ noeuds du réseau. Ainsi, soit d_{v_i,v_j} la distance séparant le noeud v_i du noeud v_j, la distance moyenne L_{v_i} du noeud v_i est obtenue par :

$$L_{v_i} = \frac{1}{|V|-1} \sum_{j=1}^{|V|} d_{v_i,v_j} \tag{2.2}$$

  Cette mesure apporte une indication sur le rôle et l'influence d'un noeud au sein du réseau.

- **Le coefficient de clustering** d'un noeud v_i, noté C_{v_i}, est la probabilité que deux voisins v_j et v_k du noeud v_i soit eux-mêmes voisins. Considérons une disposition "*géométrique*" des noeuds. On pose t_{v_i} le nombre de triangles dont le noeud v_i fait partie. Le coefficient de clustering est donné par la formule :

$$C_{v_i} = \frac{2 \times t_{v_i}}{k_{v_i} \times (k_{v_i}-1)} \tag{2.3}$$

  D'une certaine façon, cette mesure peut être vue comme la densité locale d'un noeud. Elle permet par exemple de déterminer si un noeud est central ou périphérique. Elle est par exemple utilisée par de nombreux algorithmes d'affichage de réseaux.

Mesures globales

Contrairement aux mesures précédentes, les mesures globales décrivent l'ensemble de la structure en mettant en évidence certaines propriétés statistiques.

- **La densité** p d'un réseau $G = (V, E)$ est l'une des premières mesures utilisées pour caractériser la structure. Elle est égale au nombre de liaisons $|E|$, divisé par le nombre de liaisons possibles E_{max}. Pour un réseau non-orienté contenant $|V|$ noeuds, le nombre de liaisons possibles est donné par $E_{max} = \frac{1}{2} \times |V| \times (|V|-1)$.

$$p = \frac{|E|}{E_{max}} = \frac{2 \times |E|}{|V| \times (|V|-1)} \tag{2.4}$$

La densité apporte une information sur la connectivité globale à l'intérieur du réseau.

- **Le degré moyen** K d'un réseau $G = (V, E)$, correspond à la moyenne des degrés individuels k_{v_i} de chaque noeud v_i.

$$K = \frac{1}{|V|} \sum_{i=1}^{|V|} k_{v_i} \tag{2.5}$$

Le degré moyen est souvent comparé aux degrés individuels, pour déterminer comment un noeud donné est connecté par rapport à la moyenne. Il apporte également une information globale sur la connectivité des noeuds.

- **La distance moyenne** L, correspond à la distance moyenne séparant deux noeuds quelconque dans le réseau. Elle est obtenue, en faisant la moyenne des distances moyennes L_{v_i} obtenues en chaque noeud v_i.

$$L = \frac{1}{\frac{1}{2} \times |V| \times (|V| - 1)} \sum_{i<j} d_{ij} = \frac{1}{|V|} \sum_{i=1}^{|V|} L_{v_i}$$

Cette mesure fournit une information sur la proximité des noeuds dans le réseau ainsi que leur facilité à communiquer et échanger.

- **Le diamètre** Q d'un réseau est la plus grande des distances pouvant séparer deux noeuds au sein du réseau.

$$Q = \max \ d_{v_i, v_j} \quad \text{avec } (v_i, v_j) \in V \times V \tag{2.6}$$

Tout comme la distance moyenne, elle apporte une information sur la facilité qu'ont les noeuds du réseau à communiquer.

- **Le coefficient de clustering moyen** C correspond à la moyenne des coefficients de clustering C_{v_i} obtenus en chaque noeud v_i.

$$C = \frac{1}{|V|} \sum_{i=1}^{|V|} C_{v_i} \tag{2.7}$$

Il apporte une information sur la tendance qu'ont les noeuds à former des amas densément connectés.

- **La distribution des degrés** $P(k)$ est définie comme la probabilité qu'un noeud, choisi aléatoirement, ait un degré de k. Autrement dit, $P(k)$ représente le pourcentage de noeuds dans G ayant k connexions.

$$P(k) = \frac{|\{v_i \in V \ ; \ k_{v_i} = k\}|}{|V|} \tag{2.8}$$

Comme nous le verrons dans la section suivante, la courbe de distribution des degrés des réseaux du monde réel a souvent des formes très particulières. Cette information est utilisée pour caractériser les réseaux et comprendre comment se répartit la connectivité au sein de la structure.

D'autres mesures telles que la *distribution des distances* ou la *distribution des coefficients de clustering* peuvent également être trouvées. Finalement, qu'elles soient globales ou locales, toutes ces mesures permettent de caractériser les réseaux de façon statique, en fournissant des informations pertinentes sur l'état de leur structure topologique. Les premières méthodes d'analyse des réseaux ont exploité uniquement ce type de mesures [Bott 1957, Milgram 1967].

2.1.4 Structures et modèles de génération

L'étude détaillée des propriétés structurelles de très grands réseaux a permis de mettre en évidence les structures topologiques particulières communes aux réseaux du monde réel. De nombreux travaux récents ont en effet montré qu'au-delà de leur différence sémantique, la plupart des réseaux issus du monde réel sont caractérisés par des propriétés topologiques analogues telles qu'une distribution des degrés qui suit une loi de puissance, une distance moyenne relativement faible ou la présence de communautés. Ces caractéristiques sont radicalement différentes des celles observées classiquement sur des réseaux réguliers ou aléatoires étudiés traditionnellement dans le domaine de la théorie des graphes.
Intéressons-nous aux quatre types de structure identifiés et aux modèles permettant leur génération.

Réseaux réguliers

Les *structures régulières* sont les structures de réseau les plus simples, souvent utilisées dans des modèles d'automates cellulaires. Dans un réseau régulier, chaque noeud possède un nombre identique de liaisons. La densité du réseau est souvent faible, alors que le coefficient de clustering est, lui, relativement élevé. Il est d'ailleurs intéressant d'observer que comme la structure est régulière, le coefficient de clustering ne varie pas avec la taille du réseau. Enfin, la distance moyenne dans ces réseaux est souvent élevée, puisque la structure ne présente pas de "*ponts*" permettant de relier des individus fortement éloignés.

Il y a plusieurs façons d'obtenir de tels réseaux. L'une des plus simples consiste à disposer les noeuds équitablement sur un cercle et à créer, pour chaque noeud, des connexions avec les x premiers noeuds situés à gauche et à droite de sa position. Naturellement, pour garantir une structure régulière, x doit être identique pour tous les noeuds v_i du réseau. Ainsi, $\forall v_i \in V,\ k_{v_i} = 2 \times x$.
Dans un tel réseau, la distribution des degrés est donc définie en un seul point $k = 2 \times x$, tel que $P(k) = 1$.
Bien que ce type de structure s'observe en réalité très peu dans la nature, elle est en revanche souvent utilisée comme base pour la formation de réseaux plus réalistes.

Réseaux aléatoires

Le terme de *réseaux aléatoires* fait référence à des structures au sein desquelles l'existence d'un lien entre deux noeuds est le résultat d'un processus aléatoire. Un réseau aléatoire se caractérise par une distribution des degrés dite "*homogène*", c'est-à-dire une loi de poisson en forme de cloche. Cela traduit le fait que dans le réseau, une forte proportion de noeuds est moyennement connectée, alors qu'une plus faible proportion est fortement et faiblement connectée. On observe également que les réseaux aléatoires ont des distances moyennes relativement faibles. Cela s'explique par le fait que quand des liens sont créés aléatoirement entre des noeuds, la probabilité qu'un noeud se retrouve isolé des autres est plutôt faible.

Le modèle de génération de réseaux aléatoires le plus connu est celui proposé par Erdos et Rényi [Erdos 1960]. Dans le modèle *Erdos-Rényi*, chaque lien potentiel du réseau est créé avec une probabilité p, indépendante de l'existence des autres liens. Autrement dit, chaque couple de noeuds (v_i, v_j) a une probabilité p d'exister dans le réseau.
Une variante de ce modèle consiste à choisir uniformément au hasard un réseau G dans l'ensemble de tous les réseaux possibles contenant N noeuds.

Bien que les réseaux aléatoires aient été l'objet de recherches intensives, ce type de structure ne reproduit pas totalement les caractéristiques observées dans les réseaux du monde réel. Barabasi et Bonabeau [Barabasi 2003] expliquent par exemple que "*malgré le placement aléatoire des liens, la plupart des noeuds ont environ le même nombre de connexions. En effet, dans un réseau aléatoire, les degrés suivent une distribution de Poisson avec une forme de cloche et il est extrêmement rare de trouver des noeuds qui ont, de manière significative, plus ou moins de liens que la moyenne*".

Réseaux petit-monde

Un *réseau petit-monde* désigne à l'origine une structure dans laquelle les chemins entre deux noeuds quelconques sont généralement très courts, c'est-à-dire que la distance moyenne dans le réseau est relativement faible. Ce concept a notamment été mis en évidence par la célèbre expérience de Milgram [Milgram 1967] (détaillée dans la section suivante), qui montra que le nombre d'intermédiaires nécessaires pour atteindre deux individus dans un réseau était d'environ 6. Aujourd'hui un réseau de type petit-monde est caractérisé par deux propriétés : une distance moyenne relativement faible dans le réseau et un coefficient de clustering élevé.

Les premiers travaux à s'intéresser à la génération de réseaux petit-monde sont ceux de Watts et Strogatz [Watts 1998], qui ont proposé un modèle de génération simple, connu sous le nom de modèle *Watts-Strogatz* et basé sur l'extension d'un réseau régulier.
Comme illustrée sur la Figure 2.4, la procédure est la suivante. (1) Un réseau régulier est généré selon la méthode présentée précédemment. (2) Chaque lien subit ensuite un processus de *réécriture* selon une probabilité p_r. La procédure de réécriture consiste à remplacer un lien (v_i, v_j) par (v_i, v_k) avec k choisi aléatoirement, tel que $k \neq i$ et $(v_i, v_k) \notin E$.

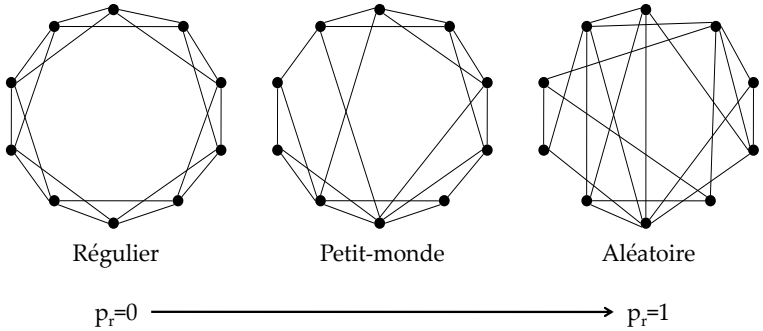


FIGURE 2.4 – Modèle *Watts-Strogatz* : D'un réseau régulier vers un réseau aléatoire

Comme nous l'avons expliqué précédemment, les réseaux réguliers présentent des coefficients de clustering et des distances moyennes relativement élevés. Ainsi, la réécriture aléatoire de quelques liens introduit des connexions entre des noeuds potentiellement situés sur de longues distances dans le réseau régulier initial, ce qui réduit considérablement la distance moyenne entre les noeuds. Quand la probabilité p_r reste relativement faible, de

nombreux noeuds conservent leurs connexions avec leurs voisins initiaux, c'est-à-dire ceux du réseau régulier. Le coefficient de clustering global reste donc relativement élevé alors que la distance moyenne est, elle, réduite, permettant ainsi l'émergence de la propriété petit-monde. Watts et Strogatz ont également montré que plus p_r se rapproche de 1, plus le réseau tend vers une structure aléatoire, puisque tous les liens sont réécrits aléatoirement.

La propriété petit-monde est aujourd'hui couramment observée dans de nombreux réseaux du monde réel. Toutefois, bien que le modèle de *Watts-Strogatz* permette de générer aisément des réseaux de type petit-monde, il a souvent été critiqué pour son incapacité à produire des noeuds possédant un degré élevé, une caractéristique souvent observée dans les réseaux du monde réel.

Réseaux scale-free

Une autre découverte fondamentale dans le domaine des réseaux a été faite en 1999 par Barabasi et Albert [Barabasi 1999], alors qu'ils étudiaient une partie du réseau de pages WEB. Ils montrèrent en effet que contrairement aux réseaux aléatoires, le réseau étudié présentait une distribution des degrés hétérogène, dans laquelle seule une faible proportion de noeuds, appelés "*hubs*", avait beaucoup plus de connexions que les autres. De telles structures ont par la suite également été observées dans de nombreux autres réseaux tels que le réseau Internet, les réseaux de citations d'articles scientifiques et certains réseaux sociaux.
Ainsi, un *réseau scale-free* est un réseau dont la distribution des degrés suit une loi de puissance. Cela se traduit par le fait que dans le réseau, une forte proportion de noeuds est faiblement connectée, alors qu'un très faible pourcentage de noeuds concentre à eux seuls un nombre élevé de connexions. D'une façon générale, la proportion $P(k)$ de noeuds dans le réseau ayant k liens peut être approché par :

$$c \times k^{-\lambda} \quad \text{avec } \lambda > 1 \tag{2.9}$$

Plusieurs modèles ont été proposés pour la génération de réseaux scale-free. Le plus connu d'entre eux est le modèle *Barabasi-Albert* [Barabasi 1999] qui introduit la notion d'*attachement préférentiel*.
(1) Un réseau initial contenant N_0 noeuds est tout d'abord crée. Dans ce réseau, nous devons garantir que $N_0 > 2$ et que $\forall v_i \in V$, $k_{v_i} \geq 1$; autrement dit, chaque noeud du réseau initial doit posséder au moins une connexion (le nombre initial de liens n'influence pas les propriétés résultantes).
(2) Une fois ce réseau initial obtenu, les actions suivantes sont itérées : un nouveau noeud v_i est ajouté au réseau et connecté à m autres noeuds v_j avec une probabilité p_j qui croît proportionnellement avec le degré k_{v_j} de v_j :

$$p_j = \frac{k_{v_j}}{\sum_{v_m \in V} k_{v_m}} \tag{2.10}$$

Naturellement, aux premières itérations, le nombre de liens est faible et la probabilité d'établir une connexion est plus ou moins équivalente pour tous les noeuds. Cependant, au fur et à mesure de l'évolution du réseau, on observe l'apparition de noeuds fortement connectés, avec lesquels un nouvel arrivant a une forte probabilité d'établir une connexion. Le mécanisme caractérisé par l'équation 2.10 est appelé *attachement préférentiel*. Il conduit à ce que les noeuds "préfèrent" établir une connexion avec les noeuds déjà fortement connectés.

Le réseau scale-free est aujourd'hui le type de réseau le plus fréquemment retrouvé dans les études menées sur les réseaux du monde réel.
Un comparatif de ces différentes structures et de certaines mesures associées, inspiré de [Borner 2007], est présenté sur la Figure 2.5.

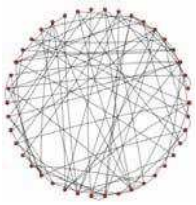
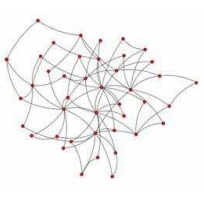
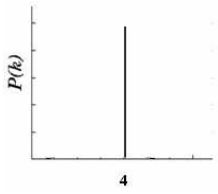
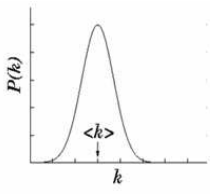
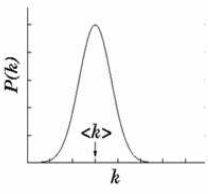
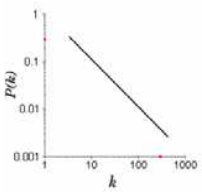
Structure	Régulier	Aléatoire	Petit-monde	Scale-free
Représentation				
#Noeuds	42	42	42	42
#Liens	84	84	84	81
Distance moy.	5.6	2.76	2.9	2.6
Coeff. Cl.	0.5	0.13	0.16	0.22
Deg. Moy.	4	4	4	3.86
Distribution des degrés				

FIGURE 2.5 – Comparatif des différentes structures et mesures associées
Source : Borner *et al.* [Borner 2007]

La structure utilisée pour le réseau petit-monde présente une courbe de distribution des degrés en forme de cloche. Cependant, la propriété petit-monde ne détermine par une forme particulière pour la courbe de distribution.
En effet, il est important de préciser que les propriétés petit-monde et scale-free ne sont pas exclusives. Certains réseaux du monde réel peuvent donc être à la fois petit-monde et présenter une distribution des degrés qui suit une loi de puissance.

2.1.5 Réseaux complexes et réseaux sociaux

Comme nous l'avons expliqué précédemment, les réseaux du monde réel peuvent être caractérisés par différents types de structure. D'une façon générale, le terme de *réseau complexe* [Albert 2002, Boccaletti 2006] est utilisé pour faire référence aux réseaux dont l'évolution conduit à l'émergence de propriétés structurelles non-triviales, telles qu'une structure petit-monde, scale-free, ou même les deux à la fois.

Les réseaux complexes sont généralement identifiés comme une sous-classe des systèmes complexes. En effet, un système complexe est en particulier un système dans lequel les interactions d'un ensemble d'entités conduisent à l'émergence d'un comportement global qui ne peut pas être déduit de leur comportement individuel. Ainsi, la "*complexité*" des réseaux ne tient pas tant de la structure en elle-même (nous avons d'ailleurs pu observer que la génération des telles structures était relativement aisée), mais vient plutôt des

difficultés qu'on rencontre à expliquer l'émergence de ces structures particulières quand on s'intéresse uniquement aux comportements individuels.

De notre point de vue, l'une des notions les plus ambiguës aujourd'hui dans la littérature sur les réseaux concerne le terme de *réseau social*.
D'un côté, l'intuition voudrait qu'un réseau social ne fasse référence qu'à un réseau possédant une sémantique sociale, c'est-à-dire des réseaux d'individus ou d'animaux liés entre eux par un ensemble de relations de natures sociales : amitiés, travail, activité commune, échange et partage, relations intimes, lien de parenté, etc.
D'un autre côté, on observe que par abus de langage, ce terme est aujourd'hui associé aux sites communautaires tels que Facebook, Twitter ou Google+.
Enfin, dans le domaine de la recherche, ce terme semble parfois être utilisé en lieu et place de "*réseaux complexes*", comme en témoigne par exemple le domaine dit de "*l'analyse des réseaux sociaux*", mais qui trouve en réalité des applications sur des réseaux de natures très différentes tels que des infrastructures matérielles de communication [Daly 2007].

Revenons sur l'évolution de la notion de "*réseau social*". Ce terme a été introduit pour la première fois en 1954, dans le domaine des sciences sociales, par un article de l'anthropologue J. A. Barnes [Barnes 1954] pour désigner un ensemble de relations entre des individus. L'objectif de Barnes était de rendre compte de l'organisation sociale d'une petite communauté, à travers l'analyse de l'ensemble des relations que ses membres entretenaient les uns avec les autres : connaissances, amis, voisins ou parents. Cette notion s'est ensuite largement répandue à l'intérieur des sciences sociales telles que l'anthropologie, la sociologie, la psychologie sociale ou l'économie, en trouvant une interprétation mathématique à travers la théorie des graphes et en donnant ainsi naissance au domaine de l'analyse des réseaux sociaux, domaine précurseur de la science des réseaux actuelle.

Aujourd'hui, il est couramment observé que de nombreux réseaux sociaux (réseaux d'amitiés, réseau technologique, réseau de collaboration, réseaux professionnels, etc.) sont également des réseaux complexes [Albert 2002, Newman 2003, Borner 2007].
Dans ce mémoire, les réseaux auxquels nous nous intéressons sont toujours des *réseaux sociaux*, dans le sens où, qu'ils soient générés ou non, ils sont assimilables à un ensemble d'interactions sociales entre des agents et présentent pour la plupart des propriétés structurelles caractéristiques.

2.2 Exploitation des données : Analyse et fouille de données sociales

Les *données sociales* font référence à toutes les données qui représentent l'activité sur un réseau social. Deux grandes familles de méthodes d'analyse de ces données peuvent être distinguées. Les *méthodes traditionnelles*, qui s'appuient uniquement sur des propriétés structurelles locales ou globales pour caractériser les noeuds et la structure, et les *méthodes d'extraction de connaissances* qui appliquent les principes de la fouille de données aux réseaux sociaux.

La Section 2.2.1 s'intéresse aux méthodes d'analyse traditionnelle et la Section 2.2.2 présente le domaine de la *fouille de données sociales*, ou "*social mining*", ainsi que les tâches associées.

2.2.1 Premières approches

L'analyse traditionnelle des réseaux sociaux a eu recours aux méthodes qui exploitent uniquement les propriétés structurelles des réseaux, obtenues à partir des indicateurs présentés en Section 2.1.3, pour hiérarchiser ou classifier les noeuds, identifier leur position et leur rôle, détecter des situations ou des configurations particulières, étudier la structure dans laquelle évolue les noeuds, etc. Parmi les travaux les plus populaires, qui ont suscité l'engouement pour l'analyse des réseaux sociaux, nous pouvons citer des exemples relativement anciens comme (i) l'analyse menée par Bott [Bott 1957] sur les familles, ou (ii) l'expérience conduite par Milgram [Milgram 1967] sur l'effet petit-monde, que nous présentons ci-dessous.

(i) Répartition des tâches domestiques. Élizabeth Bott est une psychologue canadienne qui a publié en 1957 une étude sur les relations au sein de différentes familles [Bott 1957]. Elle a proposé une *"approche relationnelle"* de la famille, selon laquelle toute famille s'insère dans un réseau social qui comprend à la fois des relations entre ses différents membres, et des relations avec des personnes extérieures. L'objectif était d'établir une corrélation entre la structure du réseau *interne*, et celle du réseau *externe*.
Dans son étude, elle s'est intéressée à un échantillon d'une vingtaine de familles londoniennes, dont elle décrit les relations entre époux, entre parents et enfants, et entre les membres de la famille et des personnes extérieures. Elle a distingué ainsi deux catégories de famille : soit le couple assume des tâches domestiques séparées, soit les activités sont effectuées ensemble par les conjoints. En étudiant la densité des réseaux impliqués, elle a mis en évidence une corrélation directe entre les relations avec l'extérieur et la répartition des tâches domestiques. En effet, Bott observe que les familles qui ont des réseaux de relations avec l'extérieur les plus denses sont celles qui adoptent le plus souvent des tâches domestiques distinctes. De cette observation naîtra la célèbre hypothèse de Bott "*le degré de séparation des tâches entre mari et femme varie dans le même sens que la densité du réseau*".
En somme, cette étude a pu montrer que le taux de répartition des activités domestiques croît avec la densité du réseau externe. Ce travail est particulièrement intéressant, car des comportements individuels ont pu être mis en évidence à partir de mesures très simples.

(ii) Expérience du petit monde. Plus populaire cette fois, l'expérience dite du *"petit monde"* [Milgram 1967] proposée par le psycho-sociologue Stanley Milgram en 1967, a pour objectif d'étudier l'hypothèse *"des six degrés de séparation"*, selon laquelle toute personne sur la planète peut être en contact avec n'importe quelle autre, au travers d'une chaine relationnelle comprenant au plus six individus.
Pour étudier cette hypothèse, Milgram a élaboré une expérience qui visait à calculer le nombre moyen de liens qui séparent une personne de n'importe quelle autre. L'expérience suggère que deux personnes, choisies au hasard parmi les citoyens américains, sont reliées en moyenne par une chaîne de six relations (voir Figure 2.6). Milgram a ainsi envoyé 60 lettres à des recrues de la ville d'Omaha dans le Nebraska. L'expérience consistait pour chacun des participants à faire passer la lettre à des connaissances personnelles, qu'il pensait être capables de faire parvenir directement ou indirectement la lettre à son destinataire.
Bien que de nombreux facteurs soient en mesure de modifier les résultats de l'expérience (groupes ethniques, statut social, catégorie socio-professionnelle, etc.), Milgram confirma l'hypothèse des six degrés de séparation en constatant qu'en moyenne cinq à sept intermédiaires furent nécessaires pour acheminer correctement les lettres. Milgram montra donc que la distance moyenne séparant deux individus était d'environ 6.
L'effet *petit-monde* est aujourd'hui couramment observé sur les réseaux du monde réel.

Les travaux plus récents sont basés sur ce même principe, c'est-à-dire le calcul de diverses

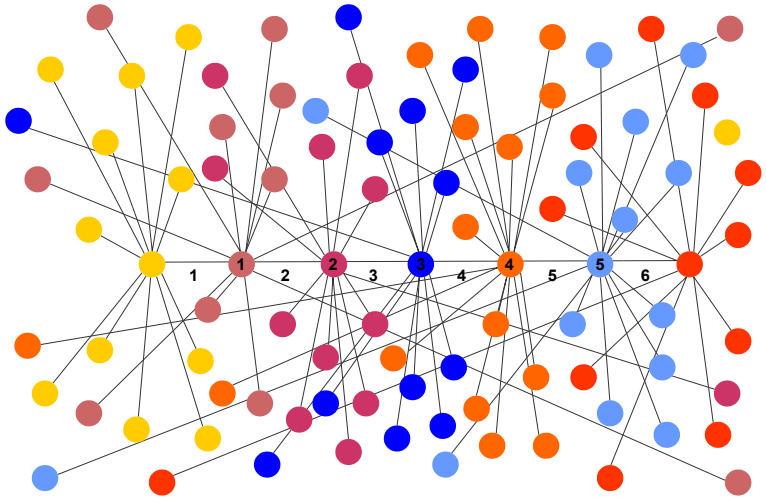


FIGURE 2.6 – Représentation du petit monde
(source : Wikimedia)

propriétés structurelles pour mettre en évidence des caractéristiques des noeuds ou de la structure. Typiquement, l'une des tâches les plus couramment abordées est celle qui consiste à classer les noeuds d'un réseau selon leur importance ou leur centralité. Par exemple, la méthode PageRank [Brin 1998] utilisée par le moteur de recherche Google, permet de classer les sites web en mesurant leur popularité.
D'autres mesures ont également été proposées, telles que *le degré de centralité*, qui permet d'identifier les noeuds les plus actifs, *la centralité d'intermédiarité*, qui mesure combien de fois un noeud se trouve sur les chemins géodésiques de tous les autres couples de noeuds, ou *la centralité de proximité*, qui identifie les noeuds les plus rapidement joignables.

Des travaux similaires ont également trouvé des applications dans le domaine de la diffusion de maladies. Par exemple, Christley *et al.* [Christley 2005] comparent plusieurs mesures pour déterminer celles qui caractérisent le mieux les individus à risque.

2.2.2 Fouille de données sociales

Les méthodes traditionnelles de fouille de données reposent sur l'hypothèse implicite selon laquelle les données sont indépendantes et identiquement distribuées (*IID*). En effet, les jeux de données classiques correspondent le plus souvent à des collections de n-uplets mutuellement indépendants. Pourtant, si cette restriction s'avère être cohérente au regard du problème classique d'inférence statistique, elle ignore toutefois l'influence des interactions entre entités dans les phénomènes étudiés.
Ainsi, l'un des défis de la recherche dans le domaine de l'étude des réseaux est de proposer de nouveaux algorithmes, ou adaptations d'algorithmes existants, capables d'extraire efficacement de la connaissance à partir de données sociales.

Getoor et Diehl [Getoor 2005] définissent la *fouille de réseaux sociaux*, comme "*l'ensemble des techniques de data mining qui considèrent explicitement les liens lors de la construction de modèles descriptifs ou prédictifs à partir de données relationnelles*". Ainsi, la fouille de données sociales aborde quatre grandes tâches que nous détaillons ci-après : (i) la classification des noeuds, (ii) l'identification de groupes, (iii) la prédiction de liens et (iv) la recherche de motifs fréquents.

(i) La classification basée sur les liens regroupe les méthodes qui se donnent pour objectif d'affecter à chaque noeud du réseau une classe. Contrairement aux méthodes traditionnelles qui se basent uniquement sur les valeurs des attributs, les méthodes de classification basées sur les liens prédisent la classe d'un noeud en intégrant les informations connues sur les noeuds, mais également sur la structure du réseau [Lu 2003]. La difficulté vient ici du fait que les classes des noeuds connectés tendent souvent à être corrélées. Les algorithmes doivent pouvoir tenir compte de ces corrélations.
La classification de pages WEB est un des exemples les plus représentatif de ce problème. L'objectif est de prédire la catégorie d'une page selon la fréquence des mots présents sur la page et celle des autres pages liées. La structure de ce réseau est aisément obtenue en analysant les liens hypertextes.

(ii) L'identification de groupes, ou **clustering sur les noeuds**, fait référence à une famille de méthodes qui a pour objectif d'identifier des groupes de noeuds qui partagent des caractéristiques communes. Dans le contexte des réseaux, les groupes sont souvent définis comme des amas de noeuds densément connectés ; on parle également de *communautés* [Fortunato 2009, Combe 2012]. Nous revenons sur ces méthodes au Chapitre 5, car elles sont en relation avec la problématique que nous abordons.
D'un point de vue pratique, ces méthodes apportent des informations pertinentes sur l'organisation de la structure du réseau. Typiquement, sur des réseaux sociaux en ligne, ces méthodes permettent d'identifier les différentes communautés d'utilisateurs. Cette information peut ensuite être utilisée pour rechercher des corrélations entre l'appartenance à une communauté et les comportements observés (opinions, achats, activités, etc.).

(iii) La prédiction de liens est un problème étroitement lié à la dynamique des réseaux. L'objectif est de prédire, entre deux états du réseau, la formation de liens entre deux noeuds [Liben-Nowell 2007]. Les algorithmes peuvent se baser uniquement sur la structure du réseau, ou prendre également en compte les attributs des noeuds. Ces méthodes peuvent aussi être utilisées pour prédire l'existence d'un lien, c'est-à-dire un lien qui n'est pas présent dans le jeu de données, mais qui existe dans la réalité.
Ces méthodes trouvent des applications dans le domaine du marketing par exemple, où la formation d'un lien entre un utilisateur et un produit peut être prédite de façon à mener des actions ciblées.

(iv) La recherche de motifs fréquents. Dans le contexte des réseaux, un motif est traditionnellement défini comme un sous-graphe. Les méthodes de recherche de motifs fréquents identifient donc les sous-réseaux qui se retrouvent fréquemment soit dans un ensemble de réseaux, ou au sein d'un unique réseau très large [Cheng 2010]. Une présentation détaillée de ces méthodes est faite dans le Chapitre 5.
Une application classique de ces méthodes concerne les réseaux de produits achetés conjointement, où les méthodes d'extraction de motifs fréquents dans les réseaux permettent de mettre en évidence les sous-ensembles de produits fréquemment achetés ensemble.

2.3 Modélisation des phénomènes de propagation

Des phénomènes de propagation peuvent être observés partout : maladie infectieuse, virus informatique, phénomène de mode, rumeur, ou plus généralement information. Pourtant, bien que ces phénomènes puissent sembler de prime abord très différents, ils sont tous des exemples types de processus qui ont pour support un réseau d'interactions entre des entités.
Ainsi, en raison de leur intérêt dans de nombreux domaines, les phénomènes de propagation ont fait l'objet de recherches actives et ont été étudiés à travers deux problèmes duaux :

(i) celui de la percolation, qui s'intéresse à la structure du support et à sa capacité à "*inonder*" un maximum d'entités, et (ii) celui de la diffusion qui se focalise sur l'évolution du processus en tentant de comprendre les différentes phases de son évolution.

Cette section présente ces deux types de problème. La Section 2.3.1 expose le problème de la percolation et la Section 2.3.2 est consacrée aux modèles de diffusion.

2.3.1 Percolation dans les réseaux

La théorie de la percolation a été introduite en 1957 par Broadbent et Hammersley [Broadbent 1957], pour analyser la pénétration d'un gaz dans un labyrinthe formé de passages ouverts ou fermés. À l'origine, l'objectif était de comprendre comment les masques à gaz des soldats devenaient inefficaces. Ces masques sont en effet constitués de petites particules de carbone poreuses qui forment un réseau aléatoire de tunnels interconnectés. Si les pores sont assez larges et suffisamment connectés, le gaz passe à travers les particules. En revanche, si les pores sont trop petits ou s'ils sont imparfaitement connectés, les émanations ne peuvent plus traverser le filtre. L'efficacité de la solution dépend donc d'un seuil critique qui est caractéristique du phénomène de percolation.

D'une façon générale, l'étude de la percolation vise à mettre en évidence les phases de transitions sur des structures aléatoires. Ces transitions sont généralement liées à la valeur critique d'un paramètre clé, appelé *seuil de transition* ou *seuil de percolation*, à partir duquel un système subit un changement brutal d'état qui permet la pénétration d'un élément.

Des cas d'études classiques sont fournis par les réseaux de communication [Hammersley 1980]. Supposons par exemple que dans un réseau téléphonique, où toutes les stations sont connectées de proche en proche, des liens soient détruits de manière aléatoire, soit à cause d'un mauvais entretien, soit par l'action d'un "saboteur stochastique". Au fur et à mesure de la suppression, il devient de plus en plus difficile de maintenir un chemin dans le réseau capable de relier deux stations données, et ce, jusqu'à ce qu'un seuil critique de suppression soit atteint qui éclaterait le réseau en plusieurs composantes. La valeur p_c, associée au pourcentage critique de liaisons actives nécessaires pour que deux points quelconques soient reliés définit le *seuil de percolation*.

Dans le contexte de l'étude des phénomènes de propagation sur les réseaux, la théorie de percolation a été utilisée pour répondre à des questions telles que : la structure du réseau permet-elle une propagation du phénomène ? ou quel pourcentage d'individus peut potentiellement être affecté ?

Il s'agit d'évaluer la probabilité d'existence d'un ensemble de liens, permettant la connexion directe ou indirecte entre deux entités du réseau. Plus précisément, on s'intéresse au seuil de paramètres critiques qui garantissent la connexité de la structure, c'est-à-dire le maintien d'une composante principale géante capable de supporter un phénomène de propagation et d'affecter un maximum de noeuds. Quand une telle composante est maintenue, on dit que le réseau "*percole*".

Soit $G = (V, E)$ un réseau aléatoire dans lequel chaque lien du réseau existe selon une probabilité p. La probabilité que ce graphe admette une unique composante connexe est appelée probabilité de percolation et est notée $\theta(p)$. Les travaux menés par Kesten [Kesten 1982] montrent qu'il existe un seuil critique $p_c = 0.5$, tel que :

$$\begin{cases} \theta(p) = 0 & \text{si } \ p < p_c \\ \theta(p) > 0 & \text{si } \ p > p_c \end{cases}$$

Un des résultats les plus intéressants de la littérature sur la percolation est apporté par Cohen *et al.* [Cohen 2000], qui s'intéressent au pourcentage critique de noeuds p_c à

supprimer aléatoirement pour déconnecter des réseaux scale-free. Ils montrent que pour des distributions de degrés suivant une loi de puissance de la forme $P(k) \approx c \times k^{-\lambda}$, avec $\lambda \leq 3$, la transition n'a jamais lieu, suggérant ainsi que le réseau possède toujours une unique composante ; autrement dit, le réseau *percole* toujours. Ces résultats viennent confirmer ceux obtenus par Albert *et al.* [Albert 2000] qui démontraient également la forte robustesse de ce type de réseau face à la suppression aléatoire de noeuds.
D'autres résultats sont en revanche venus compléter ces travaux, en montrant que si la suppression cible uniquement les noeuds les plus connectés du réseau, le pourcentage de noeuds fortement connectés à supprimer pour éclater la composante s'exprime comme une fonction de l'exposant λ. Cohen *et al.* [Cohen 2001] montrent par exemple que dans la plupart des configurations, le pourcentage de noeuds à supprimer est inférieur à 3%. Quand $\lambda > 3$, celui-ci s'abaisse à moins de 1%. Ils montrent ainsi que la plupart des réseaux du monde réel sont vulnérables à ce type d'attaques ciblées.

La percolation a souvent été définie comme le problème dual de celui de la diffusion [Newman 2003, Pajot 2001]. En effet dans le cas de la percolation, le mécanisme stochastique tient du milieu à travers lequel le processus évolue et non du processus lui-même. À l'inverse, les travaux menés sur les problèmes de diffusion s'intéressent à l'évolution aléatoire du phénomène dans un milieu, cette fois, déterministe. Le tableau de la Figure 2.7 résume ces observations.

	Percolation	**Diffusion**
Évolution de phénomène	Déterministe	Aléatoire
Structure du réseau	Aléatoire	Déterministe

FIGURE 2.7 – Dualité des problèmes de percolation et de diffusion
Source : Pajot [Pajot 2001]

Prenons l'exemple de la transmission d'une information sur un réseau de communication. Dans le cas de la diffusion, l'information quitte un point d'origine et passe à chaque instant t, d'un noeud à l'autre avec une probabilité α. Lorsque le nombre de pas tend vers l'infini, l'information a ainsi une probabilité de 1 de visiter chaque noeud du réseau si le réseau est connexe.
Dans le cas de la percolation cette fois, le mécanisme stochastique s'applique à la structure du réseau et plus à l'information. Typiquement, chaque noeud a, indépendamment des autres, une certaine probabilité d'être connecté aux autres noeuds du réseau. A chaque instant t, l'information se propage ainsi d'un noeud à l'autre avec une probabilité de 1 selon les connexions du noeud sur lequel elle se trouve. En ce sens, la propagation est entièrement déterminée par la structure du milieu. Lorsque le nombre de pas tend vers l'infini, seul un nombre fini de noeuds peut avoir connaissance de l'information selon la structure du réseau. Sur un réseau statique, la question de la diffusion ne se pose que si la percolation est assurée par la structure du réseau.

2.3.2 Approche réseau des problèmes de diffusion

Avec le développement des nouvelles technologies de l'information et de la communication (smartphones, tablettes, lunettes à réalité augmentée, etc.), l'émergence de nouveaux médias sur l'Internet (blog, site d'échanges et de partages, sites communautaires, etc.), et le large accès aux transports en commun, qui permettent aujourd'hui de diffuser une information à très grande échelle et en très peu de temps, l'étude des phénomènes de diffusion est devenue un enjeu majeur dans de multiples contextes. Typiquement, il est crucial pour

une entreprise de comprendre et de maitriser comment une nouvelle ou un récit peuvent se propager et affecter son image. De même, dans le domaine du marketing, il devient essentiel de savoir quels sont les individus à cibler pour maximiser les ventes par du marketing viral. Pour les états et les professionnels de santé, nous avons pu constater, à travers l'actualité récente (SARS, H5N1), à quel point il pouvait être important d'identifier les comportements à risque et de comprendre comment ils affectent la propagation de maladies infectieuses. Dans le domaine de la sécurité, étudier comment se propage un récit dans une population peut permettre de comprendre comment il affecte un climat social. En informatique, des questions similaires sont également abordées pour comprendre comment un virus peut perturber le bon fonctionnement d'un réseau.

Les modèles de diffusion épidémiques sont ceux qui ont reçu le plus d'attention dans la littérature pour leur intérêt dans de nombreux domaines (géographie [Eliot 2006], sociologie [Wallace 1991], biologie [Meyers 2005], mathématiques [Kermack 1927] ou informatique [Eubank 2005]), mais également pour leur facilité à s'adapter à d'autres types de phénomènes de diffusion tels que la propagation de virus informatiques [Wierman 2004] ou de rumeurs [Zanette 2002]. Dans cette section, nous présentons les deux grandes familles de modèles de diffusion : les *modèles à compartiments*, modèles mathématiques précurseurs basés sur le concept de mélange homogène, et les modèles plus récents *basés sur les réseaux*, qui visent à intégrer la complexité des interactions humaines impliquées.

Les modèles à compartiments

Les modèles à compartiments considèrent une population comme un ensemble de groupes (c'est-à-dire des *compartiments*), caractérisés chacun par l'état des individus au regard de l'épidémie. Dans ce type de modèle, on suppose que les individus au sein des différents compartiments changent de compartiments de façon homogène ; on parle de *mélange homogène* ou d'*action de masse*. D'une certaine façon, cette approche suppose que les individus au sein d'un même compartiment entretiennent une structure relationnelle régulière avec les individus des autres compartiments. Intéressons-nous aux principaux modèles.

Le modèle *SI* (*Susceptible-Infected*) est l'un des modèles épidémiques les plus simples. Dans ce modèle, deux groupes d'individus peuvent être identifiés : les *susceptibles* (S) et les *infectés* (I). Les *susceptibles* sont les individus qui peuvent contracter la maladie s'ils entrent en contact avec des individus infectés. Les *infectés* sont, eux, des individus porteurs de la maladie, qui peuvent la transmettre lors de contacts avec des susceptibles. Les individus infectés ont une probabilité α d'établir un contact avec un individu susceptible. Cette approche simpliste modélise la propagation de la maladie par le passage de l'état *susceptible* à l'état *infecté*. Ce modèle suppose qu'un individu *infecté* reste dans cet état (voir Figure 2.8).

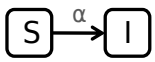


FIGURE 2.8 – Modèle de diffusion épidémique *SI*

Plus formellement, soit S le nombre d'individus susceptibles et X le nombre d'individus infectés. La taille n de la population est obtenue par $n = S + X$.
Posons α la probabilité qu'un individu infecté établisse un contact avec un individu susceptible. La maladie ne se propage que lorsqu'un tel contact est établit. Ainsi, si la population est composée de n individus, la probabilité qu'une personne rencontrée aléatoirement soit susceptible est de $\frac{S}{n}$. Par conséquent, chaque individu *infecté* a potentiellement, par unité

de temps, $\alpha \times \frac{S}{n}$ contacts avec une personne *susceptible*. Puisqu'il y a X individus infectés à chaque unité de temps, le nombre de nouvelles infections peut être estimé par $X \times \alpha \times \frac{S}{n}$, ce qui nous permet d'exprimer le taux de changement du nombre d'individus *infectés* par l'équation différentielle :

$$\frac{dX}{dt} = X \times \alpha \times \frac{S}{n} \tag{2.11}$$

De la même façon, le taux de changement du nombre d'individus *susceptibles* peut s'exprimer par l'équation différentielle :

$$\frac{dS}{dt} = -S \times \alpha \times \frac{X}{n} \tag{2.12}$$

Ainsi, en posant $s = \frac{S}{n}$ et $x = \frac{X}{n}$, nous obtenons :

$$\frac{ds}{dt} = -\alpha \times s \times x, \qquad \frac{dx}{dt} = \alpha \times s \times x \tag{2.13}$$

En posant, $s = (1 - x)$, nous obtenons $\frac{dx}{dt} = \alpha \times (1 - x) \times x$ et

$$x(t) = \frac{x_0 e^{\alpha \times t}}{1 - x_0 + x_0 e^{\alpha \times t}} \tag{2.14}$$

ou x_0 est la valeur de x à l'instant $t = 0$.

Comme le montre la Figure 2.9, ce type de modèle produit des courbes d'infection avec un point d'inflexion, pour lesquelles nous pouvons observer que le nombre d'individus infectés connait une forte croissance pendant les premiers instants. Cela correspond à la phase initiale du processus où la plupart des individus sont encore susceptibles. Après cette étape, le nombre d'individus infectés croît toujours, puisqu'aucun individu n'est supposé sortir de son état d'infection, mais la croissance diminue, jusqu'à affecter ainsi l'ensemble de la population.

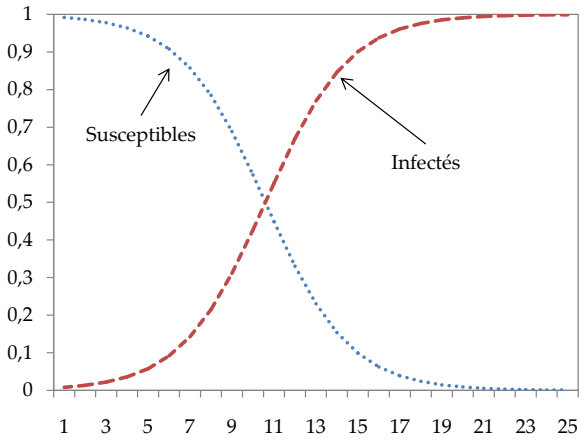


FIGURE 2.9 – Courbes de diffusion dans le modèle *SI*

Le modèle *SIR* (*Susceptible-Infected-Recovered*) est également l'un des modèles de diffusion épidémique les plus simples et les plus fréquemment retrouvés dans la littérature. Il constitue une évolution directe du modèle *SI*. Dans le modèle *SIR*, les deux états du modèle *SI* sont conservés (*Susceptibles* et *Infected*), auquel est ajouté le troisième état R (*Recovered*), qui n'est atteint que par les individus infectés selon une certaine probabilité β

(voir Figure 2.10). Dans ce modèle, le terme "*Recovered*" indique que l'individu est sorti de son état d'infection et ne peut plus contracter la maladie, soit parcequ'il devient immunisé, soit à la suite d'un décès. Ce modèle suppose qu'un individu dans l'état R conserve son immunité. Par conséquent, un individu *recovered* ne peut pas être de nouveau *susceptible* ou *infecté*.

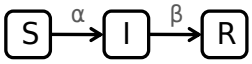


FIGURE 2.10 – Modèle de diffusion épidémique *SIR*

Le modèle est basé sur deux paramètres : la probabilité α qu'un individu *infecté* établisse un contact avec un individu *susceptible* et le taux de rétablissement β. Comme dans le modèle *SI* classique, un individu *susceptible* contracte la maladie quand un contact est établi avec un individu infecté selon la probabilité α. De plus, tout individu infecté à une probabilité β d'atteindre l'état *Recovered*. Ainsi, sur le même principe que le modèle *SI*, les taux de changement peuvent s'exprimer par les équations différentielles suivantes :

$$\frac{ds}{dt} = -\alpha \times s \times x, \qquad \frac{dx}{dt} = \alpha \times s \times x - \beta \times x \qquad \frac{dr}{dt} = \beta \times x \tag{2.15}$$

Sur la Figure 2.11, nous présentons des exemples de courbes de diffusion obtenues avec le modèle *SIR*. Nous pouvons observer que la courbe représentant le nombre d'individus infectés (on parle également de *courbe d'incidence*) est en forme de cloche. Cela signifie que le nombre d'individus infectés croît jusqu'à atteindre une valeur maximale, avant de connaitre une phase de décroissance.

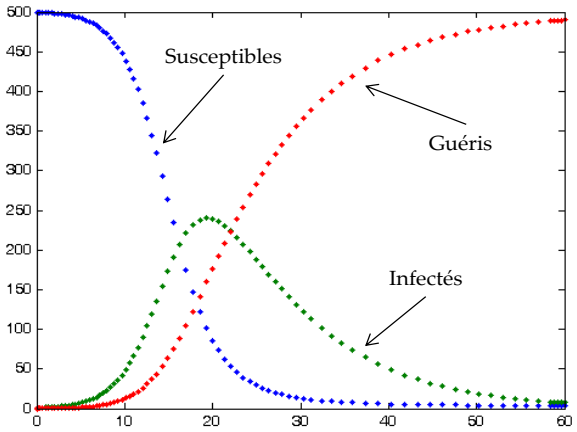


FIGURE 2.11 – Courbes de diffusion dans le modèle *SIR*

Basés sur le même principe, on peut également retrouver dans la littérature le modèle *SIS* (*Susceptible-Infected-Susceptible*), qui est également une évolution directe du modèle *SI* et qui prend en compte la possibilité de ré-infection des individus, ou le modèle *SIRS* (*Susceptible-Infected-Recovered-Susceptible*) qui combine les modèles *SIR* et *SIS* en autorisant la ré-infection des individus après une phase d'immunité plus ou moins longue.

Enfin, des modèles beaucoup plus élaborés ont également été proposés, tels que les modèles *SEIS*, *SEIR*, *MSIR*, etc., qui représentent des situations plus complexes, comme celle des individus qui ne sont pas totalement guéris, mais qui continuent de propager

la maladie. Une présentation détaillée de ces différents modèles mathématiques peut être trouvée dans [Easley 2010, Newman 2010].

Les modèles basés sur les réseaux

Bien que les modèles à compartiments aient été largement utilisés dans l'étude des phénomènes de diffusion, l'hypothèse selon laquelle les individus ont une même probabilité d'établir des contacts s'avère être irréaliste. En effet, dans la réalité les contacts qu'entretiennent les individus sont souvent hétérogènes, puisque les individus ne sont généralement connectés qu'à une petite proportion d'individus et cette proportion n'est jamais choisie aléatoirement.

C'est en 1985 qu'a été utilisé pour la première fois le concept de réseau social pour étudier la diffusion du SIDA [Klovdahl 1985]. Dans ce travail, Klovdahl montre la pertinence d'une approche réseau dans le suivi de la transmission de l'agent infectieux et la mise en place de stratégies visant à réduire sa propagation. Ce travail précurseur pose ainsi les bases de l'étude des phénomènes de diffusion par la modélisation réseau, en mettant en avant le caractère central des relations entretenues par les entités dans le processus de transmission. En effet, prenons par exemple le cas de la propagation de maladies infectieuses telles que la grippe ou les maladies sexuellement transmissibles. Il est aujourd'hui reconnu que les contacts de proximité ou les relations intimes qu'entretiennent les individus sont les principaux vecteurs de transmission. Cette hypothèse se vérifie également lorsque l'on s'intéresse aux autres types de diffusion. Alors qu'une information sera naturellement véhiculée à travers un réseau de communication, voire même de l'infrastructure matérielle sous-jacente, la propagation d'une rumeur sera elle, influencée par les liens d'amitié, de croyance ou d'influence que maintiennent les individus entre eux.

Ainsi, contrairement aux modèles à compartiments, les modèles à base de réseau tentent de prendre en compte l'hétérogénéité des interactions humaines en se basant sur l'hypothèse que la structure et la nature du réseau dans lequel évoluent les individus sont les principaux facteurs déterminant le comportement du processus.

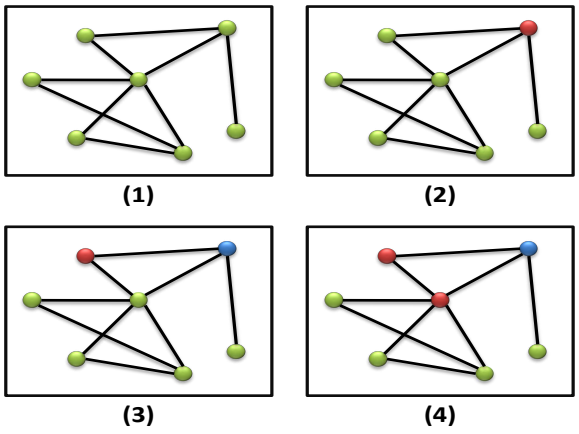


FIGURE 2.12 – Modèle *SIR* sur un réseau
Vert : Susceptibles, Rouge : Infectés, Bleu : Guéris

La plupart des modèles classiques de diffusion (*SI*, *SIR*, *SIS*, *SIRS*, etc.) ont donc été adaptés aux réseaux. Prenons le cas du modèle *SIR* (voir Figure 2.12) qui est aujourd'hui l'un des plus utilisés [Christley 2005, Read 2008, Salathe 2010a] sur les réseaux.

Soit $G = (V, E)$ un réseau avec V, l'ensemble des noeuds et E l'ensemble des liens.
On pose $F_t : V \to \{S, I, R\}$ la fonction qui renvoie, à chaque itération t, l'état d'un noeud v_i.
On définit également N_t^i comme étant le nombre de voisins infectés du noeud v_i à l'itération t, c.-à-d. $N_t^i = |\{v_j \in V; (v_i, v_j) \in E \text{ et } F_t(v_j) = I\}|$
Une fois qu'un premier noeud est infecté, à chaque itération t :

- Chaque noeud susceptible v_i a une probabilité $1 - (1 - \alpha)^{N_t^i}$ d'être infecté, ou α est la probabilité de transmission par contact. Ainsi plus un noeud possède de voisins infectés, et plus sa probabilité de devenir lui-même infecté augmente.
- Un noeud infecté passe à l'état *Recovered* avec une probabilité β

L'étude des phénomènes de diffusion à travers l'approche réseau à fait l'objet d'une activité de recherche intense. Les problèmes couramment abordés sont : (i) La compréhension des phénomènes, qui vise à étudier le comportement du processus de diffusion selon différentes configurations. (ii) L'identification de situations à risque, avec un intérêt particulier porté aux facteurs favorisant l'émergence et l'évolution du phénomène ; les résultats obtenus sont souvent utilisés pour mettre en place des stratégies d'intervention adaptées. (iii) La recherche de motifs, qui tente de mettre en corrélation les attributs des noeuds et les tendances observées.

(i) Comprendre les phénomènes. Plusieurs travaux se sont intéressés à l'effet des structures topologiques sur le processus de propagation. Nous avons déjà pu observer l'intérêt porté aux composantes du réseau pour comprendre si la structure permet la propagation [Cohen 2000, Cohen 2001].

D'autres travaux se sont intéressés à l'impact de la distribution du degré sur le processus. Par exemple, en comparant la diffusion sur un réseau aléatoire et sur un réseau scale-free, Lloyd et May [Lloyd 2001] montrent que pour des processus dont le taux de transmission par contact est faible (paramètre α dans les modèles *SI*, *SIR*, etc.), seul le réseau scale-free produit un pic épidémique. Cela s'explique par la présence d'individus fortement connectés (*hubs*) dans un réseau scale-free, qui garantissent que même une maladie relativement bénigne se propagera largement si un *hub* est infecté.

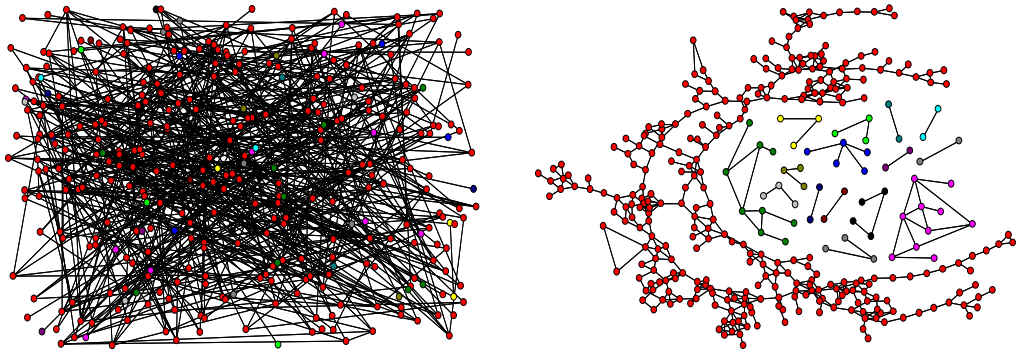
Des résultats similaires sont obtenus sur les réseaux petit-monde. En effet, Watts et Strogatz [Watts 1998] montrent que des probabilités très faibles de réécriture aléatoire des liens (voir modèle de Watts-Strogatz présenté dans la Section 2.1.4 pour la génération de réseaux petit-monde), suffisent à réduire de façon significative le taux de transmission par contact nécessaire pour obtenir un pic épidémique.

Le diamètre et la distance moyenne dans le réseau offrent également d'autres indicateurs sur la capacité de l'épidémie à traverser plus ou moins facilement la structure, et donc à infecter un maximum d'individus.

Un autre moyen simple d'aborder le problème consiste à représenter efficacement le réseau. Il ne s'agit cependant pas de proposer des représentations agréables pour les yeux, mais plutôt de mettre en évidence des éléments utiles pour que le cerveau puisse identifier les caractéristiques structurelles pertinentes.

L'étude de méthodes d'affichage efficaces est un problème bien connu de la théorie des graphes, appelé *graph layout problem*, et pour lequel de nombreux algorithmes ont été proposés. Un des plus connus est l'algorithme *Spring Embedding* [Quinn 1979] proposé par Quinn et Breuer, qui considère les noeuds comme reliés par des ressorts pouvant se repousser ou s'attirer. Nous pouvons également citer l'algorithme Kamada-Kawai [Kamada 1989] qui tente d'éviter le chevauchement des liens ; ou encore l'algorithme Fruchterman-Rheingold [Fruchterman 1991] qui cherche à rapprocher les noeuds appartenant à une même communauté.

Dans l'exemple proposé sur la Figure 2.13, nous montrons la représentation d'un réseau



(a) (b)

FIGURE 2.13 – Visualisation pour la compréhension de la structure
(a) Disposition aléatoire des noeuds, (b) Après application de Spring Embedding

selon un placement aléatoire, puis après l'application de l'algorithme *Spring Embedding*. A partir de cette vue du réseau, deux observations utiles concernant le phénomène de diffusion peuvent être faites. La première est que le réseau contient plusieurs composantes, donc si une personne est infectée, la maladie ne peut se propager à partir d'elle, qu'au sein de sa propre composante. On peut également observer que les noeuds du réseau sont globalement très peu connectés et que la distance moyenne est élevée, ce qui complique l'évolution du processus, puisque pour infecter les individus la maladie doit pouvoir traverser le réseau.

(ii) Identifier des situations à risques. Une des tâches classiques abordées dans les travaux sur la diffusion consiste en l'identification de situations à risque pour la mise en place de stratégies d'intervention visant à réduire l'évolution du processus. Ces situations sont souvent caractérisées par le rôle que jouent les individus dans le réseau.

Par exemple, Christley *et al.* [Christley 2005] comparent plusieurs mesures locales (degré, centralité, distance moyenne, etc.) et montrent que le degré est la mesure la plus discriminante pour l'identification d'individus à risque. De même, Pastor-Satorras et Vespignani [Pastor-Satorras 2001] comparent une vaccination aléatoire des noeuds, à une vaccination ciblée selon la connectivité des noeuds et montrent que la vaccination ciblée sur les noeuds les plus connectés réduit considérablement la vulnérabilité d'un réseau scale-free face aux épidémies. Le même constat est également fait par Dezsho et Barabasi [Dezsho 2002]. D'une certaine façon, ces résultats sont une conséquence des études menées sur la percolation qui montrent que les réseaux scale-free sont vulnérables aux attaques ciblées sur les noeuds les plus connectés. Par conséquent, quand ces noeuds sont vaccinés contre l'épidémie, la propagation est naturellement réduite.

D'autres travaux se sont intéressés à des indicateurs globaux. Par exemple, l'étude menée par Salathe et Jones [Salathe 2010a] s'intéresse aux réseaux possédant plusieurs communautés. Ils montrent que sur de tels réseaux, des stratégies d'interventions ciblant des individus faisant le pont entre plusieurs communautés sont plus efficaces que celles qui s'intéressent uniquement aux individus les plus fortement connectés. Un exemple de cette situation est illustré sur la Figure 2.14, où l'on observe les individus faisant le pont entre plusieurs communautés en rouge.

La plupart des stratégies d'intervention considère l'ensemble de la structure du réseau. Typiquement, l'identification des individus les plus connectés nécessite une connaissance

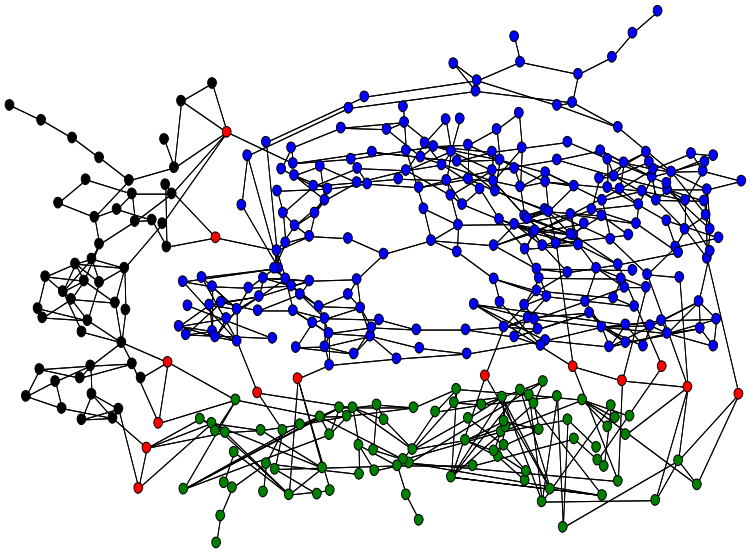


FIGURE 2.14 – Individus faisant le pont entre plusieurs communautés

globale du réseau et des connexions de chaque noeud ; une information qui peut être très difficile à obtenir dans la réalité. Certains travaux ont proposé des méthodes ne nécessitant qu'une connaissance partielle du réseau. C'est par exemple le cas des travaux menés par Christakis and Fowler [Christakis 2010] qui proposent une stratégie d'intervention alternative qui ne requiert qu'une connaissance du voisinage de certains individus choisis aléatoirement. Cette méthode est basée sur un postulat simple : "*les voisins d'un noeud ont plus de voisins que lui*". Par conséquent, ils supposent que les voisins d'un individu choisi aléatoirement seront infectés avant lui, et ils les utilisent ainsi comme "*capteurs*" pour la détection précoce de l'épidémie.

(iii) Rechercher des motifs. Des motifs complexes peuvent également être mis en évidence. En effet, quand ils sont disponibles, il est possible d'intégrer les propriétés des noeuds (age, sexe, nationalité, etc.) dans l'analyse, pour déterminer si elles contribuent, d'une façon ou d'une autre, à l'évolution du phénomène.

L'une des méthodes consiste par exemple à rechercher des corrélations entre les propriétés des noeuds (topologiques, individuels et démographiques) et l'évolution du processus de diffusion. En effet, les individus connectés entre eux dans un réseau ont souvent des caractéristiques similaires, formant ainsi des groupes densément connectés d'individus aux caractéristiques plus ou moins analogues. Ces groupes peuvent être mis en évidence par les méthodes d'affichage. Concrètement, un noeud peut être représenté selon la valeur de ses attributs en modifiant par exemple sa couleur, sa forme ou sa taille. Ce type de représentations a été utilisé dans diverses études menées sur des animaux. Croft *et al.* [Croft 2008a] utilisent par exemple une représentation visuelle dans une étude consacrée à une population de Guppys sauvages de Trinidad. Deux attributs sont ainsi considérés : le sexe, qui détermine la forme du noeud (cercle vide pour les femelles et cercle plein pour les mâles) et la taille du corps, qui détermine la taille du noeud. Ils mettent ainsi en évidence un cluster de cercles ouverts, c'est-à-dire de femelles, qui apparaissent comme les individus centraux du réseau.

Dans le domaine de la diffusion, ces représentations ont été utilisées pour identifier des

motifs pertinents et rechercher leurs implications sur le processus. Par exemple, l'étude menée par Christakis *et al.* [Christakis 2007] sur la diffusion de l'obésité utilise une représentation dans laquelle les noeuds sont représentés par des cercles vides de couleur rouge pour les femmes et de couleur bleue pour les hommes. La taille de chaque noeud est proportionnelle à leur indice de masse corporelle et la couleur à l'intérieur de chaque cercle indique si l'individu est obèse ou pas (jaune : obèse, vert : non-obèse). Ils observent ainsi des clusters de personnes obèses et montrent que la probabilité qu'une personne devienne obèse augmente de 57% si elle fait partie d'un tel cluster.

2.4 Collecte automatique de données sociales

La collecte de données sociales fait référence à toutes les méthodes qui se donnent pour objectif de collecter des interactions sociales entre des entités. Comme nous l'avons présenté dans le Chapitre 1, ce problème est un axe fondamental de la recherche sur les réseaux. En effet, en raison des difficultés que rencontrent les scientifiques à obtenir et à exploiter des données réelles, les travaux ont souvent eu recours à des modélisations pour aborder les phénomènes portés par les réseaux. La collecte de données réelles constitue aujourd'hui un enjeu majeur pour la confrontation, la validation ou l'amélioration des modèles existants.

Cependant, collecter les données qui témoignent d'interactions survenant entre des entités dans la réalité n'est pas une tâche facile. Le type d'interaction à collecter dépend du type d'entités, de leur environnement, et du phénomène que l'on souhaite étudier. La collecte de données soulève donc des problèmes variés touchant à la fois à la complexité des méthodes à mettre en place, et aux types d'interaction qu'elles sont en mesure de collecter.

Dans cette section, nous nous intéressons aux méthodes de collecte, et plus particulièrement aux méthodes récentes qui tirent parti de l'amélioration des performances des microcontrôleurs pour une collecte automatique des données sociales.

2.4.1 Collecte à partir de données en ligne

Avec le développement du WEB 2.0 et la multiplication des sites d'échanges et de partages tels que les forums, les blogs, les sites communautaires ou les sites de e-commerce, divers jeux de données ont pu être générés à partir des interactions crées via ces médias et de l'activité des entités concernées. Les différents réseaux pouvant être obtenus à partir de ces sites sont les suivants :

- **Réseaux d'intérêt :** Réseaux bipartis créés à partir de forums et impliquant les utilisateurs et leurs sujets d'intérêt. Par projection, ce réseau peut ensuite être utilisé pour générer un réseau d'utilisateurs, dans lequel deux utilisateurs sont connectés s'ils partagent les mêmes sujets d'intérêt.
- **Réseaux de blogs :** Les articles publiés sur les blogs font souvent référence à d'autres blogs par le biais des liens hypertextes. Ces informations peuvent être utilisées pour générer des réseaux de blogs connectant deux blogs quand l'un fait référence à l'autre.
- **Réseaux d'amitié/intérêt :** À l'origine, les sites communautaires tels que Facebook, Twitter ou Google +, permettaient d'obtenir des informations sur les liens d'amitié ou de connaissance entretenus par les individus. Aujourd'hui, face à la diversité d'information présente sur ces sites (individu, entreprise, parti politique, artiste, produit, évènement, etc.), nous observons que la sémantique du lien s'assimile désormais à "*porter un intérêt à*".
- **Réseaux d'achats :** Trois types de réseaux peuvent être obtenus à partir des sites de commerce en ligne. (i) Des réseaux bipartis, impliquant les consommateurs aux pro-

duits qu'ils achètent. (ii) Des réseaux de consommateurs, connectant deux consommateurs quand ils ont acheté un même produit. (iii) Des réseaux de produits, connectant deux produits quand ils sont achetés conjointement par un même consommateur.

- **Réseaux de co-auteurs :** Le cas le plus répandu concerne les bases de données d'articles scientifiques. Par exemple, en utilisant les données issues de DBLP [2], un réseau de co-auteurs peut être obtenu, dans lequel deux auteurs sont connectés s'ils sont co-signé un même article.

Le principal inconvénient de ces approches réside dans le fait qu'elles ne permettent d'obtenir qu'un type particulier d'interactions, sémantiquement lié au média utilisé. Les réseaux extraits de ces sources d'informations sont par exemple difficilement exploitables dans le cadre de l'étude de la diffusion d'une épidémie ou d'une information. Dans le cas de la diffusion de certaines maladies infectieuses telles que la grippe, les contacts de proximité géographique entretenus par les individus lors de leurs déplacements ou dans les lieux qu'ils fréquentent ne peuvent être extraits à partir d'aucune source d'information connue. C'est la raison pour laquelle des méthodes de collecte alternatives ont été proposées.

2.4.2 Microcontrôleurs pour la collecte de données sociales

Aujourd'hui, l'évolution des moyens techniques et notamment le fort développement que connaissent les micro-dispositifs, capables de capturer des données sonores, visuelles ou spatio-temporelles, a récemment permis d'entrevoir de nouvelles possibilités dans la collecte de données sociales. Les microcontrôleurs sont composés de circuits intégrés, qui rassemblent les éléments fondamentaux d'un ordinateur : un processeur, une mémoire morte pour le programme, une mémoire vive pour les données et des interfaces d'entrées-sorties permettant l'interaction avec l'extérieur. Les microcontrôleurs présentent des capacités limitées en mémoire et en calcul et sont fréquemment utilisés pour la conception de périphériques divers : capteurs (voir Figure 2.15), téléphones portables, tablettes, dispositifs de localisation (GPS) ou puces RFID.

FIGURE 2.15 – Exemple de microcontrôleur : capteur MicaZ

Malgré leurs limitations, ils ont depuis quelques années connu un succès grandissant qui s'explique par différents facteurs. Tout d'abord, la miniaturisation extrême des composants permet d'obtenir des dispositifs de plus en plus petits. Leur coût de fabrication peu élevé et leur capacité à communiquer entre eux facilitent leur utilisation à grande échelle lors

2. Site référençant les publications en informatique

d'expériences scientifiques d'envergure ou dans des domaines pour lesquels une observation humaine permanente est requise.

Ainsi, les perspectives intéressantes offertes par les microcontrôleurs ont permis leur utilisation récente pour la collecte de données sociales. Des travaux récents ont en effet eu recours à ces dispositifs pour collecter les traces de différents types d'interactions sociales entre des humains ou des animaux. Nous distinguons essentiellement deux types de dispositifs de collecte.

Dispositifs mobiles

Les dispositifs mobiles sont les plus répandus dans le domaine de la collecte de données sociales. Ils font référence à des dispositifs placés sur les agents et pouvant se déplacer dans l'espace : capteurs mobiles, téléphones portables, tablettes, collier GPS ou puces RFID. Les données collectées sont souvent des données sur la position géographique et les dispositifs situés à proximité. Deux types de configurations peuvent être envisagés. (i) Soit les dispositifs sont capables d'envoyer directement leur information à un serveur central. C'est par exemple le cas des téléphones portables, des tablettes, ou des colliers GPS qui peuvent utiliser le réseau Internet, et plus particulièrement la technologie 3G, pour acheminer les données collectées. (ii) Soit les dispositifs sont organisés en réseau (capteur sans-fils ou puce RFID) et communiquent leurs informations à des antennes relais, chargées d'acheminer les données collectées vers le serveur (voir Figure 2.16).
Cette méthode présente l'avantage de fournir directement des données individuelles, puisque le dispositif est rattaché à un individu unique, le "*porteur du capteur*".

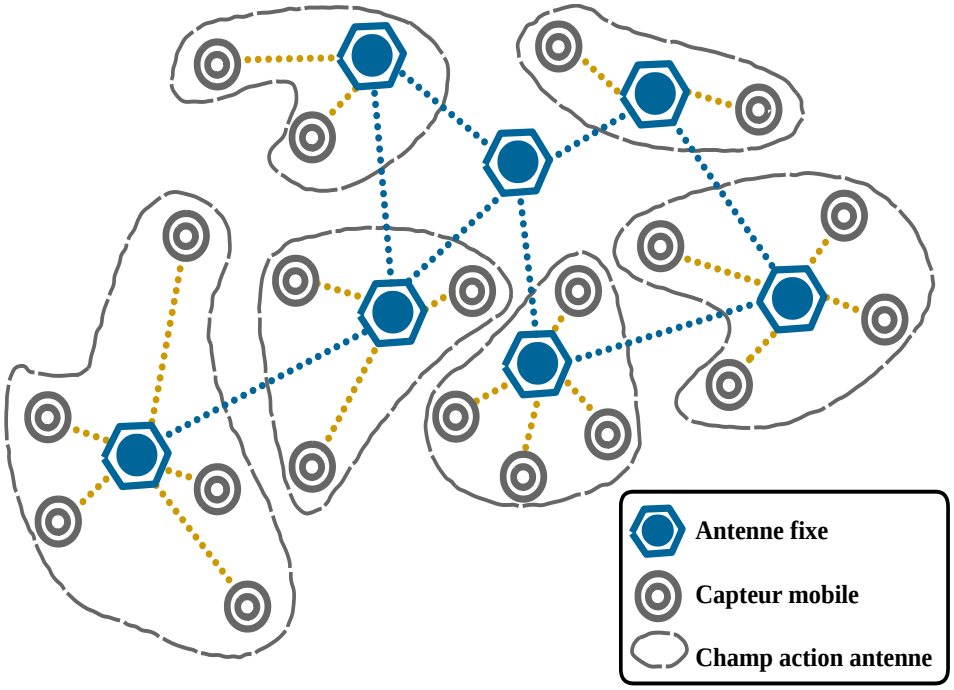


FIGURE 2.16 – Architecture réseau de capteurs mobiles

Par exemple, Stehle *et al.* [Stehle 2011] proposent une architecture de capteurs mobiles pour collecter les interactions sociales au sein d'une école et étudier comment se propage une

maladie sur la population observée. Cette solution a été utilisée dans une école primaire de Lyon, dans laquelle 232 enfants et 10 enseignants ont été équipés de puces RFID. Il était demandé aux participants de porter les badges au niveau de leur poitrine, de façon à ce qu'un contact de proximité ne soit détecté que si deux individus sont face à face, dans un rayon inférieur à $1,5m$ et durant une période d'au moins $20s$. Les données étaient collectées par des antennes relais disposées au sein de l'école, puis agrégées et sauvegardées sur un serveur local, qui construisait le réseau de contacts en temps réel. 77602 contacts de proximité ont ainsi été collectés.

L'expérience menée par Laibowitz *et al.* [Laibowitz 2006] est très similaire. Ils ont conçu des capteurs sous forme de badges et les ont utilisé lors d'une conférence pour collecter les interactions sociales entre les participants. Contrairement à l'étude de Stehle, leurs badges étaient équipés de périphériques supplémentaires leur permettant d'afficher des messages, de réaliser des enregistrements sonores ou de mesurer l'accélération du porteur.

Olguin et Pentland [Olguin 2008] ont proposé des capteurs beaucoup plus perfectionnés pour la détection d'interactions sociales. Les capteurs, qu'ils qualifient de "*capteurs sociométriques*", sont capables d'identifier des interactions de proximité, des comportements liés aux déplacements (une personne avec une forte activité ou une activité faible), d'extraire des caractéristiques dans la voix pour identifier l'intérêt ou l'excitation et la position géographique. Ils ont utilisé ces capteurs dans plusieurs études, dont une qui a consisté à générer un réseau social à partir des contacts de proximité au sein d'un hôpital. L'objectif était d'identifier les goulots d'étranglement potentiel et les défaillances dans la gestion des agents à partir de l'analyse du réseau obtenu.

Dispositifs fixes

Les dispositifs fixes font référence à des capteurs figés dans l'espace, capables d'effectuer des relevés localement ou dans leur environnement proche. Les capteurs sont disposés sur la zone que l'on souhaite étudier et fournissent des informations à un serveur central qui est chargé de réaliser les analyses nécessaires. Les capteurs sont le plus souvent équipés d'au moins un périphérique, capable d'effectuer des relevés de natures diverses : sons, image, vidéo, luminosité, température, humidité, détecteur de mouvement, etc. Quand l'espace à couvrir est étendu, ils peuvent également communiquer entre eux pour s'échanger des informations et les acheminer vers le serveur central ; on parle alors de *réseau de capteurs*. Un exemple de réseau de capteurs disposé dans une forêt est illustré sur la Figure 2.17.

Le déploiement de dispositifs fixes pour la collecte de données sociales est un domaine nouveau. La première difficulté à surmonter est en général d'identifier les individus et les interactions avec précision. En effet, les sources de données peuvent être de nature différente (image, son, vidéo, etc.), il faut souvent pouvoir reconnaitre efficacement les individus et leurs interactions en combinant ces différentes sources à l'aide d'algorithme d'analyse d'images ou de flux sonores. Quand cette reconnaissance est possible, les dispositifs fixes permettent d'extraire des interactions sociales beaucoup plus variées que les dispositifs mobiles, qui se limitent pour la grande majorité aux contacts de proximité géographique.

Ainsi, en raison des difficultés liées au processus d'identification des individus et des contacts, ce domaine de recherche reste encore très ouvert.
Nous pouvons citer les travaux pionniers de Chen *et al.* [Chen 2007a], qui ont expérimenté l'utilisation de capteurs fixes pour la collecte d'interactions sociales au sein d'une maison de retraite. Les capteurs, installés dans deux salles et un couloir, étaient équipés d'enregistreurs audio et vidéo. Ainsi, en utilisant des méthodes d'analyse d'images et de flux audio, divers types d'interactions sociales sont détectées telles que "deux personnes discutent", "se serrent la main", "s'embrassent", "se rapprochent", "s'éloignent" ou "marchent ensemble".

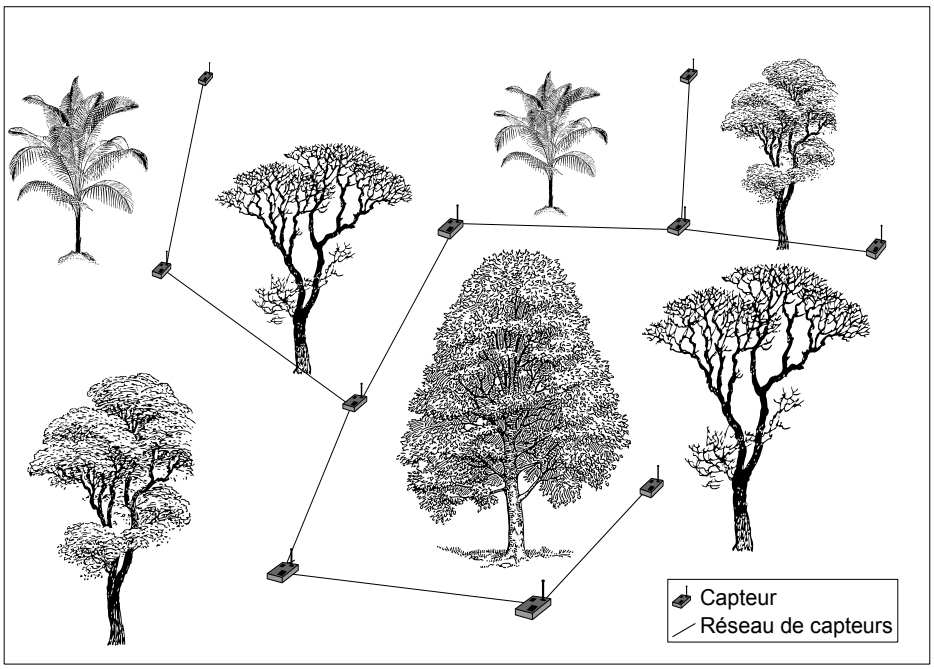


FIGURE 2.17 – Réseaux de capteurs fixes disposé au sol dans une foret

Évidement, quand les espaces à couvrir sont plus étendus, la question de la configuration optimale de ces solutions doit être étudiée pour maximiser la qualité des données sociales collectées. Par exemple, combien de capteurs sont nécessaires pour couvrir efficacement la zone d'étude ? Comment placer les capteurs sur la zone les uns par rapport aux autres ? Quels types de périphériques doivent être connectés et comment les paramétrer ? Comment répartir les calculs sur les capteurs, ou limiter les échanges sur le réseau ? etc.

Pauwels *et al.* [Pauwels 2007] apportent une visibilité sur les champs d'application de cette très vaste solution. Ils montrent en effet l'intérêt d'un réseau de capteurs fixes pour permettre de déduire : (i) **Qui joue un rôle**, par l'identification des acteurs du système (personnes ou animaux), (ii) **Ou et Quand**, en fournissant des fenêtres temporelles pour déterminer le contexte, (iii) **Quoi**, par la reconnaissance des activités et des interactions sociales des entités étudiées, (iv) **Pourquoi**, en associant une sémantique et un objectif aux actions, et (v) **Comment**, par la reconnaissance des expressions, des mouvements ou des gestes.
Selon les auteurs, de tels dispositifs, intégrés de façon durable dans notre quotidien, permettraient de créer une "*Intelligence Ambiante*", capable d'identifier les individus, de reconnaitre leurs actions, leurs émotions et leurs intentions, de façon à les assister selon leurs préférences individuelles et leurs besoins.

2.5 Conclusion

Dans ce chapitre, nous avons présenté les champs de recherche principaux qui constituent le contexte des travaux présentés dans ce mémoire.

Dans un premier temps, nous avons pu observer que la science des réseaux trouvait ses origines dans le domaine mathématique de la *théorie des graphes*, une discipline vieille

d'environ 300 ans, et qui a été alimentée par de nombreuses communautés scientifiques. Nous avons ainsi présenté les fondements théoriques de cette discipline, en présentant les principales définitions, notations, propriétés et structures couramment rencontrées dans la littérature sur les réseaux. Nous avons ainsi pu définir le concept de *réseau social*, qui constitue le point central des travaux de recherche présentés dans ce mémoire.
Une fois les bases posées, cet état de l'art a ensuite été orienté selon les travaux qui seront abordés dans les chapitres suivants.

La première section a été consacrée aux deux grandes approches utilisées pour l'analyse des réseaux sociaux : (i) les *approches traditionnelles* qui reposent sur l'exploitation des propriétés structurelles pour caractériser les noeuds, les liens ou la structure, et (ii) les *approches récentes* qui tentent d'extraire de la connaissance des réseaux sociaux, par l'application des concepts issus de la fouille de données.

La section suivante à été consacrée au cas spécifique des méthodes de modélisation et d'étude des phénomènes de propagation. Nous avons montré que cette problématique a été étudiée à travers deux problèmes duaux : celui de la *percolation*, qui cherche à comprendre si la structure permet au phénomène de "traverser" le réseau, et celui de la *diffusion*, qui s'intéresse aux caractéristiques du phénomène comme son amplitude ou aux moyens de le contenir. Dans ce deuxième type de problème, nous avons montré comment les modèles mathématiques fondateurs, conçus à base de compartiments, ont été étendus aux réseaux pour prendre en compte les structures relationnelles réelles qui offrent un média aux processus de diffusion.

Enfin, une dernière section s'est intéressée aux travaux menés sur la collecte automatique des données sociales pour la génération de réseaux sociaux. Nous avons pu observer que les premières approches ont exploité les données issues des sites d'échanges et de partage pour générer des réseaux sociaux basés sur l'activité des utilisateurs. Les méthodes récentes ont, elles, exploité les capacités offertes par les microcontrôleurs, à travers des dispositifs fixes ou mobiles, pour la collecte automatique des données de terrain qui ne peuvent pas être extraites uniquement à partir de sites web.

Chapitre 3

Phénomènes de diffusion dans les réseaux sociaux dynamiques

Sommaire

Comme introduit dans le Chapitre 2, une approche prometteuse qui apporte une alternative aux modèles mathématiques initiaux pour étudier les phénomènes de diffusion (épidémie, virus informatique, phénomène de mode, rumeur, etc.) consiste à s'intéresser aux interactions entre les individus. Cependant, il est extrêmement difficile d'obtenir des informations précises sur l'évolution réelle de ces processus, tant les composantes impliquées sont nombreuses, complexes et variables. En effet, il faut pouvoir identifier les acteurs, l'ensemble des interactions et facteurs impliqués, les changements d'état liés à la fois au réseau, mais également aux individus ainsi que leurs causes et effets, tout ceci au cours du temps et à une échelle suffisamment grande pour permettre une étude efficace du phénomène. De telles données restent encore extrêmement rares et souvent très partielles et biaisées.

Ainsi, de nombreux travaux [Christley 2005, Christakis 2010, Salathe 2010a] reposent sur des simulations ou des modèles très simples, qui pour la majeure partie, considèrent la diffusion comme évoluant sur des réseaux statiques. Or, nous savons que la plupart des réseaux du monde réel ne sont pas fixes. Ce sont au contraire des objets animés, en évolution constante, au sein desquels des noeuds et des liens peuvent, à chaque instant, apparaitre ou disparaitre. Cette dynamique joue d'ailleurs un rôle essentiel, puisque c'est elle qui définit les comportements des noeuds, favorise l'émergence ou l'isolement de certains individus et est à l'origine des caractéristiques structurelles globales souvent observées sur les réseaux du monde réel.

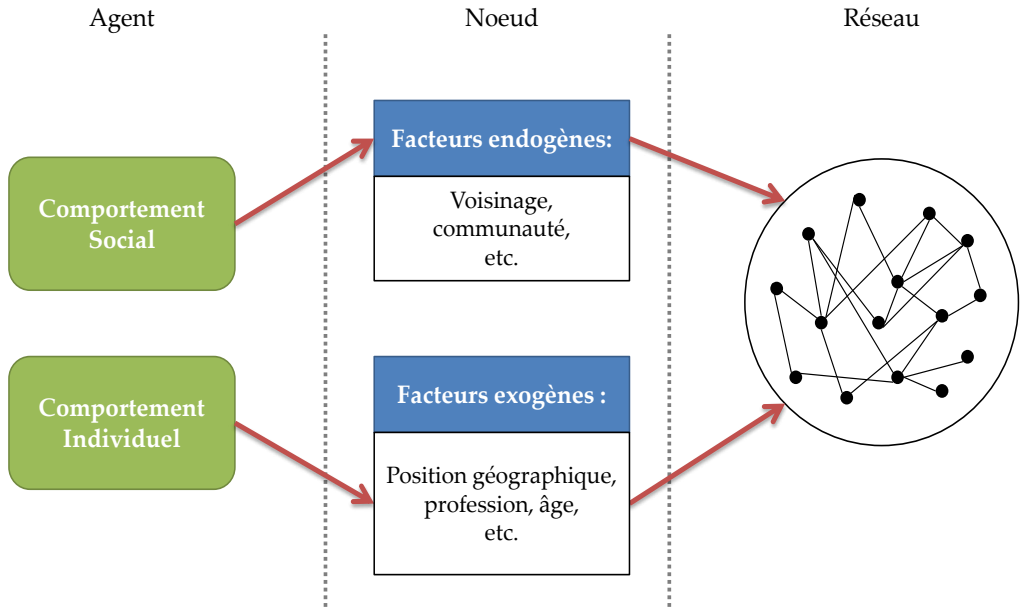


FIGURE 3.1 – Facteurs impliqués dans la dynamique

En ce qui concerne les phénomènes de diffusion, des travaux très récents ont déjà exploré ponctuellement l'effet de la dynamique du réseau sur l'ampleur du processus [Li 2004, Albano 2012]. En effet, pour comprendre pleinement ces phénomènes cette dimension doit naturellement être prise en compte. Prenons simplement l'exemple des stratégies d'intervention actuellement utilisées en épidémiologie ; celle qui donne les meilleurs résultats consiste à vacciner les individus ayant le degré le plus élevé [Christley 2005]. Cependant, les individus de plus fort degré à l'instant t, sont susceptibles de ne pas être les mêmes à l'instant $(t + \Delta)$ en raison des changements survenus sur le réseau.
Ainsi, la dynamique des réseaux doit être considérée pour tenter d'apporter des réponses à des questions émergentes telles que :

1. Quels sont les processus impliqués dans la dynamique du réseau ?
2. Comment influencent-ils le phénomène de diffusion ?
3. Quels sont les facteurs qui influencent ou pas la diffusion ?

Les premiers travaux sur ce sujet se sont restreints à l'impact de changements topologiques du réseau sur les processus de diffusion [Yoneki 2008, Read 2008, Albano 2012]. Cependant, ces travaux abordent souvent la dynamique du réseau à travers des stratégies d'évolution extrêmement simplifiées qui ne reflètent pas la complexité réelle des facteurs impliqués dans l'évolution des réseaux du monde réel. En effet, les changements survenant localement sur un noeud peuvent être le résultat (i) de comportements sociaux, qui font intervenir des facteurs endogènes comme le voisinage du noeud, ou la communauté à laquelle il appartient, ou (ii) de comportements individuels liés à un ensemble de facteurs exogènes tels que sa position géographique, sa profession, son age, etc. (voir Figure 3.1). L'état de l'art en la matière est assez éclectique, puisque les différents travaux menés se sont focalisés pour l'instant sur des points de vue et des situations très particulières.

Dans ce chapitre, nous présentons une approche qui prend en compte la dimension dynamique des réseaux de manière plus générale, pour l'étude des phénomènes de diffusion.
L'une des principales difficultés consiste à intégrer de façon cohérente d'une part **la dyna-**

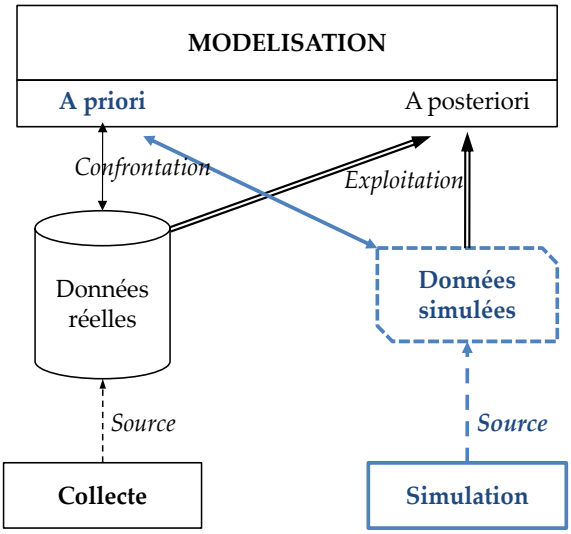


FIGURE 3.2 – Des données sociales aux modèles : modèle *a priori* et données simulées

mique du réseau, et d'autre part les différents aspects liés à **la dynamique du processus de diffusion** lui-même.

Pour étudier les phénomènes de diffusion sur les réseaux en évolution et répondre aux questions soulevées par cette nouvelle problématique, nous proposons le modèle *D2SNet*, une solution unifiée, qui combine ces deux types de dynamique : réseau et diffusion. Nos travaux s'inscrivent sur l'axe de la modélisation (voir Figure 3.2) et ont pour objectif d'explorer l'effet, sur le processus de diffusion, des différentes composantes qui induisent des modifications structurelles du réseau.
Le modèle que nous définissons formellement est générique, dans le sens où il n'est pas limité à un type particulier de diffusion d'information au sens large, mais au contraire, permet de simuler tout type de diffusion sur des réseaux en évolution. On peut en effet considérer tous ces phénomènes comme étant très similaires, malgré des sémantiques différentes. Nous verrons que les principes de modélisation sont sensiblement analogues.

Dans ce chapitre nous commençons par présenter *D2SNet*, le modèle unifié que nous avons défini pour l'étude des phénomènes de diffusion sur les réseaux dynamiques.
D2SNet est ensuite utilisé pour explorer, à travers deux études, les effets de changements structurels induits par deux types de facteur :
(1) Dans la première étude, nous nous intéressons à une dynamique du réseau induite par des comportements individuels des noeuds. Contrairement aux études qui simulent l'évolution du réseau par des stratégies d'évolution synthétiques, ce travail a pour objectif de mesurer l'impact de comportements sociaux individuels couramment observés sur les réseaux du monde réel. Cette première approche permet de mettre en évidence que les comportements, à la base de la formation des liens dans de nombreux réseaux sociaux, ont des effets différents sur le processus de diffusion.
(2) Dans la seconde étude, nous nous intéressons à une forme d'évolution plus réaliste, qui considère l'environnement social comme le principal moteur de la formation des liens au sein des réseaux sociaux. Cette étude vient compléter les résultats obtenus en (1), en montrant que l'impact de la dynamique n'est pas systématique, et varie selon les caractéristiques de l'environnement social dans lequel évoluent les noeuds.

Ce chapitre est organisé de la façon suivante. Dans la Section 3.1, nous présentons les

travaux récents menés sur les processus de diffusion dans les réseaux dynamiques. Dans la Section 3.2, nous présentons formellement *D2SNet*, et discutons sa complexité et sa flexibilité. Les Sections 3.3 et 3.4 sont respectivement consacrées aux études (1) et (2). Dans la Section 3.5, nous présentons *DynSpread*, l'outil graphique qui implémente ce modèle. Enfin, nous concluons ce chapitre dans la Section 3.6.

3.1 Dynamique des réseaux et phénomènes de diffusion

Ces dernières années, l'étude de l'évolution des réseaux a été un domaine de recherche très actif [Barabasi 1999, Dorogovtsev 2002, Boguna 2003, Toivonen 2009]. En effet, de nombreuses études ont été menées pour comprendre comment des mécanismes d'évolution pouvaient conduire aux caractéristiques topologiques particulières observées sur de nombreux réseaux du monde réel. Pourtant, bien que des avancées aient été faites sur la compréhension de ces mécanismes, on peut s'étonner d'observer que peu de travaux menés sur la dynamique des réseaux ont contribué à l'étude des phénomènes de diffusion. C'est ce qu'observent également Gross *et al.* [Gross 2006], qui expliquent que les études sur "*la dynamique du réseau et la dynamique sur le réseau*" ont été menées séparément, détachant ainsi deux aspects fondamentaux et intrinsèquement liés de nombreux phénomènes survenant sur les réseaux du monde réel.
Dans cette Section, nous présentons les différents travaux menés sur ce sujet.

L'étude des phénomènes de diffusion sur des réseaux en évolution est un domaine relativement récent. Nous pouvons par exemple citer l'approche mathématique de Gross *et al.* [Gross 2006] qui s'intéresse au phénomène de diffusion sur un réseau dont la dynamique est liée à un comportement "*adaptatif*" des noeuds, c'est-à-dire qu'à chaque itération, les individus susceptibles, connectés à un individu infecté, ont la possibilité de casser leur lien avec l'individu infecté et d'en recréer un nouveau avec un individu susceptible choisi aléatoirement selon une probabilité w. Gross *et al.* étudient cette stratégie d'évolution selon deux points de vue. Ils montrent tout d'abord qu'au niveau de la structure, la stratégie tend à produire deux clusters composés respectivement des individus susceptibles et des infectés. En ce qui concerne la diffusion, ils observent que l'augmentation de la probabilité w tend à réduire la taille de l'épidémie.
De façon très similaire, Volz et Meyers [Volz 2007] s'intéressent à la diffusion dans un réseau évoluant selon le modèle NE (*Neighbour Exchange*), basé sur la permutation de noeuds dans deux liens. Dans ces deux approches, le nombre de noeuds et de contacts reste figé, seule la composition des contacts est modifiée.

Par simulation, Li *et al.* [Li 2004] étudient la diffusion sur un réseau petit-monde dynamique. Pour cela, ils étendent le modèle de Watts et Strogatz [Watts 1998] en y ajoutant une probabilité de mouvement des noeuds notée p_{move}. Ce processus peut être vu comme une décision de reconnecter un lien. La reconnexion s'effectue exactement comme dans le modèle de Watts et Strogatz. Leurs résultats montrent que le temps caractéristique de la diffusion diminue quand p_{move} augmente ; autrement dit, plus il y a de l'activité sur le réseau et plus l'épidémie survient précocement et croît vite dans le temps.
Dans leurs travaux, Read *et al.* [Read 2008] étudient comment la fréquence des rencontres entre les individus peut influencer le processus de diffusion. Pour ce faire, ils étudient un groupe de 49 volontaires pour lesquels ils collectent les interactions durant 14 jours. Les données obtenues leur permettent de générer un réseau contenant 8661 contacts et 3528 individus différents. Sur ces contacts, ils identifient deux catégories ainsi que leurs fréquences : (a) les contacts basés sur les conversations et (b) les contacts physiques. Pour comprendre l'impact de la fréquence des rencontres sur le processus de diffusion, ils utilisent un modèle

SIR et simule une diffusion sur le réseau pondéré avec les fréquences des rencontres. Il est intéressant d'observer que contrairement aux travaux précédents, qui modifient à chaque itération des éléments du réseau, ce travail résume d'une certaine manière la dynamique du réseau par une pondération des liens.

Très récemment, la disponibilité de données réelles sur des phénomènes de diffusion prenant place sur le réseau Internet a permis de mesurer l'impact de la dynamique du réseau sur ces processus. Nous pouvons par exemple citer les travaux de Albano [Albano 2012], qui utilise des données issues de la transmission réelle de fichiers dans un réseau pair-à-pair et montre que lorsque la dynamique est prise en compte, le phénomène semble connaitre une décroissance avec le temps. Elle introduit ainsi un modèle de diffusion basé sur une probabilité d'infection décroissante avec le temps qui permet une meilleure approximation de la diffusion réelle sur ce type de réseau.
Toujours basée sur des données réelles, l'étude menée par Guille et Hacid [Guille 2012] s'appuie sur des données issues du site communautaire Twitter pour proposer un modèle de diffusion basé sur les comportements individuels. Leur approche repose sur l'apprentissage automatique sur des attributs extraits de trois dimensions (sociale, sémantique et temporelle) pour calculer la probabilité qu'un utilisateur u_1 transmette une information à un utilisateur u_2 à chaque instant t.

Des outils de simulation ont également été conçus pour simuler différents types de scénarios d'infection sur des populations. Christensen *et al.* [Christensen 2010] proposent par exemple un outil capable de générer des réseaux construits par différents types de relation : famille, travail, école, etc. La dynamique du réseau est abordée ici selon une combinaison de changements démographiques et topologiques, dus à des évènements tels que des naissances, des mariages, des mouvements d'immigration, des situations dans lesquelles des individus quittent ou rejoignent un école, un emploi, etc. La fréquence de ce type d'évènement est déterminée à partir de données statistiques.
De même, Barrett *et al.* [Barrett 2008] ont proposé l'outil *EpiSimdemics*, capable de reproduire les déplacements d'individus dans la ville de Portland. Cet outil, qui s'intéresse essentiellement aux contacts induits par la proximité géographique, est par exemple capable de prendre en compte les activités des individus, leur profession, leur âge, les axes de circulation et de communication.

3.2 Dynamique sur le réseau et dynamique du réseau : Vers un modèle unifié

Les phénomènes de diffusion peuvent être de natures très différentes : épidémie, virus informatique, diffusion de fichiers, rumeurs, informations, etc. Pourtant, il est intéressant de constater que bien que ces processus aient été étudiés par des communautés scientifiques diverses (sociologie [Wallace 1991], biologie [Meyers 2005], physique [Rhodes 1996], mathématiques [Kermack 1927] ou informatique [Eubank 2005]), les principes de modélisation de ces phénomènes s'avèrent être très similaires quel que soit le type de diffusion considéré.

C'est ce point de convergence que nous analysons dans la Section 3.2.1 et qui est au coeur du modèle unifié que nous proposons. Ce modèle est détaillé dans la Section 3.2.2 et nous le discutons dans la Section 3.2.3.

3.2.1 Modèles de diffusion : états et transitions

Quel que soit le phénomène de diffusion considéré, nous observons que les principes de modélisation sont basés sur un même concept : une transition d'état chez les entités étudiées

(humains, animaux ou machines). En effet, les modèles standards de diffusion représentent les individus comme pouvant prendre plusieurs états. Nous trouvons généralement un état initial, qui représente le fait que l'individu ne soit pas atteint par le phénomène, puis des états intermédiaires qui modélisent l'évolution du processus selon la perception que l'individu a du phénomène, ou l'effet qu'il a sur lui.

Typiquement en épidémiologie, les individus sont généralement catégorisés selon qu'ils soient "*Susceptible*", "*Infected*" ou "*Recovered*" [Christley 2005, Christakis 2010]. Un individu *susceptible* est un individu qui n'est pas encore infecté par la maladie, mais qui peut le devenir s'il entre en contact avec un individu infecté. Les individus dans l'état *Infected* désignent ceux qui transmettent la maladie en infectant les individus *Susceptible*. Enfin, les *Recovered* sont les individus qui ont guéri de la maladie et qui ne peuvent plus ni la contracter, ni la diffuser.
Dans le contexte de la diffusion de rumeurs, la plupart des modèles considèrent également trois états : *Ignorant* pour ceux qui ignorent la rumeur, *Spreader* pour ceux qui la propagent et *Stiffler* pour ceux qui tentent de l'étouffer [Daley 1965, Nekovee 2007].
Pour la diffusion de connaissances ou d'innovations, on retrouve les états : *Interested*, et *Adopted* [Granovetter 1978, Gruhl 2004].

Ainsi, bien que tous les modèles tendent à représenter au niveau individuel l'évolution du processus par un ensemble d'états, des divergences peuvent être observées dans (i) les mécanismes de transitions, qui peuvent être liés à la nature du phénomène, l'environnement des noeuds, le temps passé d'un état et même à la connaissance actuelle que l'on a du phénomène et (ii) l'objectif de la modélisation, qui est souvent dépendant de la sémantique du problème.
Pour (i) par exemple, la probabilité qu'un individu dans l'état *Susceptible* passe à l'état *Infected* croît avec son nombre de voisins infectés. De même, un individu dans l'état dans *Infected* passe à l'état *Recovered* selon une certaine probabilité. Cet état traduit le fait que la maladie ne sera plus transmise par cet individu soit par l'application de soins, la mise en quarantaine ou le décès de l'individu. En revanche, quand il s'agit de rumeurs, un *Spreader* peut devenir *Stiffler* de deux façons : quand il est en contact avec un autre *Spreader*, ou quand il a un contact avec un *Stiffler*.
En ce qui concerne le point (ii), les objectifs peuvent être radicalement différents selon le type de phénomène considéré. Par exemple, les modèles de diffusion d'épidémies ou de virus informatique, tentent souvent de reproduire l'évolution réelle de l'infection, dans le but de mesurer l'effet de stratégies d'intervention visant à limiter, voire même idéalement éradiquer complètement le processus [Salathe 2010b]. Dans le contexte de la diffusion de rumeurs, d'opinions, ou plus généralement de connaissances, l'objectif est au contraire de diffuser aussi vite et efficacement que possible sans entraves à l'évolution du phénomène [Zanette 2001]. Les interventions visent dans ce cas à maximiser le nombre d'individus affectés par le phénomène.
Le tableau de la Figure 3.3, présente une comparaison des principaux modèles utilisés dans chaque type de diffusion.

Finalement, de par leur sémantique large et leur facilité à être transposés à d'autres types de diffusion, les modèles épidémiques sont ceux qui ont été les plus utilisés.
Si cette transposition semble plus naturelle quand on s'intéresse à la diffusion de virus informatique [Kephart 1993, Wierman 2004], les modèles épidémiques ont également été appliqués à la diffusion d'informations [Eugster 2004, Kermarrec 2003], la diffusion de fichiers dans des réseaux pair-à-pair [Albano 2012] ou même la diffusion de rumeurs [Noymer 2001, Zanette 2002]. Par exemple, Nekovee *et al.* [Nekovee 2007] expliquent que la diffusion d'une rumeur peut être perçue comme une "*infection de l'esprit*" et qu'elle

	Epidémie	**Virus informatique**	**Rumeur**	**Information/ Innovation**
Nom *(en)*	*S-I-R*	*S-I-R*	*I-S-S*	*I-A*
Etats *(fr)*	1. Susceptible 2. Infected 3. Recovered	1. Susceptible 2. Infected 3. Recovered	1. Ignorant 2. Spreader 3. Stiffler	1. Interested 2. Adopted
Transitions	▪S→I: Quand un Infecté rencontre un Susceptible. Dépend généralement d'une probabilité de transmission par contact α et du nombre de voisins infectés k $1 - (1 - \alpha)^k$ ▪I→R: Dépend du temps écoulé, ou d'une certaine probabilité β	▪Idem SIR	▪I→S: Quand un Diffuseur est en contact avec un Ignorant, avec une probabilité λ ▪S→S: Quand un Diffuseur est en contact avec un autre Diffuseur ou un Stoppeur, avec une probabilité α	▪I→A: A chaque fois qu'un voisin *j* d'un nœud *i* adopte l'innovation, i à une certaine probabilité de l'adopter également. Dépend d'une probabilité liée au lien $p_{j,i}$ (Certains modèles utilisent des comportements associés aux nœuds).
Objectifs	▪Reproduire diffusion réelle ▪Limiter/Eradiquer la diffusion	▪Limiter/Eradiquer la diffusion ▪Tester la robustesse du système	▪Maximiser la diffusion ▪Identifier/traiter les situations pouvant limiter la diffusion	▪Etudier/comprendre évolution du phénomène ▪Maximiser diffusion

FIGURE 3.3 – Comparaison des principaux modèles de diffusion

montre ainsi des "*ressemblances intéressantes avec celle d'une épidémie*".

Ainsi, dans notre tentative de proposer un modèle unifié, capable de simuler tout type de diffusion, nous devons conserver ces deux aspects fondamentaux : une représentation du phénomène sous forme d'états au niveau des noeuds, et des transitions entre ces états pouvant être fonction de nombreux facteurs : probabilité fixe, voisins d'un noeud et leur état, environnement social, temps écoulé, etc.

3.2.2 Le modèle unifié *D2SNet*

Le réseau est le support sur lequel la dynamique de l'ensemble du système peut être observée. Lorsque l'on considère les phénomènes de diffusion sur des réseaux en évolution, deux types de dynamique doivent être considérées. (i) D'un côté **la dynamique de la diffusion**, qui repose sur l'état des noeuds et leur évolution. (ii) De l'autre **la dynamique du réseau**, qui s'exprime par des changements au niveau de la topologie du réseau, des attributs des noeuds, et leur occurrence.

Ainsi, il s'agit de concevoir un modèle qui combine ces deux processus dynamiques qui s'appliquent à une ressource commune : le réseau. Nous devons également prendre en compte le fait que ces processus peuvent également être interdépendants. Par exemple, l'évolution de la topologie du réseau peut être le résultat du comportement des individus, alors que la transition d'un état à l'autre peut dépendre du voisinage du noeud et le comportement peut lui même évoluer selon l'état des voisins. La difficulté majeure dans l'élaboration

d'un modèle unifié réside donc dans la fusion cohérente de ces deux aspects et de leur co-dépendance éventuelle.

Dans le modèle unifié que nous proposons, *D2SNet* (*Diffusion in Dynamic Social Networks*), nous étendons le concept d'automate à états finis et modélisons le changement d'état d'un noeud par des *automates à états des noeuds* au sein desquels la transition d'un état à l'autre peut être fonction de divers facteurs tels que le comportement du noeud ou son environnement. Le modèle intègre également la notion d'itération qui permet de faire évoluer, sur un même intervalle de temps, le réseau et la diffusion. Le modèle est dit "*unifié*" car il combine les deux processus dynamiques (réseau et de diffusion) et permet de modéliser différents types de diffusion.

Plus formellement, soit $G = (V, E)$ le réseau sur lequel le phénomène de diffusion est étudié. V est l'ensemble des noeuds et E l'ensemble des liens avec $E \subseteq V \times V$.

Soit $S = \{s_1, s_2..., s_n\}$ un ensemble fini non-vide d'états et V^s l'ensemble des couples *(noeud,état)* qui définissent l'état de chaque noeud selon sa perception du phénomène ou l'effet qu'il a sur lui, c.-à-d. $V^s = \{v^s = (v, s_i)\ ;\ v \in V, s_i \in S\}$.

Nous définissons le *réseau d'états* $G^s = (V^s, E, S)$, qui représente le réseau initial G, dans lequel chacun des noeuds est dans un état donné.
Nous appelons G le *graphe sous-jacent* de G^s.
Ainsi, en accord avec les différents modèles de diffusion, les états peuvent par exemple être "*Susceptible*", "*Infected*" et "*Recovered*" s'il s'agit de maladies infectieuses, ou plutôt "*Ignorant*","*Spreader*" et "*Stiffler*" s'il s'agit d'une rumeur.

Soit $\mathscr{G}$ l'ensemble des *réseaux d'états* sur lesquels le phénomène de diffusion est étudié. Nous définissons "*l'automate à états des noeuds*" $A_{\mathscr{G}} = (T, \phi, SI, SF)$ sur le *réseau d'états* G^s appartenant à $\mathscr{G}$ tel que :
$T \subset S \times S$ est l'ensemble des transitions,
$\phi : V^s \times S \longrightarrow [0..1]$ est la *fonction de transition* qui renvoie pour un noeud $v^s = (v, s_i) \in V^s$ dans l'état s_i la probabilité de passer à l'état s_j.
SI et SF sont respectivement les ensembles d'états initiaux et finaux avec $SI \subseteq S$ et $SF \subseteq S$.
Dans certains cas, SI et SF peuvent être réduits à des singletons, voire même un ensemble vide dans le cas de SF. Le cas typique est celui des formes de diffusion caractérisées par des modèles cycliques. C'est par exemple le cas des maladies qui peuvent être modélisées par le modèle *SIS*, au sein duquel un individu dans l'état *Recovered* peut être de nouveau infecté. Dans de tels types de diffusion, il n'y a pas d'état final.

Plutôt que de définir le changement d'état avec une probabilité fixée, nous introduisons une fonction de transition ϕ pour deux raisons. (i) La probabilité de changer d'état varie en fonction du noeud et du temps. (ii) Le changement peut également être dépendant de l'environnement du noeud. Par exemple, toujours dans le contexte des maladies, la probabilité pour un noeud donné de passer de l'état "*Susceptible*" à celui d'"*Infected*", croît avec le nombre de voisins infectés.
Dans une telle configuration, ϕ sera par exemple défini sur ces deux états par :

$$\phi((v,\text{"}Susceptible\text{"}),\text{"}Infecté\text{"}) = 1 - (1 - \alpha)^{N^v}$$

où α est la probabilité de transmission par contact et N^v le nombre de voisins infectés du noeud v.

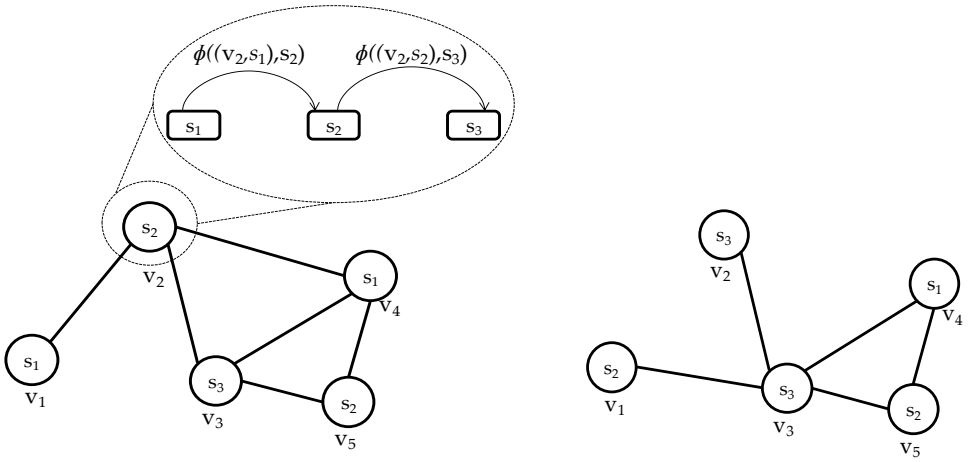


FIGURE 3.4 – Réseaux d'états selon le modèle *D2SNet*
Exemple des réseaux (a) $G^s_{k_l}$ et (b) $G^s_{k_{l+1}}$

Nous définissons $\Phi : \mathscr{G} \longrightarrow \mathscr{G}$ la *fonction de transition du réseau*, qui renvoie l'image du réseau d'états G^s par ϕ, c.-à-d. le réseau d'états dans lequel les noeuds sont les images des noeuds de G^s par ϕ.

Nous définissons le *modèle d'évolution du réseau*, $EM = (\Psi, Q)$ ou :
$\Psi : \mathscr{G} \longrightarrow \mathscr{G}$ est la *fonction d'évolution* qui décrit l'évolution du réseau à chaque itération. Évidemment, cette fonction peut être non-déterministe. Elle définit la façon dont des noeuds et des liens sont ajoutés ou supprimés dans le réseau. Elle peut également définir les changements qui surviennent au niveau des attributs des noeuds ou des liens. De cette façon, Ψ manipule uniquement le *graphe sous-jacent*.
Q est la vitesse d'évolution du réseau. En d'autres termes, le nombre de fois que la fonction Ψ est appliquée au réseau entre chaque itération.
Il est important de garder à l'esprit que EM modélise la dynamique de l'ensemble du réseau.

Soit $K = \langle k_0, k_1, ..., k_m \rangle$, la séquence de temps sur laquelle le phénomène de diffusion est étudié, tel que $\forall l \in [0..m],\ k_l < k_{l+1}$.
Nous définissons la séquence de *réseaux d'états* $G^s_K = \langle G^s_{k_0}, G^s_{k_1}, ..., G^s_{k_m} \rangle$, dans laquelle chaque $G^s_{k_l} = \Phi(\Psi(G^s_{k_{l-1}}))$ est le *réseau d'états* qui résume le phénomène de diffusion à l'instant k_l. Ainsi, chaque $G^s_{k_l}$ est défini par $G^s_{k_l} = (V^s_{k_l}, E_{k_l}, S)$ ou $V^s_{k_l}$ et E_{k_l} sont respectivement l'ensemble des noeuds du réseau et l'ensemble des liens à l'instant k_l.

Le modèle *D2SNet* que nous proposons est défini comme le *modèle d'évolution fusionné* $(G^s_0, A_{\mathscr{G}}, \mu, EM, NB_{init}, ToS, T_{max})$ dans lequel :
G^s_0 est le *réseau d'états* initial,
$A_{\mathscr{G}} = (T, \phi, SI, SF)$ est *l'automate à états des noeuds* défini sur G^s,
μ est le nombre de mises à jour des états des noeuds entre chaque itération (pour une évolution conjointe du réseau et de la diffusion),
$EM = (\Psi, Q)$ est le *modèle d'évolution du réseau* défini sur $\mathscr{G}$.
NB_{init} est le nombre de noeuds qui initie le processus de diffusion,

ToS est l'instant à partir duquel commence la diffusion, et
T_{max} est la période durant laquelle le phénomène est étudié, c.-à-d. $T_{max} = |K|$.

L'Algorithme 1 détaille le modèle. Nous commençons par créer un ensemble L contenant $|S|$ listes vides (une liste pour chaque état), ou chaque liste L^i sera chargée de stocker, à chaque itération, le pourcentage de noeuds à l'état i (Ligne 1 de l'Algorithme 1). Chaque élément L^i_k représente ainsi le pourcentage de noeuds dans l'état i à l'instant k.
Ensuite, en utilisant *l'automate à états des noeuds* $A_{\mathscr{G}}$, le *réseau d'états* initial G^s_0 est généré (Ligne 3).
Tant que le moment marquant le début de l'infection n'est pas atteint, c.-à-d. tant que $k < ToS$, le réseau évolue selon le *modèle d'évolution* EM (Lignes 4-7). Une fois ce moment atteint, NB_{init} noeuds choisis aléatoirement initient le processus de diffusion par une évolution de leur état (Ligne 8).
Par la suite, à chaque itération k, (i) l'évolution conjointe du réseau et du processus de diffusion affecte le réseau d'états (Lignes 12-21) et (ii) l'ensemble L est mis à jour (Lignes 22-25). Ces deux opérations sont répétées jusqu'à ce que T_{max} soit atteint, c.-à-d. tant que $k < T_{max}$ (Lignes 10-27).

La Figure 3.4 montre deux exemples de *réseaux d'états* : (a) $G^s_{k_l}$ et (b) $G^s_{k_{l+1}}$. Nous pouvons observer qu'à l'instant k_{l+1} l'état des deux noeuds situés les plus à gauche évolue, alors que deux liens sont supprimés et un autre est créé.

3.2.3 Discussion

La dynamique d'un phénomène est principalement caractérisée par la nature des changements et la fréquence à laquelle ils surviennent. Dans notre tentative de proposer un modèle unifié de diffusion et d'évolution du réseau, les deux processus sont définis par un ensemble de variables paramétrables. Cela nous permet d'observer l'impact relatif de chaque type de dynamique sur le processus résultant en faisant varier ces paramètres.

En effet, la dynamique du réseau peut être plus ou moins lente, par rapport au phénomène de diffusion. Dans le cas des rumeurs par exemple, l'information semble parfois se répandre plus rapidement que le réseau qui la supporte. En revanche, quand on considère la diffusion de certaines maladies telle que la grippe, le réseau de contacts de proximité qui permet la diffusion semble, lui, évoluer beaucoup plus vite que la maladie elle-même.

La dynamique liée au processus de diffusion se traduit localement par un changement d'état des noeuds. Ce sont les paramètres du modèle de diffusion utilisé (par exemple α et β dans le cas du modèle *SIR*) qui définissent la probabilité qu'aura un noeud de passer d'un état s_i à un état s_j selon la fonction ϕ. A un niveau global, la fréquence des changements est contrôlée par le nombre de mises à jour des noeuds μ à chaque itération.

En ce qui concerne l'évolution du réseau, les changements surviennent selon le modèle d'évolution appliqué Ψ et la fréquence de ces changements dépend de la vitesse d'évolution Q à chaque itération. Le cas extrême étant $Q = 0$, ou la diffusion se produit sur un réseau statique.

D'un point de vue heuristique, les paramètres μ et Q reflètent la dynamique de l'ensemble du système. Ils traduisent le fait qu'entre μ mises à jour des états des noeuds, Q modifications sur le réseau sont opérés et inversement ; pour Q modifications sur le réseau, μ mises à jour des états des noeuds sont effectuées. Ainsi, nous appelons le rapport μ sur Q le *taux de dynamicité* λ du système (cf. Equation 3.1).

$$\text{si } Q \neq 0 \qquad \lambda = \frac{\mu}{Q} \tag{3.1}$$

algorithme 1 Évolution *D2SNet*

Précondition : $G = (V, E)$: **Réseau**, $A_{\mathscr{G}} = (T, \phi, SI, SF)$: **Automate à états des noeuds**, μ : **Entier**, $EM = (\Psi, Q)$: **Modèle d'évolution**, NB_{init} : **Entier**, ToS : **Unité de temps**, T_{max} : **Unité de temps** ; avec $ToS < T_{max}$

1. L : **Ensemble de listes** $L^i \leftarrow$ ajouter $|S|$ listes vides à L
2. k : **Iteration** $\leftarrow 0$
3. Générer $G_0^s = (V_0^s, E, S)$ tel que $\forall v^s = (v, s_l) \in V_0^s$, $s_l \in SI$
   %Le réseau évolue selon EM
4. **tant que** $(k < ToS)$ **faire**
5. $\quad G_{k+1}^s \leftarrow \Psi^Q(G_k^s)$
6. $\quad k \leftarrow k + 1$
7. **fin tant que**
8. Sélectionner aléatoirement NB_{init} noeuds et passer leur état à s_j si $(s_l, s_j) \in T$ et $s_l \in SI$
   %Évolution conjointe du réseau et de la diffusion selon EM et $A_{\mathscr{G}}$
9. max **: Entier** $\leftarrow MAX(\mu, Q)$
10. **tant que** $(k < T_{max})$ **faire**
11. $\quad t \leftarrow 0$
    %Le réseau évolue Q fois alors que la diffusion μ *fois*
12. $\quad$ **tant que** $t < max$ **faire**
13. $\quad\quad G_{k+1}^s \leftarrow G_k^s$
14. $\quad\quad$ **si** $Q \neq 0$ et $t \ mod \ \lfloor \frac{max}{Q} \rfloor = 0$ **alors**
15. $\quad\quad\quad G_{k+1}^s \leftarrow \Psi(G_{k+1}^s)$
16. $\quad\quad$ **fin si**
17. $\quad\quad$ **si** $t \ mod \ \lfloor \frac{max}{\mu} \rfloor = 0$ **alors**
18. $\quad\quad\quad G_{k+1}^s \leftarrow \Phi(G_{k+1}^s)$
19. $\quad\quad$ **fin si**
20. $\quad\quad t \leftarrow t + 1$
21. $\quad$ **fin tant que**
22. $\quad$ **pour tout** état $i \in S$ **faire**
23. $\quad\quad V_{k+1}^i \leftarrow \{v \in V \ ; \ (v, s_l) \in V_{k+1}^s \quad \text{et} \quad s_l = i\}$
24. $\quad\quad$ add $\frac{|V_{k+1}^i|}{|E_{k+1}|}$ à L_{k+1}^i
25. $\quad$ **fin pour**
26. $\quad k \leftarrow k + 1$
27. **fin tant que**
28. **retour** L

Si $\lambda > 1$ le processus de diffusion évolue plus vite que les modifications sur le réseau. À l'inverse, si $\lambda < 1$, les modifications sur le réseau surviennent plus fréquemment que le phénomène n'évolue. Enfin, si $\lambda = 1$, nous estimons que les phénomènes évoluent autant l'un que l'autre sur une même période.

Précisons tout de même que le *taux de dynamicité* ne donne qu'une indication sur les changements effectués au niveau macroscopique. Les modèles impliqués, à la fois au niveau de la diffusion et au niveau de l'évolution du réseau, étant souvent non-déterministes, il peut arriver qu'entre deux itérations, aucune modification ne soit observable sur le *réseau d'états*.

Il est assez difficile d'évaluer la complexité de notre modèle puisque les traitements effectués varient selon le modèle de diffusion considéré Φ, la complexité de la fonction

d'évolution Ψ et les paramètres associés μ et Q. Soit
$\Theta(\Psi)$ la complexité du modèle d'évolution appliqué au réseau et
$\Theta(\Phi)$ la complexité du modèle de diffusion utilisé.
La complexité de l'algorithme *Évolution D2SNet*, notée $\Theta(D2SNet)$, peut s'exprimer comme une fonction de $\Theta(\Psi)$ et $\Theta(\Phi)$ comme le montre l'équation 3.2.

$$\begin{aligned} \Theta(D2SNet) &= ToS \times Q \times \Theta(\Psi) + (T_{max} - ToS) \times (Q \times \Theta(\Psi) + \mu \times \Theta(\Phi)) \\ &= T_{max} \times Q \times \Theta(\Psi) + (T_{max} - ToS) \times \mu \times \Theta(\Phi) \end{aligned} \tag{3.2}$$

Globalement, nous pouvons observer que la complexité vient essentiellement des modèles impliqués. En effet, l'approche *D2SNet*, qui combine ces différents modèles au cours du temps, accroît linéairement les complexités des modèles impliqués avec le temps d'étude.

En ce qui concerne la flexibilité de cette solution, nous avons montré que *D2SNet* peut être utilisé pour étudier divers types de phénomènes de diffusion tels que la diffusion de maladies, de rumeurs, d'opinions ou de virus informatique, en adaptant les états et les transitions.
Un autre avantage de notre solution est que de nombreux types de modèles d'évolution peuvent également y être intégrés ; allant des plus simples, dans lesquels la dynamique du réseau peut être décrite comme un ensemble de règles d'évolution, aux plus complexes, ou cette fois la dynamique peut être induite par des comportements individuels, tels que la confiance ou la mobilité, qui tiendraient compte de la topologie du réseau, des attributs des noeuds, de l'environnement social, etc.
Plus généralement, la dynamique du réseau peut être le résultat de comportements réels, issus de données statistiques, collectées sur le terrain à l'aide de capteurs [Laibowitz 2006, Olguin 2008], ou même obtenues à partir de simulations plus complètes qui tentent de reproduire des comportements individuels complexes [Eubank 2005]. Dans une telle configuration, la fonction d'évolution pourra être réduite à la lecture d'informations dans un fichier ou une base de données.

3.3 Mécanismes de formation de liens élémentaires

Bien que les premiers travaux présentés dans la Section 3.1 aient pu mettre en évidence l'influence de la dynamique du réseau sur le processus de diffusion, nous observons que les stratégies d'évolution utilisées sont souvent arbitraires et ne reflètent (i) ni la complexité des comportements humains conduisant à la formation des liens au sein des réseaux sociaux (ii) ni l'avancement des études menées sur la formation et l'évolution des réseaux.

Dans le Chapitre 2 nous avons présenté des modèles de génération de réseaux sociaux permettant d'obtenir des réseaux vérifiant certaines propriétés. D'autres modèles ont également été proposés pour comprendre comment les réseaux sociaux se forment et évoluent [Dorogovtsev 2002, Albert 2002, Toivonen 2009]. Alors que certains travaux ont par exemple proposé des modèles de formation complexes, basés sur la projection des individus dans des espaces multidimensionnels [Boguna 2003], d'autres ont en revanche cherché à comprendre si les structures observées dans les réseaux sociaux pouvaient être expliquées par des mécanismes locaux [Watts 1998]. Les modèles de cette dernière catégorie, sont basés sur l'hypothèse que la topologie est le résultat de mécanismes microscopiques élémentaires gouvernant l'évolution de l'ensemble du réseau [Kumpula 2007, Toivonen 2009, Barabasi 2002]. Typiquement, on peut par exemple observer que certains individus créent des liens avec les amis de leurs amis, alors que d'autres en créent plutôt avec des individus fortement connectés.

À notre connaissance, il n'existe pas d'étude comparative sur les effets de ces différents mécanismes de formation de liens sur les processus de diffusion. Ainsi dans la première étude présentée dans ce chapitre, nous nous intéressons aux principaux mécanismes de formation, et utilisons le modèle *D2SNet* pour comparer leurs effets sur le processus de diffusion.

3.3.1 Objectifs, méthode et environnement

Les études menées par Barabasi et Albert [Barabasi 1999] ont pu mettre en évidence que la formation des liens dans les réseaux scale-free était basée sur le principe "d'*attachement préférentiel*". De même Kumpula *et al.* [Kumpula 2007] ainsi que Toivonen *et al.* [Toivonen 2009] ont eux, observé que les mécanismes dits de "*fermeture triadique*" et de "*connexion globale*" étaient à la base de la formation de nombreux réseaux du monde réel.

Dans cette première étude menée sur l'effet de la dynamique du réseau sur les processus de diffusion, nous nous intéressons à quatre mécanismes de formation élémentaires, connus pour reproduire les comportements locaux observés dans la formation de nombreux réseaux sociaux du monde réel :

- **Fermeture triadique** (FT) [Granovetter 1973, Kossinets 2006], ou **fermeture cyclique**, est un des concepts les plus répandus dans le domaine des réseaux sociaux. Il renvoie à l'expression bien connue "*les amis de mes amis sont mes amis*". Plus formellement, c'est le processus à travers lequel un noeud est susceptible de créer un lien avec les voisins de ses voisins. Un tel comportement est par exemple souvent observé sur les relations d'amitié.
  Des études récentes ont par ailleurs généralisé ce concept en montrant que la probabilité qu'un lien se crée entre deux noeuds décroît à peu près exponentiellement avec la distance géodésique qui les sépare.
- **Connexion globale** (CG) [Milgram 1967, Kumpula 2007], ou également **attachement global**, correspond à un mécanisme abstrait de création de liens sociaux, à travers lequel un noeud établit une nouvelle connexion au-delà de son voisinage immédiat, c'est-à-dire en dehors des ses voisins et de leurs voisins. Plus formellement, la connexion globale représente la probabilité empirique que deux étrangers créent une connexion. Ce type d'évolution traduit souvent le fait qu'un individu tente d'accéder à des ressources éloignées, comme cela est par exemple le cas dans les réseaux de relations professionnelles, dans lesquelles les individus tentent d'élargir leur cercle de contact. Elles peuvent également provenir de rencontres lors d'activités ou sur le lieu de travail.
  La fermeture triadique et la connexion globale sont souvent utilisées conjointement pour l'élaboration de modèles plus complexes [Ebel 2003, Boguna 2003].
- **Attachement préférentiel** (AP) [Barabasi 1999] est un mécanisme de formation qui favorise la création de liens avec les individus les plus connectés du réseau. Plus formellement, c'est un processus à travers lequel un noeud "préfère" établir une connexion avec des noeuds déjà fortement connectés. Bien que ce concept soit relativement récent, il a été étudié ces dernières décennies, notamment pour son implication dans la formation de structures scale-free, un type de structure fréquemment observé dans les réseaux du monde réel.
- **Aléatoire** (AL) [Bollobas 2001, Erdos 1960] fait référence à un mécanisme de formation au sein duquel des liens sont crées aléatoirement entre les noeuds du réseau. Ce type de mécanisme est utilisé pour modéliser l'évolution de réseaux pour lesquels aucune connaissance sur le développement n'est disponible. En effet, certaines études ont montré que les réseaux possédant une topologie complexe et des principes d'or-

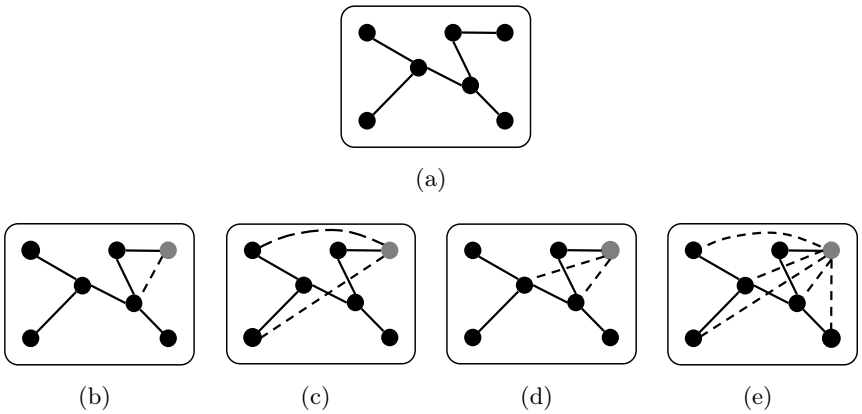


FIGURE 3.5 – Liens possibles lors de l'application de mécanismes de formation (a) réseau de référence (b) FT, (c) CG, (d) AP et (e) AL

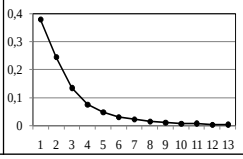
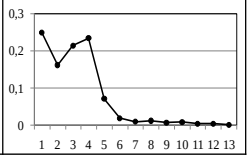
ganisation inconnus s'avèrent souvent évoluer de façon aléatoire [Albert 2002], ce qui explique que ce type d'évolution soit, encore aujourd'hui, régulièrement utilisé dans l'étude de systèmes complexes.

Quelques exemples de l'application de ces mécanismes peuvent être observés sur la Figure 3.5. La Figure 3.5(a) est le réseau de référence et les Figures 3.5(b), 3.5(c), 3.5(d) et 3.5(e) illustrent les formations possibles (en pointillés) quand on applique respectivement FT, CG, AP et AL à partir du noeud grisé.

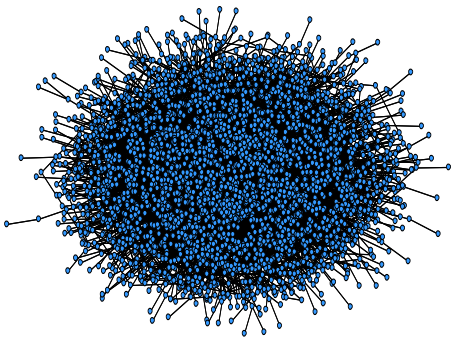
Dans cette première approche, notre objectif est de mesurer et de comparer les effets de chacun de ces mécanismes de formation sur le processus de diffusion. Pour cela, les deux réseaux initiaux décrits sur la Figure 3.6 ont été utilisés.
$N1$ est un réseau généré en utilisant le modèle *Barabasi-Albert* [Albert 2002], connu pour sa capacité à produire des réseaux *scale-free*, comme en témoigne la distribution des degrés observable sur la Figure 3.6.
$N2$ est un réseau de contacts de proximité extrait à partir d'EpiSims [Eubank 2005], un outil de simulation conçu pour reproduire statistiquement les déplacements d'individus dans la ville de Portland. Pour chacun de ces réseaux, nous supposons que nous n'avons aucun a priori sur le développement, ce qui nous permet de faire différentes suppositions quant à leur évolution en appliquant les mécanismes présentés.

Les résultats ont été obtenus en appliquant aux réseaux $N1$ et $N2$ l'approche *D2SNet* pour introduire à la fois un mécanisme de formation et le phénomène de diffusion. Le processus de diffusion est simulé en utilisant le modèle *SIR*. En ce qui concerne l'évolution du réseau, nous supposons que le comportement des individus ne change pas durant la simulation. Autrement dit, le réseau évolue toujours selon le même mécanisme de formation Ψ (FT, CG, AP ou AL).

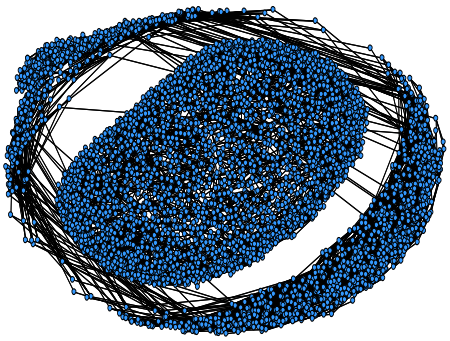
De plus, comme notre objectif est d'étudier l'effet de la dynamique du réseau sur le processus de diffusion, un certain nombre de paramètres jugés non-pertinents pour l'étude, a été fixé.
Pour le modèle *SIR*, la probabilité de transmission par contact α a été fixée à 0.1 et la probabilité de guérir β à 0.2.
De même, nous supposons que la diffusion débute à l'itération 0 (c.-à-d. $ToS = 0$). Seul un individu amorce la diffusion (c.-à-d. $NB_{init} = 1$). Une seule mise à jour des noeuds est effectuée entre chaque itération (c.-à-d. $\mu = 1$) et l'étude est menée sur une période de 120

	Réseaux	
	N1	**N2**
Origine	Généré	Simulé - Portland
Nb. nœuds	3233	4829
Nb. liens	5154	7455
Densité	0.000986	0.0006395
Degré Moyen	3.188	3.087
Degré Max.	118	17
Distribution Degré		
Coeff. Clustering	0.00427	0.60880

(a)



(b)



(c)

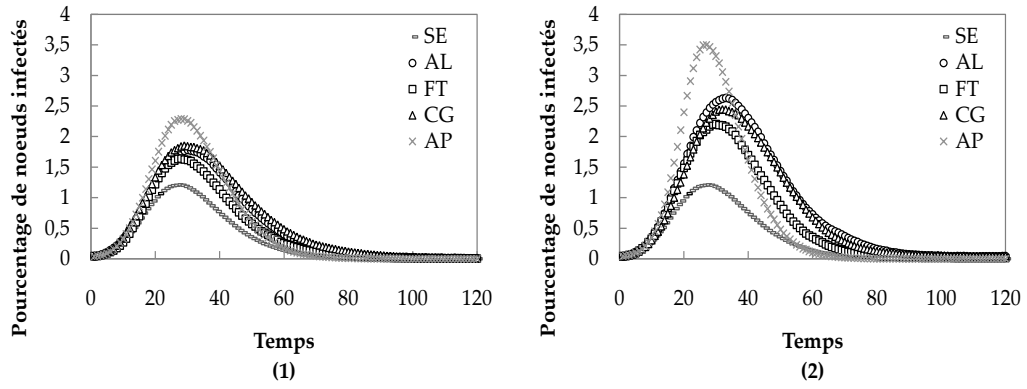
FIGURE 3.6 – Présentation des réseaux utilisés
(a) Principales propriétés des deux réseaux (b) réseau $N1$ et (c) réseau $N2$

unités de temps, c.-à-d. $T_{max} = 120$.
Enfin, *D2SNet* a été développé en JAVA et chaque test à été moyenné sur 100 exécutions. Toutes les expériences ont été menées sur la configuration matérielle suivante : Intel Core 2 Duo P8600, 2.4Ghz, 3Go Ram, Linux Ubuntu 10.10 avec JDK 1.6.

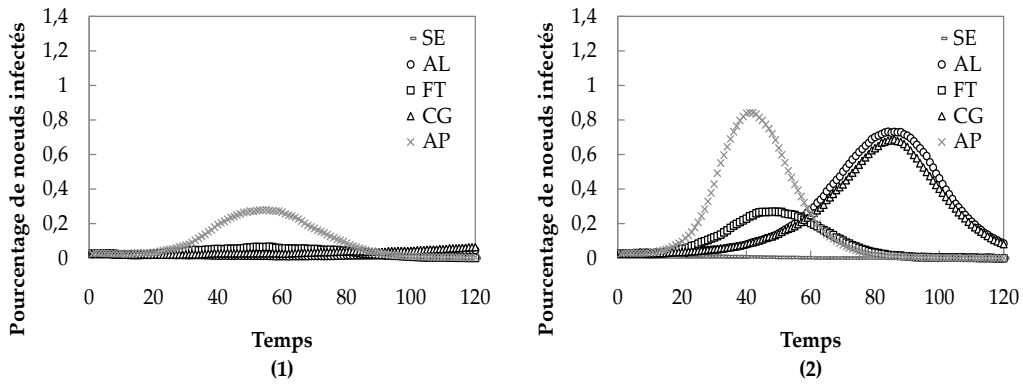
3.3.2 Résultats expérimentaux

Dans un premier temps, nous avons cherché à comprendre quels étaient les effets de ces différents mécanismes de formation (*aléatoire*, *fermeture triadique*, *connexion globale* et *attachement préférentielle*) sur les courbes d'incidences. Les courbes d'incidence sont les courbes d'écrivant l'évolution du pourcentage de noeuds infectés au cours du temps. Sur la Figure 3.7, nous comparons les courbes d'incidence obtenues avec deux vitesses d'évolution choisies arbitrairement (1) $Q = 50$ et (2) $Q = 100$. Les Figures 3.7(a) et 3.7(b) présentent

respectivement les résultats obtenus avec les réseaux $N1$ et $N2$. Sur chaque figure, nous traçons également la courbe d'incidence obtenue sans évolution du réseau (SE) comme référence.



(a)



(b)

FIGURE 3.7 – Courbes d'incidence obtenues par chaque mécanisme de formation pour (1) $Q = 50$ et (2) $Q = 100$ pour les réseaux (a) $N1$ et (b) $N2$

Le modèle *SIR* génère des courbes d'incidence en forme de *cloche*, pour lesquelles un pic est en général observable. Ce résultat, par ailleurs connu de la littérature [Christakis 2010, Newman 2010] pour $Q = 0$, indique que le nombre d'individus infectés croît jusqu'à atteindre un maximum, avant de connaitre une phase de décroissance. Dans le reste de ce mémoire, nous ferons référence à ce point de transition sur la courbe d'incidence comme étant le *pic épidémique*, ou plus simplement *le pic*.

Les observations les plus intéressantes peuvent être faites en s'intéressant à l'impact de l'évolution du réseau sur ces courbes d'incidence. A un niveau global, nous pouvons observer que la dynamique influence d'une part la valeur du pic, mais également son temps d'apparition. En effet, pour les deux réseaux considérés, le pic épidémique semble croître avec la vitesse d'évolution Q. Par exemple, lorsque le réseau $N1$ évolue selon le mécanisme AP, le pic épidémique est d'environ 2.5% quand la vitesse est de 50, contre 3.5% quand la vitesse d'évolution est de 100 (voir Figures 3.7(a)). En ce qui concerne le temps d'apparition du pic, celui-ci semble apparaitre plus tôt quand la vitesse d'évolution augmente. Cela peut s'observer sur les courbes d'incidence décalées vers la gauche entre les Figures (1) et (2).

Typiquement, pour le réseau $N2$ évoluant avec le mécanisme AP, le pic apparait environ à la $60e$ itération pour $Q = 50$, contre une apparition à la $40e$ pour $Q = 100$ (voir Figures 3.7(b)). De fortes variations sur le processus de diffusion peuvent cependant être observées selon le mécanisme de formation. Par exemple, les valeurs de pic obtenues pour le réseau $N1$ sont très proches quand $Q = 50$, alors que les différences sont beaucoup plus marquées quand $Q = 100$. De même, sur le réseau $N2$, l'**attachement préférentiel** est le seul mécanisme capable de générer un pic épidémique pour une vitesse de 50.

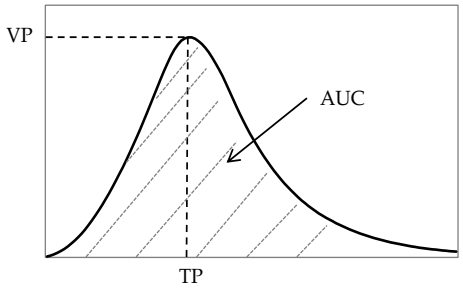


FIGURE 3.8 – Caractéristiques de la diffusion

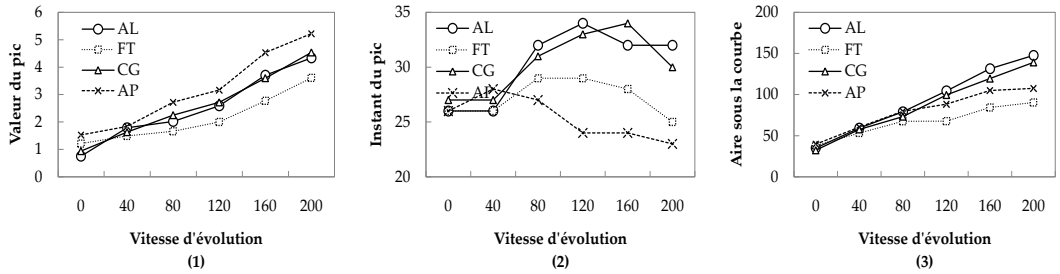
Ainsi, cette première approche nous a permis de faire deux constats. D'une part, la dynamique du réseau a un impact plus ou moins prononcé sur le processus de diffusion selon le mécanisme de formation et sa fréquence. D'autre part, nous observons que le processus de diffusion peut être caractérisé par trois indicateurs clés comme le montre la Figure 3.8 : (1) la valeur du pic VP, (2) son temps d'apparition TP et (3) l'aire sous la courbe d'incidence AUC. Dans ce contexte, l'AUC peut être interprétée comme la couverture globale du phénomène.

Ainsi, pour mieux comprendre l'effet de la dynamique sur le processus de diffusion, nous nous sommes intéressés à l'évolution de ces trois caractéristiques quand le réseau évolue. Sur la Figure 3.9, nous comparons l'évolution des indices (1), (2) et (3) pour chaque mécanisme, selon la vitesse d'évolution. La Figure 3.9(a) concerne les résultats obtenus pour $N1$ et la Figure 3.9(b) ceux obtenus pour $N2$.

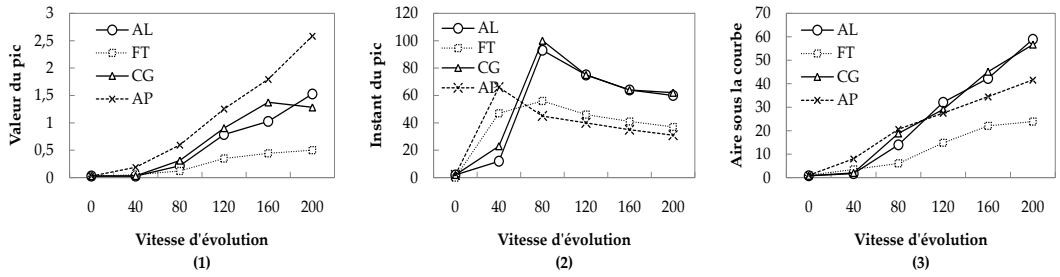
Il est intéressant de constater que des tendances communes peuvent être observées pour les réseaux $N1$ et $N2$. Premièrement, en ce qui concerne les valeurs de pic (voir Figures 3.9(1)), on observe que quel que soit le mécanisme de formation appliqué au réseau, la valeur du pic croît avec la vitesse d'évolution Q. Cependant, pour des valeurs élevées de Q, le mécanisme de formation **attachement préférentiel** tend à fournir la courbe d'incidence avec le pic le plus haut. Le pic obtenu par la **fermeture triadique** est toujours le moins élevé, alors que **aléatoire** et **connexion globale** donnent des résultats très similaires avec un pic compris entre celui obtenu par AP et celui obtenu par FT.

Quand on s'intéresse au temps d'apparition du pic (voir Figures 3.9(2)), les tendances ne sont pas les mêmes. Pour des vitesses d'évolution élevées, nous observons qu'il existe un seuil à partir duquel TP décroît. Cela peut par exemple s'observer pour le mécanisme AP quand $Q \in [80..200]$ (réseau $N1$). Si nous comparons les résultats obtenus par les différents mécanismes, nous constatons que AP donne le pic apparaissant le plus tôt. De même, alors que le pic obtenu par FT est moins élevé, nous observons qu'il est le second à survenir dans le temps. Les mécanismes AL et CG fournissent les pics qui apparaissent le plus tardivement.

Enfin, des observations intéressantes peuvent être faites quant à l'évolution de l'aire sous la courbe (voir Figures 3.9(3)). A un niveau global, quel que soit le mécanisme de



(a)



(b)

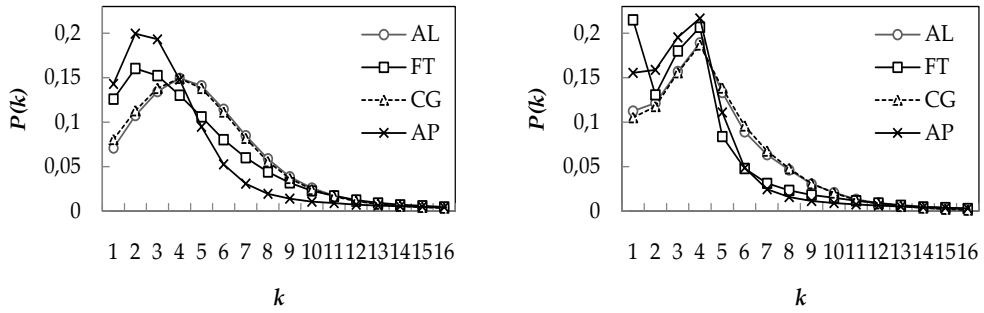
FIGURE 3.9 – Caractéristiques de la diffusion selon la vitesse d'évolution (1) VP, (2) TP et (3) AUC pour les réseaux (a) $N1$ et (b) $N2$

formation considéré, l'AUC croît avec la vitesse d'évolution. Toutefois, bien que AP soit le mécanisme qui donne le pic le plus élevé, nous pouvons observer que l'aire sous sa courbe d'incidence est plus petite que celles obtenues par AL et CG. Cela est caractéristique d'une courbe qui se resserre sur elle-même, et donc d'une diffusion certes très virulente, mais brève. FT est le mécanisme dont l'AUC est la plus petite.

Pour expliquer les tendances observées, nous avons voulu comprendre comment évoluaient les propriétés du réseau. Pour cela, nous nous sommes intéressés aux changements liés à la structure en étudiant l'évolution des principales propriétés : degré min., degré moyen, degré max., coefficient de clustering et distribution du degré. Comme nous avons pu observer sur la Figure 3.9 les fortes variations obtenues entre les différents mécanismes pour des vitesses d'évolution élevées, nous avons donc comparé de façon empirique les changements provoqués par chaque mécanisme (AL, FT, CG et AP) après le processus de diffusion en fixant $Q = 100$. Notre objectif était de comprendre vers quel type de topologie tend le réseau selon le mécanisme de formation appliqué. Ces résultats sont présentés sur la Figure 3.10. Les Figure 3.10(a) et 3.10(b) décrivent respectivement la distribution des degrés pour les réseaux $N1$ et $N2$ et la figure 3.10(c) détaille les principales caractéristiques.

Pour comprendre pleinement l'effet de ces mécanismes à la fois sur le réseau, mais également sur la diffusion, l'analyse de ces résultats doit être effectuée à deux niveaux. (i) En comparant les propriétés du réseau avant et après évolution (Figure 3.6 VS Figure 3.10) et (ii) en s'intéressant aux propriétés et structures émergeant de chaque mécanisme de formation (voir Figure 3.10). Analysons donc chaque mécanisme :

L'attachement préférentiel renforce les liens des noeuds les plus connectés. Cela peut s'observer sur la croissance du degré max. Par exemple, le degré max du réseau $N2$ avant évolution est de 17. Après diffusion, il est de 18.88 avec AL, 23.40 avec FT, 18.77



		Degré Min	Degré Moy.	Degré Max	Coefficient C.
	AL	1.0	5.673	120.27	0.0045
$N1$	FT	1.0	5.367	142.04	0.3036
	CG	1.0	5.546	119.82	0.0020
	AP	1.0	5.540	146.65	0.0142
	AL	1.0	4.54	18.88	0.3882
$N2$	FT	1.0	3.95	23.40	0.6204
	CG	1.0	4.58	18.77	0.3763
	AP	1.0	4.25	33.31	0.4588

(c)

FIGURE 3.10 – Propriétés des réseaux $N1$ et $N2$ après diffusion avec $Q = 100$ distribution des degrés pour (a) $N1$, (b) $N2$ et (c) propriétés pour $N1$ et $N2$

avec CG et 33.31 avec AP. Le mécanisme AP permet ainsi l'émergence brusque d'individus suffisamment connectés pour créer les conditions d'une diffusion rapide et virulente. Il est important de comprendre que de tels individus sont ceux qui ont la "*capacité de diffusion*" la plus élevée, puisqu'à eux seuls, ils sont capables d'infecter une grande quantité d'individus. Un tel mécanisme de formation ne fait donc qu'accélérer l'apparition de tels individus, ce qui explique les résultats obtenus : pic élevé et apparition précoce.

La fermeture triadique consolide les liens entre des groupes de noeuds. Cela se traduit par une augmentation significative du coefficient de clustering. Typiquement, il passe de 0.00427 à 0.3036 pour $N1$ et de 0.60880 à 0.6204 pour $N2$. C'est le mécanisme qui fournit le plus haut coefficient de clustering. Il est même intéressant de constater que certains des autres mécanismes utilisés peuvent même parfois réduire ce coefficient (cas de AL et CG). Cependant, bien que FT permette l'émergence de noeuds fortement connectés, sa capacité à renforcer les liens intra-communautaires limite la diffusion. Ainsi, le pic apparait relativement tôt comme nous avons pu l'observer, mais la diffusion reste essentiellement confinée à un même groupe, le modèle ne permettant pas l'émergence de connexions favorisant une diffusion d'un bout à l'autre du réseau.

L'aléatoire et la connexion globale tendent à réduire la plage de valeurs des degrés et l'effet communautaire. En effet, on peut observer sur la distribution des degrés que AL et CG génèrent une plus forte proportion de noeuds moyennement connectés, tout en réduisant la proportion de noeuds faiblement et fortement connectés (voir diminution du degré max sur la Figure3.10(c)). Nous observons également que le coefficient de clustering est systématiquement le plus bas pour ces deux mécanismes. Par exemple, sur le réseau $N1$: 0.0045 et 0.0020 pour AL et CG contre 0.3036 et 0.0142 pour FT et AP. Toutefois,

CG permet à un noeud de se connecter uniquement avec un autre au-delà de sa communauté proche. C'est ce comportement qui explique pourquoi, à l'exception du coefficient de clustering qui connait de très légères variations, les effets sur les propriétés du réseau, et donc sur le processus de diffusion, sont très similaires pour ces deux mécanismes.

Enfin, nous précisons que les tendances observées dans cette étude, ont également été confirmées pour diverses valeurs de α et β.

3.4 Stratégie avancée d'évolution du réseau

Les réseaux du monde réel ont souvent des structures topologiques non-triviales qui émergent des interactions entre les individus. En effet, nous pouvons observer que dans la plupart des réseaux sociaux, les connexions entre les noeuds ne sont ni régulières, ni totalement aléatoires. C'est cette caractéristique qui leur vaut le nom de *réseaux complexes* [Albert 2002]. Ces réseaux dits "*complexes*" sont en général le résultat de divers mécanismes microscopiques (c'est-à-dire survenant au niveau des noeuds et des liens), individuels, décentralisés et non-planifiés, aboutissant à l'émergence de propriétés statistiques macroscopiques. Par exemple, en étudiant l'évolution du réseau d'échange d'emails au sein d'une université, Kossinets *et al.* [Kossinets 2006] ont montré que les changements semblaient être dirigés par une combinaison d'effets émanant de la topologie du réseau lui-même et de la structure organisationnelle dans laquelle le réseau est intégré. Trois propriétés sont ainsi couramment identifiées [Newman 2003, Boccaletti 2006, Borner 2007] :
(i) La distribution des degrés qui suit une loi de puissance, traduisant le fait qu'il existe une faible proportion de noeuds possédant un degré plus élevé que la moyenne (propriété *scale-free*). (ii) Un coefficient de clustering élevé, qui est une caractéristique de structures topologiques de type *petit-monde* [Watts 1998]. On observe que dans de telles structures, la distance moyenne entre deux noeuds est souvent très faible. (iii) Une structure hiérarchique et communautaire [Guimera 2002] dans laquelle les individus appartiennent à des groupes (ou communautés), identifiables par une forte densité de connexions intra-communautaires et une faible densité de connexions inter-communautaires. Ces groupes appartiennent eux-mêmes à des groupes de groupes, donnant ainsi une structure naturellement hiérarchique à ces différentes communautés.

Bien que les mécanismes élémentaires étudiés dans l'étude précédente soient généralement impliqués dans la formation des liens de nombreux réseaux complexes, ils ne permettent pas à eux seuls d'expliquer les structures topologiques particulières, émergeant de l'évolution des réseaux complexes. Plusieurs modèles d'évolution ont ainsi été proposés pour reproduire de telles structures : DEB [Davidsen 2002], MVS [Marsili 2004], $KOSKK$ [Kumpula 2007], $BPDA$ [Boguna 2003], etc. Une classification et une comparaison des principaux modèles peut être trouvée dans l'article de Toivonen *et al.* [Toivonen 2009].

Dans cette deuxième étude, nous nous intéressons donc à un type particulier de modèle d'évolution, appelé *modèles spatiaux*, capables de reproduire fidèlement les propriétés observées sur la plupart des réseaux du monde réel. En utilisant *D2SNet*, nous cherchons à comprendre comment les changements survenant sur le réseau, modélisés ici par un modèle plus réaliste, affectent ou pas le processus de diffusion.

3.4.1 Le modèle spatial dynamique *DynBPDA*

Selon Toivonen *et al.* [Toivonen 2009], les modèles d'évolution des réseaux sociaux peuvent être classifiés en deux catégories. (i) Les modèles de type NEM (*Network Evolution*

*Mode*l) dont les modifications reposent uniquement sur la structure du réseau et (ii) ceux de type *NAM* (*Nodal Attribute Model*) dont le processus de formation se base sur les attributs des noeuds.
Les modèles spatiaux font partie des *NAM*. Ce sont des modèles basés sur le concept bien connu d'*homophilie* [McPherson 2001] et dont le processus de formation des liens s'appuie sur la notion de *proximité sociale*. Autrement dit, les noeuds sont projetés dans un espace social et les liens sont créés selon une probabilité qui décroît avec la distance relative dans cet espace. Ainsi, chaque noeud est identifié par un ensemble d'attributs (âge, profession, religion, centre d'intérêt, etc.) qui détermine sa position dans l'espace social. La dimension spatiale définit donc d'une part le nombre de contacts de proximité qu'un individu peut établir, mais également la structure du réseau de contacts sous-jacent qui supporte la diffusion.

L'utilisation d'un modèle spatial est motivée par le fait que l'environnement social est un aspect fondamental du processus de formation des liens chez les individus, qui influence inévitablement la structure du réseau et donc a fortiori le processus de diffusion.

Un des modèles spatiaux les plus simples et les plus efficaces est le modèle spatial *BPDA* proposé par Boguna *et al.* [Boguna 2003]. Un des plus simples parce que le modèle ne requiert qu'un petit nombre de paramètres, et un des plus efficaces parce que le modèle fournit des réseaux *scale-free* avec des structures communautaires fortes. Le point clé du modèle *BPDA* est le concept de *distance sociale*.

Plus formellement, considérons un ensemble de N individus, uniformément distribués dans un espace social à une-dimension de taille dst. La probabilité $p_{v_i v_j}$ que les noeuds v_i et v_j soient en contact décroît avec leur distance $d_{v_i v_j}$ dans l'espace social, c.-à-d. :

$$p_{v_i v_j} = \frac{1}{1 + (\frac{d_{v_i v_j}}{b})^{(1-\beta)}} \tag{3.3}$$

b est une échelle de longueur caractéristique qui permet de contrôler le degré moyen du réseau résultant. $\beta \in [0..1]$ correspond au degré de sociabilité des individus, c'est-à-dire la tendance à créer des liens plus ou moins dans l'espace social. Ainsi, quand β croît, les individus tendent à créer des liens avec des individus différents d'eux-mêmes, ou plus formellement des individus éloignés dans l'espace social. Inversement, quand β décroît, les individus tendent à créer des liens avec des individus similaires, c'est-à-dire situés relativement proches dans l'espace.
Notez que dst est un paramètre qui définit le degré de *diversité* dans le système, puisqu'il induit directement la distance maximale pouvant séparer deux individus dans l'espace social.

Précisons enfin que le modèle *BPDA* produit des réseaux non-dirigés et non-valués dans lesquels un noeud ne peut pas avoir de connexions avec lui-même. De même, deux noeuds ne peuvent pas être reliés par plus de deux liens. Le modèle *BPDA* est détaillé dans l'algorithme 2.

Le modèle *BPDA* fournit un réseau social réaliste, mais ne permet pas de simuler son évolution dans le temps. C'est la raison pour laquelle dans l'étude que nous menons, nous étendons le modèle *BPDA* pour y inclure la dimension dynamique. Dans un premier temps, notre objectif était de proposer un modèle qui pourrait décrire comment les liens des individus évoluent dans l'espace social.

Dans le contexte des modèles spatiaux, la dynamique peut être vue comme une variation survenant dans un ou plusieurs attributs d'un noeud et qui conduit à (i) un changement de sa position dans l'espace social, et (ii) la création de nouveaux liens, toujours basés sur la proximité dans l'espace social et la suppression des éventuels anciens liens.

algorithme 2 Modèle spatial *BPDA*

Précondition : N : **Nombre de noeuds**, β : **Degré de sociabilité**, dst : **Degré de diversité**, b : **Échelle de longueur**

1. $G = (V, E)$: **Réseau social**
2. ajouter N noeuds à V
3. Distribuer chaque noeud $v_i \in V$ avec une probabilité uniforme dans l'espace social de taille dst
4. **pour** i de 1 à N **faire**
5. **pour** j de $(i+1)$ à N **faire**
6. $d \leftarrow$ distance dans l'espace social entre v_i et v_j
7. ajouter $e = (v_i, v_j)$ à E avec proba. $p_{v_i v_j} = \frac{1}{1+(\frac{d}{b})^{(1-\beta)}}$
8. **fin pour**
9. **fin pour**
10. **retour** G

Évidemment, de telles variations peuvent avoir diverses origines telles qu'une évolution des croyances, l'influence d'un autre individu, une volonté personnelle, etc. Toutefois dans ce travail, nous ne nous intéressons pas aux facteurs qui peuvent causer de telles variations, mais aux effets qu'ils ont sur la structure topologique du réseau et sur le processus de diffusion.

De notre point de vue, une considération idéale de la dynamique doit conserver les concepts fondamentaux du modèle spatial, c'est-à-dire une formation des liens basée sur la notion de proximité sociale, et inclure les mouvements des individus dans l'espace selon l'évolution des attributs. Le processus d'évolution des liens peut ainsi être décomposé selon ces deux composantes élémentaires. Le degré de sociabilité qui permet aux individus de créer des liens plus ou moins loin dans l'espace social et les nouvelles opportunités de connexion qui surviennent à travers l'évolution des attributs des individus. Ces processus complémentaires sont pertinents pour l'étude du phénomène de diffusion puisque, sans perte de généralité, la dynamique du réseau peut être vue comme une combinaison permanente de sociabilité et d'évolution des attributs.

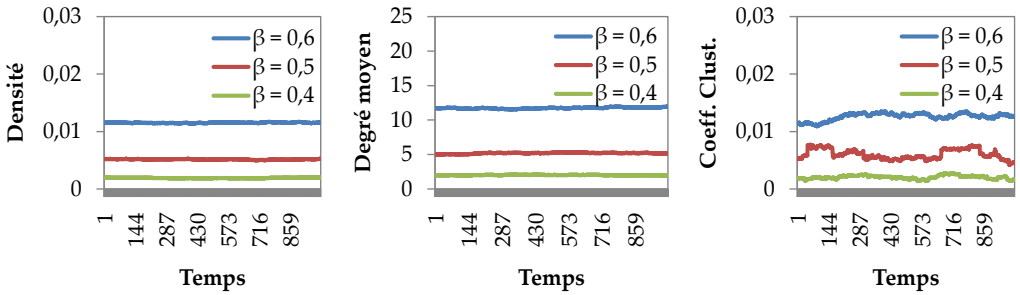
D'un point de vue heuristique, deux paramètres sont utilisés pour refléter de telles variations : (i) la fréquence des changements à un niveau global Q, c'est-à-dire le nombre d'individus affectés par une évolution de leurs attributs (correspond à la vitesse d'évolution dans le modèle *D2SNet*), (ii) la force des changements au niveau local ω, un paramètre qui contrôle la variation maximum qui peut survenir sur l'attribut.
Le modèle *BPDA* étendu, appelé *DynBPDA*, est décrit dans l'Algorithme 3.

Précisons que contrairement à certains changements qui peuvent significativement affecter les propriétés du réseau, le modèle *DynBPDA* proposé ici conserve toutes les propriétés du réseau initial.
En effet, en ce qui concerne le nombre de noeuds, celui-ci reste constant lors de l'évolution puisqu'aucun noeud supplémentaire n'est ajouté au réseau.
Lors des déplacements dans l'espace social, un noeud ne crée pas exactement le même nombre de liens. Ainsi entre deux itérations, le nombre total de liens au sein du réseau peut connaitre de très légères variations. Cependant, il est important d'observer que les nouveaux liens sont recréés exactement comme ils le seraient dans le modèle *BPDA* de base, c'est-à-dire la nouvelle position est choisie aléatoirement et les liens avec les autres noeuds sont basés sur la notion de distance sociale (voir équation 3.3). Nous garantissons ainsi que les propriétés du réseau restent stables tout au long de son évolution.

algorithme 3 Modèle spatial dynamique *DynBPDA*

Précondition : $G = (V, E)$: **Réseau social**, ω : **Variation max.** β : **Degré de sociabilité**, b : **Échelle de longueur**

1. Sélectionner un noeud $v_i \in V$
2. Supprimer tous les liens de v_i
3. Faire évoluer attribut de v_i selon ω
4. Mettre à jour position de v_i dans l'espace social
5. **pour tout** noeud $v_j \in V$ **faire**
6. **si** $v_j \neq v_i$ **alors**
7. $d \leftarrow$ distance dans l'espace social entre v_i et v_j
8. Ajouter $e = (v_i, v_j)$ à E avec proba. $p_{v_i v_j} = \frac{1}{1+(\frac{d}{b})^{(1-\beta)}}$
9. **fin si**
10. **fin pour**



(a) (b) (c)

FIGURE 3.11 – Stabilité des caractéristiques du réseau avec *DynBPDA* (a) densité, (b) degré moyen et (c) coefficient de clustering

La Figure 3.11 montre par exemple la stabilité au cours du temps de : (a) la densité, (b) le degré moyen et (c) le coefficient de clustering avec la configuration $N = 1000$, $dst = 500$, $b = 0.002$ et $\lambda = 0.5$.

3.4.2 Résultats expérimentaux

Dans le but de mesurer et de comparer l'effet de l'évolution du réseau sur le processus de diffusion, nous avons adopté la méthodologie suivante : (1) le modèle *BPDA* est utilisé pour générer le réseau initial G_0^s, puis (2) la diffusion sur le réseau en évolution est simulée en équipant le modèle *D2SNet* avec le réseau G_0^s obtenu, le modèle de diffusion *SIR* et le modèle d'évolution *DynBPDA*.

Le modèle d'évolution utilisé permet d'étudier l'impact de la dynamique du réseau selon deux aspects fondamentaux : (i) la vitesse d'évolution du réseau Q et (ii) le degré de sociabilité β. Néanmoins, le degré de sociabilité est lié à la taille de l'espace social. Ainsi, (iii) le degré de diversité dst est un autre paramètre qui doit être pris en compte. Les autres paramètres des modèles impliqués, considérés comme non-pertinents pour cette étude ont été fixés de la façon suivante :

- *SIR* : $\alpha = 0.1$ et $\beta = 0.2$
- *BPDA* et *DynBPDA* : $N = 1000$, $b = 0.02$ et $\omega = 0.5$
- *D2SNet* : $NB_{init} = 1$, $\mu = 1$, $T_{max} = 120$ et $ToS = 0$

La configuration matérielle reste inchangée : Intel Core 2 Duo P8600, 2.4Ghz, 3Go Ram, Linux Ubuntu 10.10 avec JDK 1.6.

Dans un premier temps, nous vérifions, à travers deux configurations choisies arbitrairement, que la dynamique du réseau a un impact observable sur les courbes d'incidence. Sur la Figure 3.12 nous comparons (a) l'effet de trois vitesses d'évolution quand $dst = 500$ et $sociability = 0.4$ et (b) l'effet de divers degrés de sociabilité avec $dst = 500$ et $Q = 0$ (réseau sans évolution).

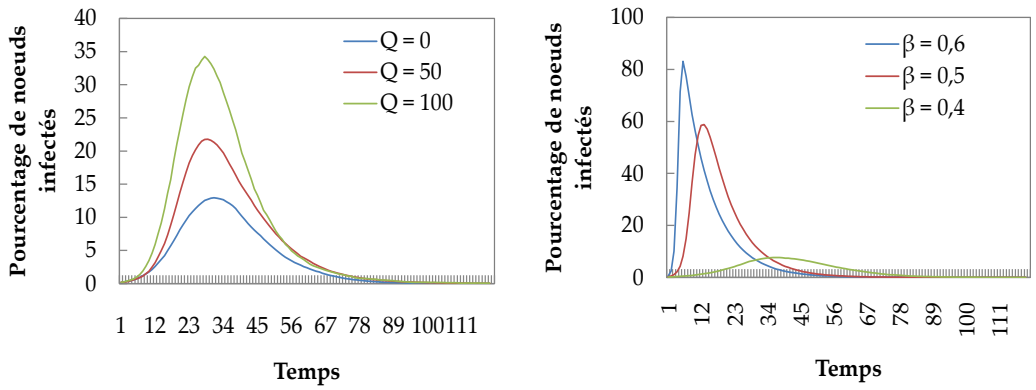


FIGURE 3.12 – Courbes d'incidence lors de l'évolution avec *DynBPDA*
(a) $dst = 500$ et $\beta = 0.4$; (b) $dst = 500$ et $Q = 0$

Comme observé précédemment, le modèle *SIR* fournit des courbes d'incidence en forme de "*cloche*", pour lesquelles un pic épidémique est facilement observable. Par ailleurs, ces résultats préliminaires confirment nos hypothèses, puisque de fortes variations peuvent être observées dans le processus de diffusion selon la configuration. En effet, la vitesse d'évolution et la sociabilité semblent être deux paramètres qui influencent significativement (i) la valeur du pic (VP), (ii) son apparition dans le temps (TP) et (iii) l'aire sous la courbe d'incidence (AUC).

Pour mieux comprendre l'implication de ces différents paramètres sur le phénomène de diffusion, nous commençons par étudier, pour différentes vitesses d'évolution, comment évolue VP selon la diversité et la sociabilité. La Figure 3.13 montre ces résultats quand (a) le réseau n'évolue pas, (b) $Q = 50$ et (c) $Q = 100$.

Plusieurs observations intéressantes peuvent être faites. À un niveau global, quelle que soit la vitesse d'évolution, nous pouvons observer que la valeur du pic décroît quand la sociabilité diminue et que la diversité augmente. Par exemple, sans évolution du réseau et avec une sociabilité comprise entre 0.41 et 0.43 : $VP \in [40..60]$ quand $dst \in [150..250]$, alors que $VP \in [0..20]$ quand $dst \in [700..900]$. Ces résultats montrent que quand les individus ont un faible degré de sociabilité dans un espace social possédant une forte diversité, le processus de diffusion est compromis.

Cependant, lorsqu'on compare les résultats obtenus avec les différentes vitesses d'évolution, nous observons que les tendances ne sont pas exactement les mêmes pour toutes les valeurs de Q. En effet, pour un degré de sociabilité élevé et une diversité faible, les différences sont plus marquées. Par exemple, pour $dst \in [500..900]$ et $\beta \in [0.41..0.43]$: $VP \in [0..20]$ sans évolution, $VP \in [0..40]$ quand $Q = 50$ et $VP \in [20..40]$ quand $Q = 100$. Cela suggère que

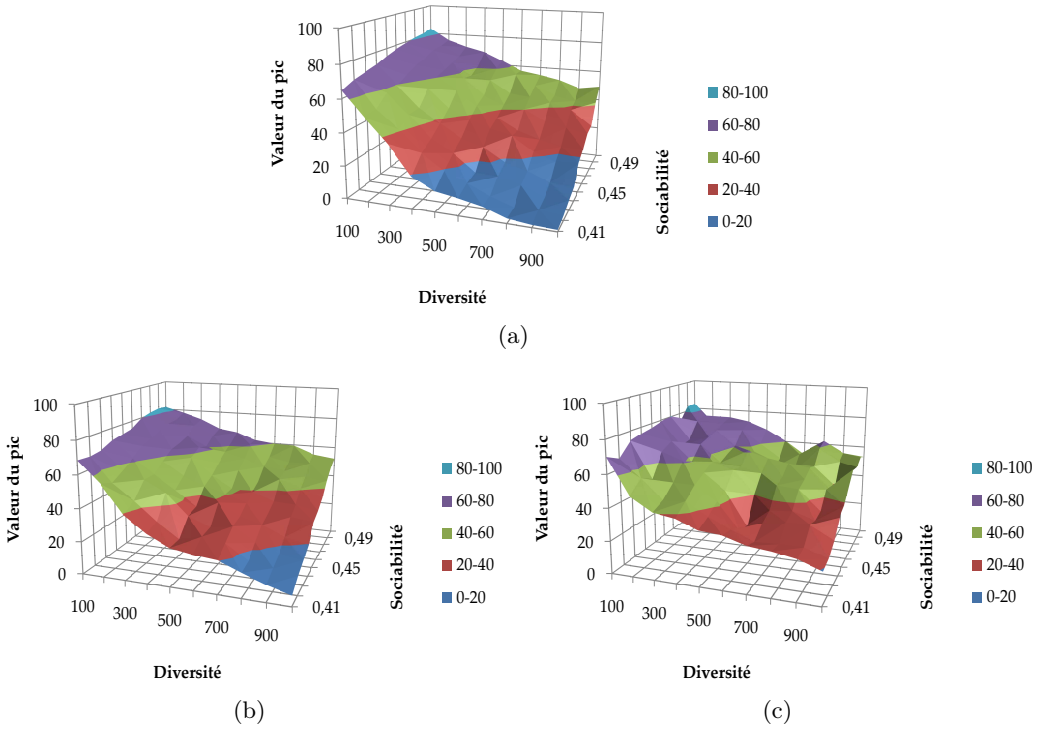


FIGURE 3.13 – Évolution de la valeur du pic selon la sociabilité et la diversité quand (a) $Q = 0$, (b) $Q = 50$ et (c) $Q = 100$

dans certaines configurations, les changements survenant sur le réseau semblent accroître la valeur du pic.

En considérant uniquement certaines valeurs de diversité, nous pouvons étudier ces variations en comparant les effets des différentes vitesses d'évolution selon l'évolution de la sociabilité. Sur la Figure 3.14, nous comparons les effets de différentes vitesse d'évolution sur (1) VP, (2) TP et (3) AUC quand (a) $dst = 100$, (b) $dst = 500$ et (c) $dst = 1000$.

En ce qui concerne la valeur du pic, des observations intéressantes peuvent être faites. Globalement, nous constatons que quelle que soit la vitesse d'évolution, des valeurs de diversité faibles tendent à réduire VP. Par exemple, sans évolution du réseau VP est d'environ 80% quand la diversité est de 100, contre 60% avec une diversité de 500. De plus, dans toutes les configurations, nous notons que VP semble croître linéairement avec la sociabilité comme le montrent les Figures 3.14(1) sur lesquelles les équations associées ont été affichées. Par exemple, sans évolution du réseau et avec une sociabilité de 0.5, la valeur du pic peut être approchée par l'équation $y = 1.879 \times \beta + 64.848$ pour $dst = 100$, alors qu'elle peut l'être par $y = 3.6285 \times \beta + 31.056$ pour $dst = 500$.
Toutefois, lorsqu'on compare les résultats obtenus entre les différentes valeurs de diversité utilisées, des effets différents sont observés selon la vitesse Q. Typiquement, quand la diversité est faible, nous pouvons observer que quelle que soit la sociabilité, l'évolution du réseau n'a aucun effet sur VP (voir Figure 3.14(a)(1)). Par exemple, pour $\beta = 0.41$, la valeur du pic est d'environ 66% quelle que soit la vitesse d'évolution. Cependant, pour une diversité moyenne ($dst = 500$), nous observons que la vitesse d'évolution du réseau a un impact qui semble être lié au degré de sociabilité (voir Figure 3.14(b)(1)). En effet, pour une sociabi-

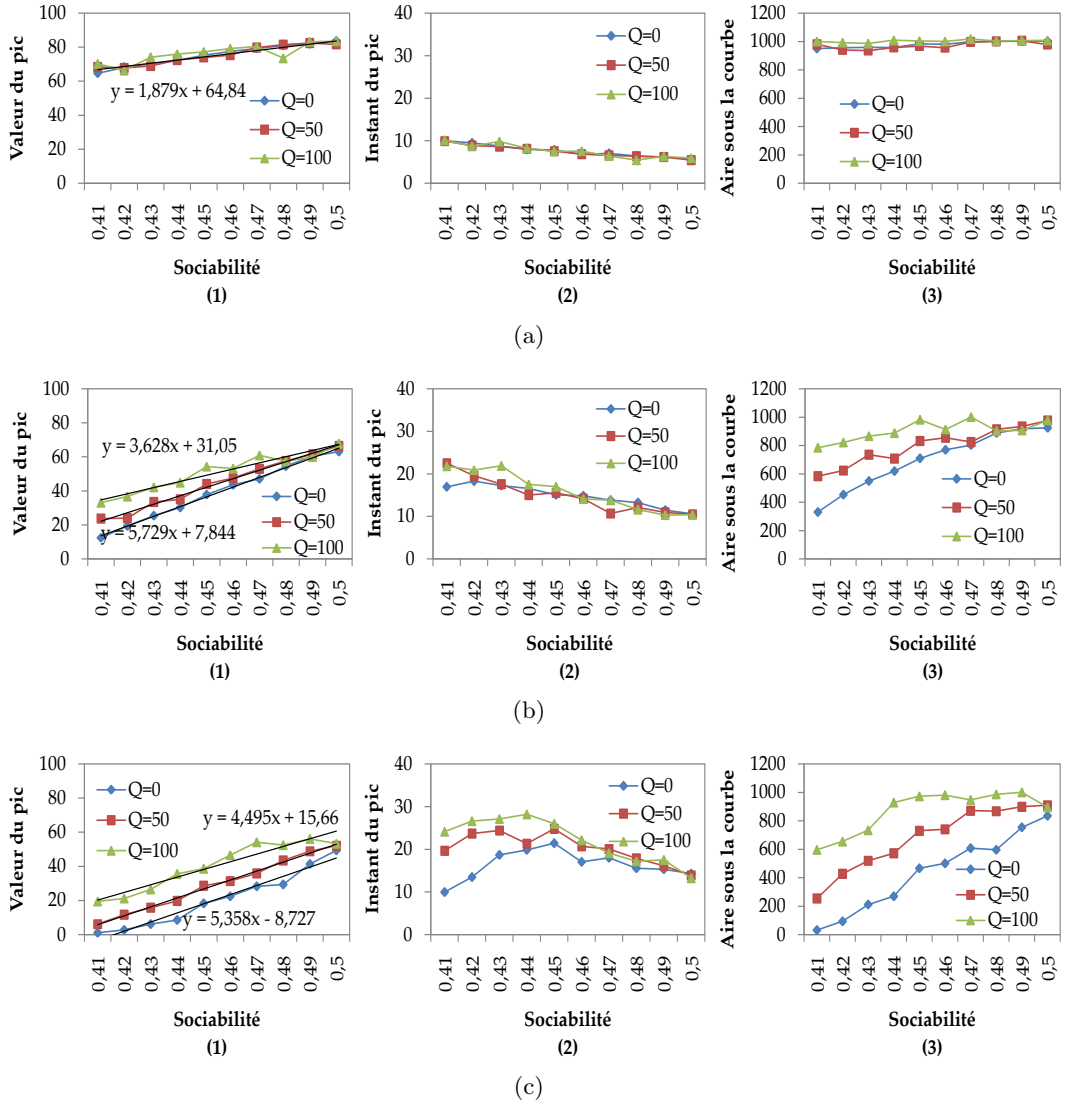


FIGURE 3.14 – Évolution des caractéristiques de la diffusion avec *DynBPDA* (1) VP, (2) TP et (3) AUC quand (a) $dst = 100$, (b) $dst = 500$ et (c) $dst = 1000$

lité faible, la valeur du pic est plus importante pour les vitesses d'évolution élevées, alors que l'écart se réduit significativement quand la sociabilité augmente. Par exemple, quand $\beta = 0.41$, VP est d'environ 12.40% sans évolution, contre 23.65% pour $Q = 50$ et 33.74% pour $Q = 100$, alors qu'à partir de $\beta = 0.48$, VP est d'environ 60% pour toutes les vitesses d'évolution. Enfin, pour un niveau élevé de diversité (voir Figure 3.14(c)(1)), la dynamique du réseau semble avoir toujours un impact sur le processus de diffusion, puisque de fortes variations peuvent être observées sur VP. Comme pour une diversité moyenne, la valeur du pic est plus importante pour des vitesses élevées. Par exemple, VP est d'environ 5% quand $Q = 50$, contre 20% avec $Q = 100$.

En ce qui concerne le temps d'apparition du pic (voir Figures 3.14(2)), nous observons globalement que TP décroît quand la sociabilité augmente et que la diversité diminue.

Cela signifie que le taux d'infection maximum est atteint plus rapidement avec un degré de sociabilité élevé et une faible diversité.

Enfin, comme nous l'avons déjà observé dans l'étude précédente, l'aire sous la courbe connait les mêmes variations que la valeur du pic (voir Figures 3.14(3)).

Finalement, bien que nous ayons pu observer que l'augmentation de la sociabilité des individus tend à accroître la valeur du pic épidémique (voir Figures 3.14(1)), nous observons également que l'impact des changements survenant sur le réseau est lui, dépendant du taux de diversité dans l'espace social.

Ainsi, ces résultats suggèrent l'implication du degré de diversité dans l'existence d'un phénomène de transition de phase au sein de la structure du réseau. Une transition de phase est une situation dans laquelle un système subit un changement brutal d'état lié à une valeur critique d'un paramètre clé, appelé *seuil de transition*. Dans ce contexte, nous formulons l'hypothèse qu'il existe une valeur de diversité à partir de laquelle l'effet des changements survenant sur le réseau devient observable.

Pour étudier cette transition, nous avons cherché à comprendre comment évolue la pente des courbes représentant la valeur du pic quand la diversité varie (selon l'observation de la croissance linéaire faite sur les Figures 3.14(1)). La Figure 3.15 compare l'évolution de cette pente pour les différentes vitesses d'évolution.

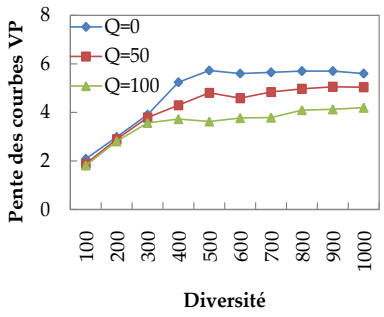


FIGURE 3.15 – Évolution des pentes des courbes VP selon la diversité

La Figure 3.15 met en évidence ce seuil de transition. En effet, nous pouvons observer que pour un taux de diversité compris entre 100 et 300, les changements survenant sur le réseau n'ont aucun effet sur le processus de diffusion. Pour toutes les vitesses utilisées, la valeur du pic peut être approchée par

$$VP_Q = (0.9124 \times dst + 1.1691) \times \beta + c$$

Ce n'est qu'à partir d'une diversité de 400 que les effets de la dynamique du réseau deviennent observables. Par exemple, pour (i) le réseau sans évolution, (ii) $Q = 50$ et (iii) $Q = 100$, la valeur du pic peut être approchée respectivement par :

$$\begin{aligned} VP_{(i)} &= (0.0089 \times dst + 5.6934) \times \beta + c \\ VP_{(ii)} &= (0.0761 \times dst + 4.3172) \times \beta + c \\ VP_{(iii)} &= (0.1206 \times dst + 3.1106) \times \beta + c \end{aligned}$$

Il est assez difficile d'expliquer ces résultats analytiquement, puisque comme expliqué lors de la présentation du modèle *DynBPDA*, les propriétés du réseau restent relativement stables durant son évolution.

Ces résultats ont montré que la sociabilité des individus tend à augmenter la valeur du pic et à accélérer son apparition. Cela peut s'expliquer par le fait que quand la sociabilité

croît, les individus tendent à créer plus de liens, et donc potentiellement avec des individus beaucoup plus éloignés dans l'espace social, ce qui facilite la diffusion et accélère le processus.

Cependant, nous avons également observé que la dynamique du réseau traduite par les mouvements dans l'espace social, n'influencent pas nécessairement le phénomène de diffusion. En effet, nos expériences ont montré qu'il existe un seuil de transition, lié au degré de diversité dans l'espace social, à partir duquel l'effet de la dynamique est observable.
Quand il y a peu de diversité dans l'espace, les individus sont dans un espace social plus confiné. Ainsi, même avec des degrés de sociabilité faibles, les individus peuvent avoir des liens avec d'autres individus situés loin dans l'espace social. Dans une telle configuration, quels que soient les mouvements effectués par les individus dans l'espace, la structure du réseau est telle qu'elle permet une diffusion forte et rapide.
En revanche, quand la diversité augmente, la densité globale et le degré moyen décroissent puisque les individus sont distribués dans un espace social plus grand. Ainsi, la dynamique du réseau permet potentiellement à certains individus de traverser le réseau et de diffuser d'un bout à l'autre lors de l'évolution.

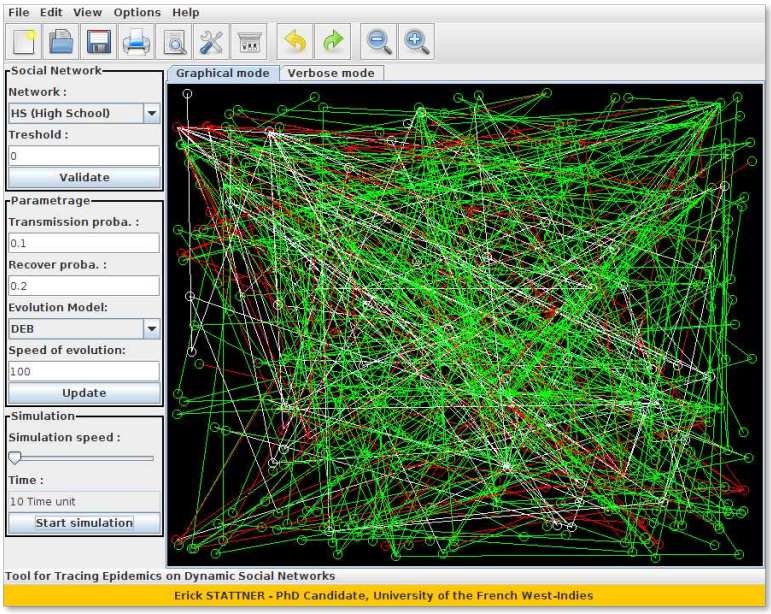
3.5 L'outil graphique *DynSpread*

Les outils traditionnellement proposés pour simuler les processus de diffusion sur les réseaux sociaux n'incluent pas la possibilité de gérer des réseaux en évolution [Corley 2008, Salathe 2010a]. Pour cette raison, nous avons développé *DynSpread*[1] (**Dyn**amics related to **Spread**ing), un outil graphique qui implémente le modèle *D2SNet* et qui vise à étudier tous les aspects liés aux diverses dynamiques impliquées dans les phénomènes de diffusion. Notre objectif avec *DynSpread* est de fournir un environnement de simulation complet, capable de reproduire tous types de diffusion sur des réseaux en évolution. Par exemple, tous les résultats présentés dans les études précédentes ont été obtenus avec cet outil.

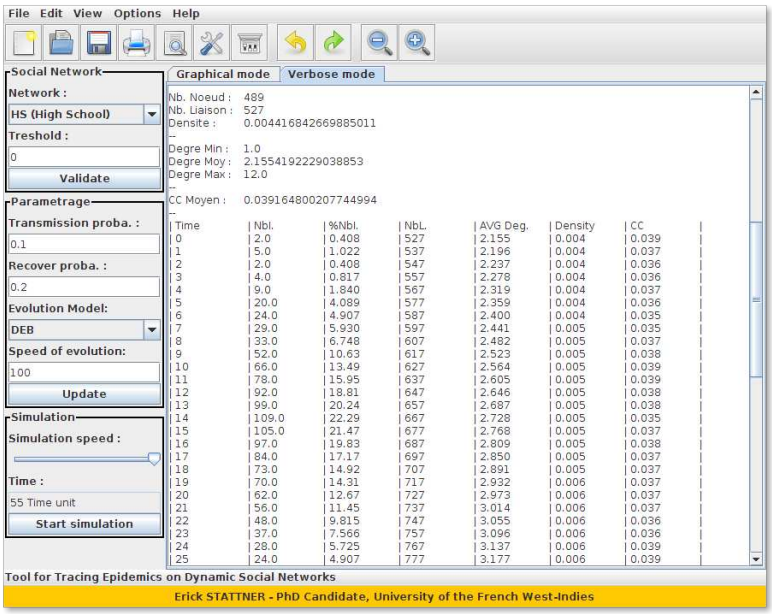
Dans sa première version, l'outil a été restreint au modèle de diffusion *SIR* et est capable de reproduire des scénarios d'infection sur des réseaux dynamiques selon différents critères. À notre connaissance, *DynSpread* est le premier outil qui permet d'appliquer divers modèles d'évolution à un réseau et d'y intégrer un processus de diffusion. L'outil a été développé en JAVA et est composé de deux panneaux principaux.

1. Le panneau de configuration situé à gauche de la fenêtre est utilisé pour charger le réseau, appliquer un filtre sur les liens (dans le cas de liens possédant des valeurs, telles que des fréquences, etc.) et calibrer la simulation. Par exemple, nous pouvons observer sur les Figures 3.16 (a) et (b) que ce panneau permet à l'utilisateur de définir la probabilité de transmission par contact α, la probabilité de guérir β, le modèle d'évolution Ψ qui sera appliqué au réseau et la vitesse d'évolution associée Q. De nombreux modèles d'évolution ont déjà été implémentés dans *DynSpread*, parmi lesquels des mécanismes simples tels que la formation aléatoire, la fermeture triadique, la connexion globale et l'attachement préférentiel, ainsi que des modèles plus complexes tels que *DEB* [Davidsen 2002], *S-DEB* [Stattner 2012j], *MVS* [Marsili 2004], *KOSKK* [Kumpula 2007], et le modèle spatial *DynBPDA* [Stattner 2012e]. Plus généralement, l'outil est également capable de lire des informations sur la dynamique des liens à partir d'un fichier.
La vitesse de simulation peut également être définie sur ce panneau. Cette option permet à l'utilisateur de réduire la vitesse de simulation de façon à pourvoir observer, à chaque itération, l'évolution du *réseau d'états* (état des noeuds, évolution liens du réseau, caractéristiques du réseau, etc.). Enfin, une fois le calibrage effectué, c'est ce panneau qui permet

1. *DynSpread* : www.erickstattner.com/DynSpread



(a)



(b)

FIGURE 3.16 – Capture des deux principales vues de *DynSpread* (a) mode visualisation du *réseau d'états* et (b) mode détaillé

de lancer la simulation et d'observer le temps écoulé.

2. Le panneau de contrôle situé au centre de la fenêtre permet de suivre la simulation soit à travers une vue $2D$ qui montre en temps réel l'évolution du réseau d'états (voir Figure 3.16 (a)) ou par un mode détaillé (verbose mode) qui fournit à chaque itération

des informations quantitatives sur l'évolution du réseau et le processus de diffusion (voir Figure 3.16 (b)). Le basculement entre les deux modes se fait en utilisant les onglets situés au-dessus de ce panneau central.
Plus précisément, la vue $2D$ offre un affichage aléatoire du réseau, puisqu'aucun algorithme de visualisation n'a pour l'instant été implémenté. Les noeuds peuvent toutefois être déplacés en utilisant la souris pour un affichage plus convivial. Sur cette vue, chaque couleur correspond à un état : vert pour les noeuds *Susceptible*, rouge pour les *Infected* et blanc pour les *Recovered*. Le même jeu de couleur est utilisé pour les liens : vert quand le lien est entre deux individus *Susceptible*, rouge quand il connecte un *Infected* à un *Susceptible* ou deux *Infected* et blanc quand le lien implique un *Recovered*.
Avec le mode détaillé, il est possible de suivre parallèlement l'évolution des caractéristiques du réseau et du processus de diffusion durant la simulation. A chaque itération, l'outil maintient une table fournissant des informations sur (i) l'instant de la mesure, (ii) le nombre de noeuds infectés, (iii) son équivalent en pourcentage, (iv) le nombre de liens du réseau, (v) le degré moyen, (vi) la densité du réseau, (vii) le coefficient de clustering, etc. Après la simulation, cette trace peut ensuite être sauvegardée en utilisant les différentes options offertes par la barre d'outils.

Nous précisons que notre outil est suffisamment souple pour y inclure des options additionnelles telles que des techniques d'analyse [Fortunato 2009], des algorithmes d'affichage [Croft 2008b], ou des méthodes de fouilles de données [Getoor 2005]. Il serait par exemple intéressant de rechercher s'il existe des corrélations entre les caractéristiques d'un individu et sa probabilité d'être infecté. Plus généralement, l'analogie existante entre le paradigme épidémique et de nombreux autres domaines permet à notre outil d'être également pertinent pour l'étude d'autres types de diffusion. Typiquement, *DynSpread* pourrait être utilisé pour étudier la diffusion d'un virus dans un parc informatique.

Nous pensons que l'utilisation de *DynSpread* par des scientifiques ou des professionnels de santé pourrait permettre d'améliorer la compréhension des processus de diffusion dans de nombreux domaines. À terme, un tel outil pourrait ainsi permettre de proposer des stratégies d'interventions adaptées qui tiendraient compte de la dynamique des réseaux, un aspect encore trop souvent négligé.

3.6 Conclusion

L'approche "réseau" est actuellement la plus répandue pour comprendre comment un processus de diffusion prend place et évolue dans une population. Bien que de nombreux travaux aient été menés pour étudier et comprendre ces phénomènes, nous observons que la plupart font l'hypothèse de structures statiques, ce qui ne reflète pas la réalité des réseaux du monde réel qui ont un caractère dynamique.

Dans ce chapitre, nous avons abordé la question de la diffusion dans les réseaux sociaux dynamiques, en cherchant à comprendre l'effet de leurs changements topologiques sur les processus de diffusion. Les contributions de ce chapitre peuvent être résumées comme suit :

(i) En partant de l'observation d'analogies entre les différents types de diffusion, tels que la diffusion de maladies, de rumeurs ou de virus informatiques, nous avons proposé **le modèle unifié *D2SNet*** pour modéliser les phénomènes de diffusion sur des réseaux en évolution, en combinant les deux types de dynamiques impliqués : la dynamique de la diffusion et celle du réseau sous-jacent. Notre modèle est suffisamment générique pour intégrer à la fois différents types de diffusion, mais également toutes formes d'évolution du réseau qui tiendraient compte de facteurs endogènes ou exogènes.

(ii) Une première étude a mis en application ce modèle pour comprendre l'impact sur le

processus de diffusion, de **mécanismes de formation de liens très simples**, identifiés comme étant à la base de la formation de liens dans de nombreux réseaux sociaux. Cette première étude a pu mettre en évidence deux résultats intéressants. (1) La fréquence des changements survenant sur le réseau influence certaines caractéristiques de la diffusion, notamment le pic épidémique, son apparition dans le temps et la couverture globale du phénomène. (2) Pour un même ajout de liens, l'effet sur ces caractéristiques varie selon le type d'évolution considéré. En effet, en cherchant à expliquer les tendances observées en s'intéressant à l'évolution des propriétés du réseau, nous avons par exemple pu observer qu'une évolution qui tend à renforcer l'effet communautaire au sein du réseau produit une diffusion dont le pic épidémique et la couverture globale restent faibles, mais dont l'apparition est relativement précoce.

(iii) Une seconde étude a de nouveau mis en application le modèle pour comprendre comment **une dynamique du réseau plus avancée**, qui considère l'environnement social des individus comme le principal moteur des changements observés sur le réseau, impacte le processus de diffusion. Alors que la première étude suggérait un impact systématique de la dynamique du réseau sur la diffusion, les résultats obtenus dans cette étude ont pu mettre en évidence que dans un contexte plus réaliste, où la formation des liens est induite par l'environnement social des individus, l'impact de la dynamique est dépendant du niveau de diversité dans cet environnement.

(iv) Une première version de la solution a été implémentée à travers **l'outil graphique *DynSpread***, qui permet de simuler la diffusion d'une maladie sur un réseau social dynamique. Notre outil peut être utilisé pour simuler une diffusion sur tout type de réseau et intègre déjà un ensemble de stratégies d'évolution proposées dans la littérature (AL, FT, CG, PA, DEB, $S\text{-}DEB$, MVS, $KOSKK$, $DynBPDA$, etc.)

Ces premiers résultats confirment qu'il est pertinent de s'intéresser à la dynamique des réseaux dans l'étude des phénomènes de diffusion. Notre approche a par exemple des implications concrètes pour l'étude de la diffusion des maladies infectieuses ou des virus informatiques. D'une part *DynSpread* peut aider les scientifiques, et notamment les professionnels de santé, dans une démarche d'amélioration de la compréhension des mécanismes de diffusion au sein des réseaux du monde réel. D'autre part, une application intéressante de notre outil serait de concevoir et de mesurer l'effet de nouvelles stratégies d'intervention, qui tiendraient compte de la dynamique des réseaux.

Plus généralement, nous commençons aujourd'hui à disposer d'ensembles significatifs de données sur des phénomènes de diffusion réels qui devraient nous permettre de confronter les données réelles à celles obtenues en simulation.

Chapitre 4

Mobilité humaine et phénomènes de diffusion : une approche multi-agents

Sommaire

Dans le chapitre précédent, nous nous sommes intéressés aux phénomènes de diffusion sur des réseaux dynamiques. Nous avons ainsi pu mettre en évidence l'impact manifeste de la dynamique du réseau sur le processus de diffusion, en simulant des stratégies d'évolution synthétiques qui ne reflètent que très simplement les comportements humains réels conduisant aux changements topologiques observés dans les réseaux du monde réel. D'une façon générale, ces stratégies modélisent ce qui se passe au niveau structurel, sans véritablement s'intéresser aux origines de ces changements. Or, les modifications qui interviennent dans les interactions sociales d'un groupe d'individus sont manifestement dues à des processus complexes d'ordre psycho-sociologiques, démographiques, politiques, etc. Parmi les comportements potentiellement facteurs de changement dans les relations sociales, nous nous sommes intéressés à la mobilité géographique, qui par ailleurs a fait l'objet d'études récentes assez nombreuses [Camp 2002, Ekman 2008, Gonzalez 2008, Belik 2011].

La mobilité des individus est en effet une dimension transversale à toute pratique sociale, qui peut impacter en profondeur un réseau d'interactions. La mobilité géographique des individus en particulier, permet de tisser des liens sociaux très variés, basés sur la proximité spatiale. Ce type de contacts de proximité entre individus a été activement étudié ces dernières années, du fait de son rôle important dans l'échange et la propagation d'informations. Il est par exemple admis que les individus évoluant dans des espaces étroits sont plus

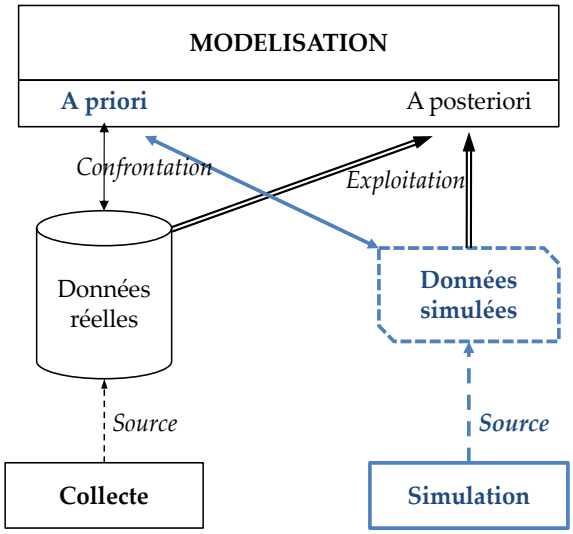


FIGURE 4.1 – Des données sociales aux modèles : modèles *a priori* et données simulées

susceptibles de développer des relations d'amitié et de créer une dynamique de contacts favorisant la diffusion de rumeurs, d'informations ou même de maladies.

Des travaux récents se sont ainsi intéressés à la mobilité des êtres humains et à ses effets à la fois sur un réseau de contacts, mais également sur les processus de diffusion. Par exemple, des modèles de mobilité synthétiques, tels que des *marches aléatoires* ou des *points de passage aléatoires*, ont fourni un support à la conception de réseaux ad hoc mobiles (*MANETs*) et de protocoles de communication adaptés [Camp 2002]. Cependant, les résultats obtenus à partir de dispositifs de géo-localisation, de téléphones portables, ou même inférés à partir de transactions bancaires, ont montré que les déplacements humains n'étaient pas tout à fait aléatoires, mais s'apparentaient, au contraire, à des schémas spatio-temporels réguliers et reproductibles [Gonzalez 2008, Song 2010, Belik 2011].

Ainsi, dans ce chapitre, qui se situe dans le contexte de la diffusion dans des réseaux dynamiques, nous nous intéressons au cas spécifique où la dynamique du réseau est induite par la mobilité géographique et le lien social est dépendant de la proximité spatiale des individus. Ces travaux s'inscrivent dans l'axe de la modélisation des phénomènes prenant place sur les réseaux sociaux à partir de données simulées (voir Figure 4.1). Notre objectif est de proposer un modèle de mobilité humaine réaliste (1) tenant compte du caractère récurrent observé dans les déplacements humains et (2) utilisable pour répondre à des questions complexes telles que :

1. Comment la mobilité des individus, influence t-elle la diffusion ?
2. Quelles sont les configurations qui favorisent ou non le phénomène de diffusion ?
3. Quelles sont les conditions qui garantissent l'émergence du phénomène de diffusion quand les agents sont en mouvement ?

Pour répondre à ces questions, nous adoptons une approche multi-agents et proposons le modèle de mobilité *Eternal-Return* (*ER*), qui représente synthétiquement les régularités spatio-temporelles observées sur les déplacements des êtres humains.
Dans le modèle, chaque agent est caractérisé par un unique attribut *mobilité*, qui représente la façon dont il se déplace dans l'espace. Cet unique paramètre permet ainsi de différen-

cier les agents selon qu'ils soient sédentaires, c.-à-d. à *mobilité faible*, ou voyageurs, à une *mobilité forte*. Le critère de mobilité est donc intrinsèquement lié à la taille de l'espace géographique dans lequel les agents évoluent. Typiquement, un agent sédentaire ne se déplace que dans son voisinage immédiat, alors qu'un voyageur explore une vaste zone avant de retourner à son lieu de résidence.
Notre modèle réduit ainsi considérablement la complexité du monde réel à un comportement social élémentaire. Ce modèle minimal a été intentionnellement adopté afin de ne conserver que les mécanismes pertinents de la mobilité, tout en ayant que très peu de paramètres.
Dans le modèle *ER*, nous supposons que les liens sociaux entre les agents sont uniquement induits par la proximité des agents dans l'espace géographique. La mobilité des individus définit d'une part le nombre de contacts de proximité qu'un agent peut établir à un instant donné, mais également la structure du réseau de contacts sous-jacent qui supporte le phénomène de diffusion. Ainsi, comme dans la réalité, un individu n'est en contact qu'avec une faible proportion de la population [Crooks 2009], c.-à-d. ceux qui sont géographiquement proches de lui. De tels contacts ont d'ailleurs une signification sociale forte, car ils surviennent quand les "*routes se croisent*" et que deux personnes se rencontrent dans un espace spatio-temporel confiné. De telles interactions sont rarement dénuées de sens, puisque chacune d'entre elles est susceptible de transmettre une information, une rumeur, une maladie infectieuse, une nouvelle, etc.

L'intérêt d'un tel modèle est double. (i) Du point de vue du réseau, le modèle de mobilité proposé permet d'étudier l'effet des déplacements des individus sur le réseau de contacts sous-jacents. (ii) Au regard du phénomène de diffusion, le modèle permet de comprendre comment et pourquoi la mobilité des individus influence les processus de diffusion.

Le contenu de ce chapitre est organisé comme suit. La Section 4.1 présente un état de l'art des travaux menés sur les modèles de mobilité. La Section 4.2 décrit formellement le modèle *ER*. Dans la Section 4.3 la pertinence de l'approche est démontrée en étudiant plusieurs aspects fondamentaux liés aux déplacements des individus selon le modèle *ER*. Dans la Section 4.4 nous nous intéressons au problème de la diffusion quand les agents sont en mouvement. Nous montrons en premier lieu qu'il existe des seuils de densité qui garantissent la percolation sur le réseau de contacts sous-jacents. Puis nous étudions l'impact de la mobilité sur le processus de diffusion. La Section 4.5 présente l'outil graphique qui implémente notre solution. Enfin, la Section 4.6 conclut ce chapitre.

4.1 Modélisation de la mobilité humaine : un état de l'art

De nouveaux défis sociétaux tels que l'aménagement urbain, l'amélioration du trafic, ou la gestion de crises sanitaires nécessitent une meilleure compréhension des comportements et des schémas de déplacements des individus dans leur environnement. Cependant, le manque d'outils généraux (juridiques, institutionnels, matériels, etc.) pour suivre ces déplacements a souvent été un obstacle dans l'extraction de toute connaissance à partir de situations réelles.

Dans le domaine de l'informatique, c'est la communauté des réseaux de communication qui a le plus contribué à la compréhension de ces processus. En effet, avec les nouvelles problématiques soulevées par les réseaux de type *MANETs*, liées essentiellement à l'émergence des périphériques mobiles tels que les capteurs sans fil, les téléphones portables, les tablettes ou les puces RFID [1], la mise en place de protocoles adaptés à la mobilité des individus est

1. Radio frequency identification

devenue un enjeu majeur. Pour mesurer efficacement les performances d'un nouveau protocole sur un réseau ad hoc composé de périphériques mobiles, il est impératif de s'intéresser avec précision, à la façon dont les individus se déplacent dans un espace géographique. Ce type d'étude permet par exemple de déterminer dans quelles configurations les protocoles ou algorithmes proposés seront les plus efficaces.

Deux types d'approches sont traditionnellement utilisées, comme nous l'avons rappelé sur la Figure 4.1 : (i) Soit la mobilité des individus est simulée à partir de données réelles, obtenues par exemple grâce aux relevés GPS ou aux traces de connexion par WiFi ou Bluetooth. [Sanchez 2001, Henderson 2004]. (ii) Soit des modèles synthétiques de déplacement sont utilisés pour représenter de façon plus ou moins réaliste, les trajectoires des individus sur une zone géographique.
Pourtant, bien que les traces obtenues à partir de l'approche (i) fournissent souvent des informations précises sur les déplacements des individus, particulièrement quand le nombre d'agents est élevé et que la période d'observation est importante, nous observons que c'est l'utilisation de modèles synthétiques qui est aujourd'hui la plus répandue.

Nous pouvons expliquer cet intérêt pour la modélisation par trois facteurs principaux. Le premier est que la collecte d'informations personnelles sur les utilisateurs pose nécessairement des problèmes de confidentialité, qui rendent souvent très difficile la collecte, l'exploitation et la publication de ces données. Le deuxième concerne la fiabilité de ces données. En effet, quand des données issues de traces sont disponibles, elles sont généralement sémantiquement liées à l'environnement dans lequel elles ont été collectées et ne peuvent donc pas être généralisées à d'autres scénarios. Par exemple, on peut supposer que des traces issues de déplacements dans une école, sont très différentes de traces issues de déplacement dans un centre commercial. Le troisième facteur d'intérêt concerne finalement la valeur ajoutée de ces modèles pour les chercheurs [Jardosh 2003]. Les simulations fournissent en effet des environnements paramétrables, permettant de simuler différents scénarios, allant par exemple des plus chaotiques aux plus favorables, d'isoler certains paramètres pour en étudier les effets, ou même de comparer les résultats obtenus avec différentes configurations pour rechercher des corrélations.

De nombreux modèles de mobilité ont ainsi été proposés. Ces modèles ont souvent pour objectif de reproduire les déplacements d'individus en tenant compte des changements de direction et de vitesse qui surviennent au cours du temps. Les modèles de mobilité les plus répandus sont basés sur un mouvement aléatoire des individus. Intéressons-nous par exemple aux trois principaux.
(i) La marche aléatoire (*random walk*) [Spitzer 2001] est certainement le modèle de mobilité le plus simple et le plus largement utilisé. Dans ce modèle, on suppose qu'un individu se déplace de sa position courante à une nouvelle position en sélectionnant aléatoirement une direction et une vitesse. La vitesse est sélectionnée dans un intervalle prédéfini $[V_{min}..V_{max}]$ et l'angle dans l'intervalle $[0..2\pi]$. Le noeud bouge ainsi dans cette direction et à cette vitesse soit pendant une période de temps donné ou pendant un certain nombre d'itérations. Dans sa version de base, la marche aléatoire suppose que l'espace dans lequel évoluent les individus n'est pas torique. Ainsi, quand un individu atteint une des limites de l'espace, il "*rebondit*" et recalcule une vitesse et une direction.
La Figure 4.2 montre des exemples de marches aléatoires obtenus avec 3 individus (un par couleur) dans un espace en 2D.
(ii) Les points de passage aléatoires (*random waypoints*) [Bettstetter 2004] constituent une extension de la marche aléatoire qui introduit des temps d'arrêt entre les changements de direction et de vitesse. Dans ce modèle, un individu est en premier lieu affecté à une position donnée pendant un certain temps, c.-à-d. le temps d'arrêt. Une fois cette

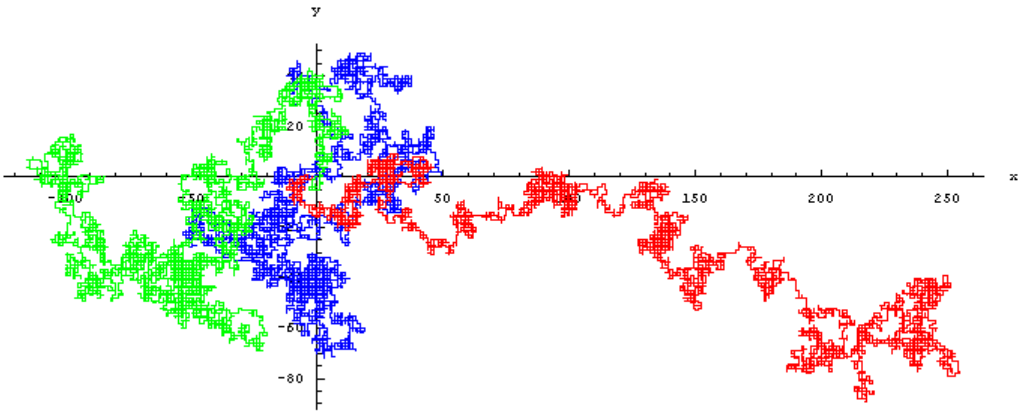


FIGURE 4.2 – Exemple de marches aléatoires obtenues avec trois agents (Source : Wikimedia)

période expirée, l'individu choisit aléatoirement une destination et une vitesse. Il se déplace ainsi jusqu'à sa nouvelle destination, à la vitesse choisie. Une fois la destination atteinte, il marque un temps d'arrêt avant de renouveler le processus en choisissant de nouveau une destination et une vitesse.
Comparé à la marche aléatoire, l'intérêt de ce modèle est qu'il introduit une dimension plus réaliste du fait que les individus passent parfois un temps plus ou moins long dans les lieux qu'ils visitent. En revanche, les traces obtenues sont évidemment sensiblement les mêmes que celles qui seraient obtenues avec une marche aléatoire classique. Il est important de souligner que le modèle est strictement équivalent à une marche aléatoire si le temps d'arrêt est fixé à 0.
(iii) La direction aléatoire (*random direction*) [Camp 2002] est un modèle qui reprend à la fois les approches (i) et (ii). Dans ce modèle, un individu choisit d'abord une direction aléatoire dans laquelle il se déplace jusqu'à atteindre la bordure de la zone. Une fois qu'une des bordures de la zone géographique est atteinte, l'individu marque un temps d'arrêt, puis choisit une nouvelle direction et se déplace jusqu'à atteindre de nouveau l'une des bordures de la zone géographique.
Dans ce modèle, les trajectoires des individus sont parfaitement rectilignes et ont l'aspect de lignes tracées d'un bord à l'autre de l'espace.

En combinant les principes de ces trois modèles fondamentaux, d'autres modèles de mobilité ont été proposés. Par exemple, des modèles basés sur des espaces toriques [Haas 1997], ou des zones urbaines [Camp 2002]. Des modèles de mobilité dits "*de groupes*", peuvent également être trouvés dans la littérature. Leur principale différence avec les précédents réside dans le fait que les déplacements d'un noeud peuvent dépendre de ceux des autres. Une bonne comparaison de ces différents modèles de mobilité peut être trouvée dans [Camp 2002] et [Musolesi 2009].

Cependant, bien que ces modèles aient été largement utilisés pour simuler les déplacements humains, nous observons que les trajectoires obtenues ne correspondent pas véritablement aux mouvements observés dans la vie réelle, ni même aux intuitions qu'on pourrait avoir de ces mouvements. Typiquement, les humains se déplacent rarement dans des directions complètement aléatoires. Dans une école, un lieu de conférence ou un centre commercial par exemple, les individus ont plutôt une destination précise et tente de suivre un chemin bien défini pour atteindre leur destination. Cette analyse est confirmée par des

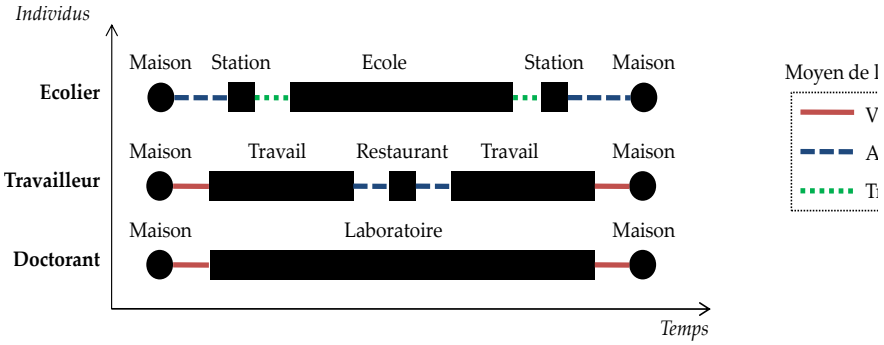


FIGURE 4.3 – Régularités spatio-temporelles des déplacements individuels (Source : Eliot et Daudé [Eliot 2006])

études récentes menées sur des jeux de données réelles. Il apparait en effet que, contrairement aux modèles basés sur la marche aléatoire, les motifs mis en évidence à partir des traces des utilisateurs suivent des schémas cycliques contenant des régularités spatio-temporelles fortes [Daude 2005, Ekman 2008, Song 2010, Belik 2011]

Par exemple, Ekman *et al.* [Ekman 2008] mettent en lumière l'aspect périodique des déplacements en montrant que dans la vie quotidienne certains schémas se retrouvent fréquemment comme le montre la Figure 4.3. En effet, la plupart des personnes vont au travail le matin, passent leur journée au travail, puis rentrent le soir à leur domicile. Les écoliers, eux, se rendent le matin à la station pour attendre le bus, passent leur journée à l'école, se rendent de nouveau à la station pour rentrer chez eux.

Les travaux menés par Belik *et al.* [Belik 2011] font également état de la **périodicité des mouvements**, en expliquant que certaines observations de trajectoires aléatoires sur des traces réelles, pouvaient être dues aux échelles de temps d'observation. Ils généralisent ainsi le concept de régularité en introduisant la notion de *base*, et proposent un modèle dans lequel un agent se déplace de base en base avant de retourner à son point de départ et d'effectuer un nouveau déplacement.

Dans leurs travaux, Daude et Eliot [Daude 2005, Eliot 2006], simplifient les déplacements périodiques en les représentant par des **trajectoires circulaires**. Toutefois, ils ne prennent pas en compte les variations d'amplitude qui surviennent d'un individu à l'autre, une composante pourtant intrinsèquement liée aux comportements humains.
En effet, les travaux menés par Gonzalez *et al.* [Gonzalez 2008] sur des traces de téléphones portables confirment la régularité des trajectoires, mais mettent également en lumière **l'hétérogénéité des déplacements** dans leur amplitude. Ils montrent que certains individus ne se déplacent que dans des espaces très confinés, alors que d'autres explorent de larges zones de leur environnement. Ce résultat confirme la notion intuitive de déplacement dans notre société moderne, puisque nous observons que dans la vie réelle, certains individus sont amenés à voyager sur de très grandes distances, alors que d'autres restent essentiellement confinés à une même ville, une même région ou un même pays.

Le modèle *ER* que nous proposons vise à représenter la régularité spatio-temporelle des déplacements, en tenant compte de tous les aspects mis en lumières par ces différents travaux : (i) la périodicité du mouvement, (ii) la circularité des trajectoires et (iii) l'hétérogénéité des déplacements, qui permet de distinguer les agents selon qu'ils soient sédentaires ou voyageurs.

4.2 Modèle de mobilité "*Eternal-Return*"

Le modèle de mobilité *Eternal-Return* (*ER*) définit un type de mobilité spatio-temporelle qui représente la façon dont les individus se comportent quand ils se déplacent d'un lieu à l'autre. Le concept de mobilité a un sens très large. Il fait souvent référence aux déplacements dans la vie réelle d'agents, d'êtres humains, d'animaux ou même de machines pouvant se déplacer dans des zones de nature ou d'échelle différentes. Dans le contexte des humains, la mobilité peut être sociale ou spatiale. En démographie, à l'échelle des populations, elle est synonyme de *migration*. La mobilité est en effet définie sur la base d'une comparaison entre le lieu de résidence de chaque individu à un instant t et le lieu de résidence un an plus tôt. Selon cette définition, les individus peuvent être classifiés en deux catégories : *migrants* et *non-migrants*. (i) Les *non-migrants* sont ceux qui habitent le même lieu du début à la fin de la période de migration. (ii) Les *migrants* sont, eux, des individus qui vivent dans des lieux différents entre le début et la fin de cette période.

Dans cette étude, nous considérons plutôt la mobilité à l'échelle individuelle comme *la circulation*, qui représente le déplacement d'individus tels que des piétons dans un espace urbain ou inter-urbain. Le modèle de mobilité *ER* est défini dans le but de simuler la tendance qu'ont les êtres humains à retourner aux lieux déjà visités. Ce type de mobilité correspond typiquement à un déplacement de type *domicile-travail*. Plus généralement, il peut être observé dans les expériences de la vie réelle sur les trajectoires humaines, qui sont souvent restreintes par la configuration des lieux (rues, signalisation, bâtiments) et qui contrastent avec le comportement stochastique modélisé lors d'une *marche aléatoire* [Spitzer 2001]. Typiquement, les individus tendent à suivre des chemins prédéfinis et à reproduire des schémas de trajectoire similaires lors de leurs déplacements dans leur environnement urbain [Daude 2005, Gonzalez 2008].

En nous inspirant de ces résultats, nous avons proposé une représentation simplifiée du déplacement des agents. Bien que le modèle de mobilité *ER* soit très restrictif, il s'avère être suffisant pour exprimer le fait que certains agents explorent de grands espaces, alors que d'autres sont confinés à de petites zones. Les agents *sédentaires* (resp. *voyageurs*) sont ainsi définis par de faibles (resp. fortes) mobilités.

Cette section détaille le modèle de mobilité *Eternal-Return*. Localement dans un premier temps, en s'intéressant à la mobilité des agents (voir Section 4.2.1), puis globalement en détaillant le fonctionnement du système dans son intégralité (voir Section 4.2.2).

4.2.1 Mobilité des agents

Puisqu'un modèle est une simplification du monde réel [Epstein 2008], le modèle *ER* peut être vu comme une représentation simplifiée de piétons se déplaçant autour de leur "*monde*" avec une vitesse égale. Chaque agent à sa propre position qui est mise à jour quand il se déplace. L'agent possède également un "*en-tête*" qui indique la direction qu'il suivra pour avancer. L'en-tête de l'agent est une valeur entre 0° et 360°. À chaque pas de temps, chaque agent se déplace tout droit sur une unité. La vitesse de déplacement est ainsi constante et identique pour tous les agents. Entre chaque pas de temps, la nouvelle position d'un agent est déterminée sur la base de sa position actuelle et sa mobilité.

Le modèle *ER* est construit sur trois principes :

1. Les **relations sociales** entre les personnes sont uniquement induites par les rencontres physiques, c'est-à-dire la proximité dans l'espace géographique. La mobilité induit donc un type particulier de liens, et donc de réseau social.
2. La **vitesse de déplacement** ne varie pas et est identique pour tous les agents.

3. Chaque agent a un comportement spatio-temporel **cyclique**, c.-à-d. que pour chaque agent a_i situé à la position p_i^t à l'instant t, il existe T_i tel que $p_i^{t+T_i} = p_i^t$.
   Ce type de comportements a été mis en évidence dans des travaux menés sur des données réelles [Ekman 2008, Belik 2011].

Une illustration concrète de ces principes peut être observée dans la vie courante. En effet, on peut constater que dans la vie réelle, les contacts sociaux sont principalement formés selon les chemins que nous suivons et les lieux que nous fréquentons (travail, école, activités, etc.). Ces chemins ont souvent un caractère cyclique, puisque dans les schémas de déplacements les plus répandus, on observe que les individus se rendent sur leur lieu de travail le matin, peuvent aller déjeuner, puis rentrent chez eux.

Sur la base de ces trois axiomes, nous proposons un **modèle minimal et synthétique** qui suppose que chaque agent se déplace sur un **polygone régulier**. Dans ce modèle de mobilité, le seul critère de différenciation entre les individus est la longueur du polygone ; il permet de considérer les individus selon une échelle qui va des **sédentaires** aux grands **voyageurs**.

Plus précisément, le modèle *ER* définit la trajectoire des agents comme un polygone régulier, avec un sommet à chaque pas de temps. Pour chaque agent a_i, l'amplitude de la déviation à chaque sommet est définie par sa constante *angle extérieur*, notée α_i. Lorsqu'un agent fait le tour du polygone, nous considérons qu'il effectue un tour complet. La somme des angles extérieurs de sa trajectoire est alors égale à 360°.
Soit fTL_i (*full Turn Length*), la longueur du chemin que doit suivre un agent a_i pour revenir à une position donnée (la taille du polygone). fTL_i est ainsi le nombre d'étapes nécessaires pour faire un tour complet. De plus, nous supposons que chaque agent a sa propre direction d_i, c.-à-d. il marche autour de son polygone soit dans le sens des aiguilles d'une montre, ou dans le sens inverse. Dans le premier cas $d_i = +1$, autrement $d_i = -1$.
Pour chaque agent, le fTL est un nombre fixé aléatoirement, et distribué uniformément dans l'intervalle [3..360], lors du processus d'initialisation. Enfin, nous normalisons cette valeur en la divisant par sa valeur maximum, c.-à-d. $\frac{fTL_i}{360}$.

Pour chaque agent a_i, nous définissons sa mobilité μ_i par l'équation suivante :

$$\mu_i = d_i \times \frac{fTL_i}{360} \tag{4.1}$$

La relation entre la *mobilité* et l'*angle extérieur* est donc :

$$\mu_i \times \alpha_i = 1 \tag{4.2}$$

La mobilité μ_i est ainsi un nombre réel appartenant à l'intervalle $[-1, 0[\cup]0, +1]$, et la valeur absolue de l'angle extérieur α_i varie de 1° à 120°. Par conséquent, les agents les moins mobiles se déplacent sur un petit triangle, alors que les plus mobiles suivent un grand polygone possédant 360 cotés. Le cas extrême de mobilité est obtenu pour $\alpha = 0$, c.-à-d. $\mu = \infty$, et correspondrait à une trajectoire linéaire.

L'Algorithme 4 décrit le processus de mobilité *ER*. Nous pouvons observer que chaque agent détermine son propre mouvement selon sa position courante et sa mobilité μ_i.
Bien que dans la vie réelle un même individu puisse vivre et se déplacer dans différentes régions, définies par exemple par son lieu de résidence ou son lieu de travail, nous considérons dans le modèle *ER* une situation beaucoup plus simpliste dans laquelle chaque agent possède une mobilité invariable. Les agents marchent autour d'un polygone régulier et, comme chacun à la même vitesse, leur seul paramètre caractéristique est leur mobilité. Par conséquent, le comportement local de chaque agent est déterministe et périodique.

Sur la Figure 4.4, nous présentons quelques trajectoires d'agents obtenues avec le modèle *ER*. Comme il y a de nombreux agents, et donc des périodes différentes, il est difficile de prédire quand et où les agents se rencontreront dans le même voisinage.

algorithme 4 Déplacement des agents selon le modèle *ER*

Précondition : *agents* : **Liste d'agents**

1. *%Déplacement des agents selon leur mobilité*
2. **pour tout** agent $a_i \in agents$ **faire**
3. $\alpha_i \leftarrow \frac{1}{\mu_i}$
4. **si** $d_i = 1$ **alors**
5. tourner à droite de α_i degrés
6. **sinon**
7. tourner à gauche de α_i degrés
8. **fin si**
9. avancer d'un pas
10. **fin pour**

Pour clarifier la terminologie et permettre de simplifier l'analyse selon la mobilité, nous définissons deux classes d'agents type : les agents *voyageurs* et les agents *sédentaires*. Dans la vie réelle, un individu sédentaire est souvent considéré comme un individu habitant la même localité toute sa vie. À l'opposé, un voyageur est une personne qui se déplace sur des distances plus ou moins grandes.
Enfin, nous précisons que bien que le modèle *ER* ne nécessite qu'un paramètre spécifique par agent, il est réaliste d'une certaine façon puisqu'il permet de différencier les agents *sédentaires* des agents *voyageurs* : les agents sédentaires (resp. voyageurs) étant définis par de faibles (resp. forts) fTl (voir Figure 4.4).

4.2.2 Implémentation

Le modèle *Eternal-Return* a été implémenté avec *NetLogo* [Wilensky 2009, Wilensky 1999, Pham 2004], un environnement de modélisation programmable permettant de simuler des phénomènes naturels et sociaux à base d'agents. Dans *NetLogo*, l'espace est représenté par une grille à deux-dimensions, connectée circulairement de façon à ce que le modèle soit similaire à un automate cellulaire en 2D, dans lequel le "*monde*" comprend plusieurs agents disposés sur une grille torique.

Supposons que le système soit simulé sur une grille $L_1 \times L_2$. La densité des agents δ est un paramètre global du modèle [2]. Ainsi, il y a $L_1 \times L_2 \times (1 - \delta)$ emplacements vides et $L_1 \times L_2 \times \delta$ agents.
Dans le but de garantir des échantillons équivalents quelle que soit la densité, les simulations présentées dans ce chapitre utilisent toutes une population de 1000 agents. La taille de la grille est donc adaptée selon la densité. Par exemple, si $\delta = 10\%$ (resp. 5%), alors $L_1 = L_2 = 100$ (resp. $L_1 = 100$ et $L_2 = 200$).

L'Algorithme 5 détaille le modèle *ER*. A l'étape initiale $t = 0$, les agents sont créés et distribués aléatoirement sur la grille (voir Lignes 2-4 Algorithme 5). Les coordonnées d'une surface unitaire (c.-à-d. une cellule dans le contexte *NetLogo*) sont représentées par des nombres entiers, alors que les coordonnées des agents sont des nombres réels. Ainsi, à un instant donné, plusieurs agents peuvent être situés sur une même cellule.

2. Dans le cas des humains, la densité de population est le nombre d'individus par unité de surface

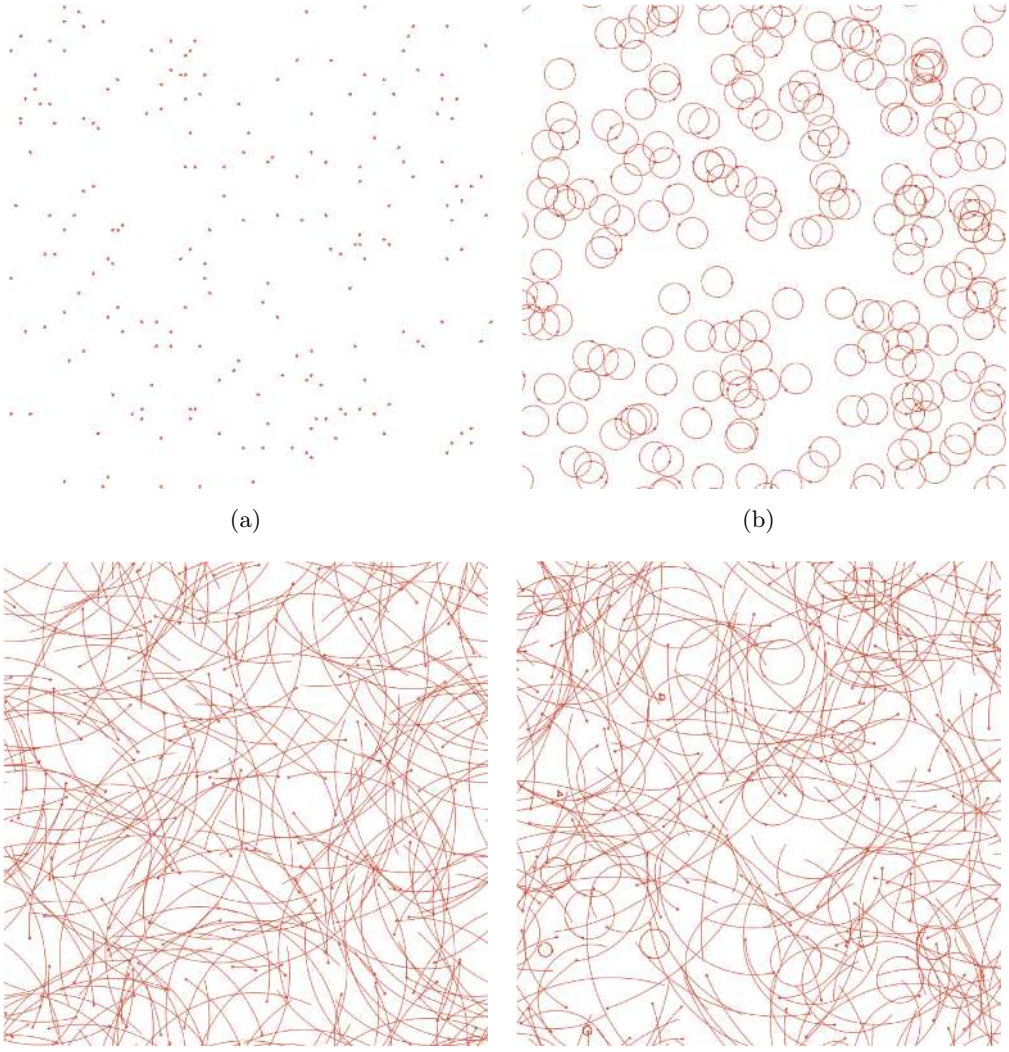


FIGURE 4.4 – Exemple de trajectoires d'agents
Agents : (a) immobiles, (b) sédentaires avec $fTL = 20$, (c) voyageurs avec $fTL = 180$, (d) mobilité mixte

Pour chaque agent, initialement un sens d_i est attribué (voir Ligne 5), et une mobilité μ_i est assignée selon une distribution aléatoire uniforme définie sur fTL (voir Ligne 6-7). Tous les résultats obtenus ont été moyennés sur 100 exécutions, avec la configuration matérielle suivante : Intel Core 2 Duo P8600, 2.4Ghz, 3Go Ram, Linux Ubuntu 10.10 avec Java JDK 1.6.

4.3 Réseau de proximité basé sur la mobilité

Avant d'utiliser le modèle pour l'étude des phénomènes de diffusion, nous devons d'abord étudier les caractéristiques résultant du déplacement des agents. Dans cette section, nous

algorithme 5 Simulation du modèle *Eternal-Return*

Précondition : : L_1, L_2 : **Taille de la grille**, $\delta \in [0..1]$: **densité**

1. t : **temps** $\leftarrow 0$
2. Créer $(L_1 \times L_2 \times \delta)$ agents
3. **pour tout** agent $a_i \in agents$ **faire**
4. Initialiser la position $(x_i(0), y_i(0))$ au hasard
5. Initialiser $d_i \in \{-1, 1\}$ aléatoirement
6. Initialiser $fTL_i \in [3..360]$ uniformément au hasard
7. $\mu_i \leftarrow d_i \times \frac{fTL_i}{360}$
8. **fin pour**
9. **loop**
10. Appeler Algorithme 4($agents$)
11. $t \leftarrow t + 1$
12. **fin loop**

cherchons à comprendre comment évoluent deux caractéristiques clés du modèle : (i) la **répartition** des agents sur les cellules lors de leur déplacement (ii) le **nombre de contacts** effectués sur le trajet. Ces deux caractéristiques sont étudiées selon la mobilité et la densité. Cela nous conduit donc à nous intéresser au réseau sous-jacent de contacts distincts entre les agents pour lequel **la répartition** des agents influence la connectivité et **le nombre de contacts** détermine la distribution du degré.

Ainsi, nous définissons ce réseau dans la Section 4.3.1. Ensuite, nous nous intéressons au nombre de visiteurs par cellule (voir Section 4.3.2) et au nombre de contacts sociaux par agents (voir Section 4.3.3).

4.3.1 Comment la mobilité induit-elle un réseau social dynamique ?

Les réseaux sociaux sont des structures regroupant des individus (*des noeuds*), connectés par un ou plusieurs types spécifiques de relations (*les liens*) possédant une sémantique sociale forte. Ces liens peuvent être de nature différente, telle que des liens d'amitié, d'intérêts communs, des relations intimes, de partage, etc. Dans le modèle *ER*, les agents peuvent être assimilés aux noeuds et leurs liens sociaux sont générés selon la proximité géographique. Ce type d'interactions, basé sur la proximité des individus dans l'espace géographique suscite beaucoup d'intérêt, puisqu'il s'agit d'une généralisation abstraite de multiples contacts réels tels qu'un contact physique, l'échange de mots ou d'information, la participation à un même événement ou la présence sur un même lieu.
En épidémiologie, les réseaux de proximité ont été largement étudiés [Chen 2007b, Salathe 2010b] pour comprendre comment divers types de contacts humains, induits par des comportements sous-jacents tels que la mobilité, facilite ou non le processus de diffusion dans une population.

La mobilité est une composante fondamentale des modèles spatiaux à base d'agents, parce qu'elle définit la configuration du voisinage de l'agent, et donc sa capacité à établir des contacts avec d'autres agents.
Dans le modèle *ER*, la mobilité permet aux agents d'explorer des zones plus ou moins importantes de leur environnement géographique, et a fortiori de générer plus ou moins de contacts de proximité. En effet, nous supposons que deux agents entrent en contact quand ils sont suffisamment proches géographiquement, c.-à-d. la distance entre eux est plus petite

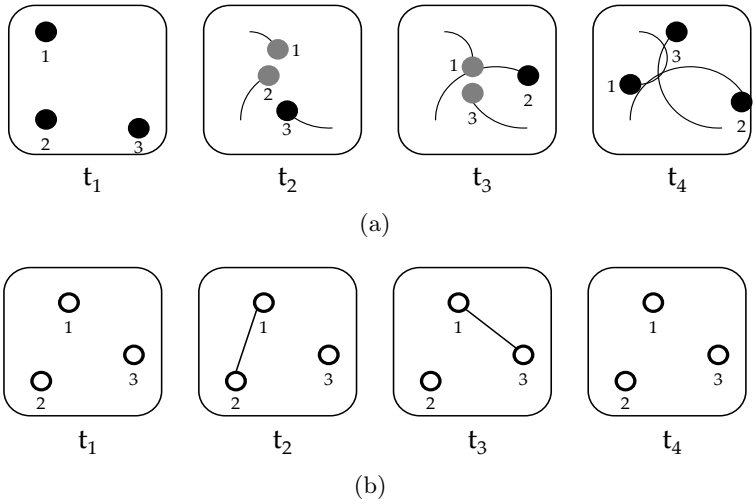


FIGURE 4.5 – Scénarios de création de contacts sociaux dans le modèle *ER*
(a) Trajectoires (b) Liens créés

ou égale à un seuil minimal, fixé à 1 dans nos expériences.

Sur la Figure 4.5, nous montrons comment des contacts de proximité instantanés surviennent et évoluent selon le modèle *ER*.
Typiquement, à t_1, aucun individu n'est suffisamment proche pour créer de contacts. À l'instant t_2, les agents 1 et 2 sont suffisamment proches pour entrer en contact, donc un lien est créé ; alors qu'à l'instant t_3 ce lien est supprimé et un lien est crée entre les agents 1 et 3. À l'instant t_4 plus aucun individu n'est en contact, tous les liens sont ainsi supprimés.
La mobilité conduit ainsi à un réseau de contacts sociaux dont la dynamique est une caractéristique très importante. En effet, à chaque mouvement des agents, de nouveaux contacts peuvent être créés ou supprimés, ce qui peut modifier considérablement la structure globale du réseau à chaque itération.

Nous définissons le réseau social de proximité basé sur la mobilité, $mSPN_T$ (Mobility-Based Social Proximity Network), comme étant le réseau cumulé de tous les contacts de proximité instantanés distincts entre les agents. Plus précisément, le réseau cumulé $mSPN_T$ représente le réseau de contacts de proximité maximum dans lequel chaque agent est lié à tous ceux qu'il rencontre sur son trajet. Ainsi à chaque instant t, le réseau de contacts de proximité instantanés, décrits sur la Figure 4.5(b), est un sous-graphe déconnecté du $mSPN_T$, et représente à l'instant t, les contacts de proximité dans lesquels les agents sont impliqués.
Évidemment, $mSPN_T$ est un réseau beaucoup plus dense que le réseau de contacts de proximité instantanés.

Plus formellement, soit $T = \langle t_0, t_1, ..., t_m \rangle$ la séquence de temps sur laquelle les agents sont en mouvement.
Nous notons le réseau de contacts de proximité instantanés à l'instant t_j par $G_{t_j} = (V, E_{t_j})$, ou V est l'ensemble des agents et E_{t_j} l'ensemble des contacts de proximité survenus à l'instant t_j.
Ainsi, $mSPN_T = (V, E_T)$ est obtenu par l'union de tous les liens des réseaux de contacts de

proximité instantanés, de l'instant t_0 à l'instant t_m, c.-à-d. :

$$E_T = \bigcup_{j=0}^{m} E_{t_j} \tag{4.3}$$

Quand le fTL est constant, le mouvement périodique généré par le modèle *ER* garantit qu'il y a convergence du $mSPN_T$ vers un état de stabilité $G^* = (V, E^*)$, pour le lequel plus aucun contact ne peut être créé.
Cet état de stabilité est atteint quand tous les agents sont retournés à leur point de départ, c.-à-d. quand $t_j = fTL$.

$$E^* = \bigcup_{j=0}^{fTL} E_{t_j} \tag{4.4}$$

Sur la Figure 4.6, trois exemples de $mSPN_T$ obtenus avec $\delta = 15\%$ et (a) $fTL = 3$, (b) $fTL = 10$ et (c) $fTL = 20$, sont représentés.
Comme attendu, la mobilité a un impact direct sur le nombre global de contacts, puisque nous pouvons observer que la densité du réseau s'accroît clairement avec le fTL.

Finalement, le $mSPN_T$ peut être vu comme un réseau qui résume la dynamique de l'ensemble des contacts survenus lors des déplacements des agents. Par conséquent, ses propriétés donnent une idée de ce que pourra être l'évolution d'un phénomène de diffusion lors de déplacements selon le modèle *ER*.
Typiquement, nous pouvons observer que des populations ayant une faible mobilité ne garantissent pas la connectivité du $mSPN_T$ (voir Figures 4.6(a) et (b)). Dans la Section 4.4.1, nous étudions plus formellement ce phénomène et montrons que cela induit une incapacité du système à "*percoler*".

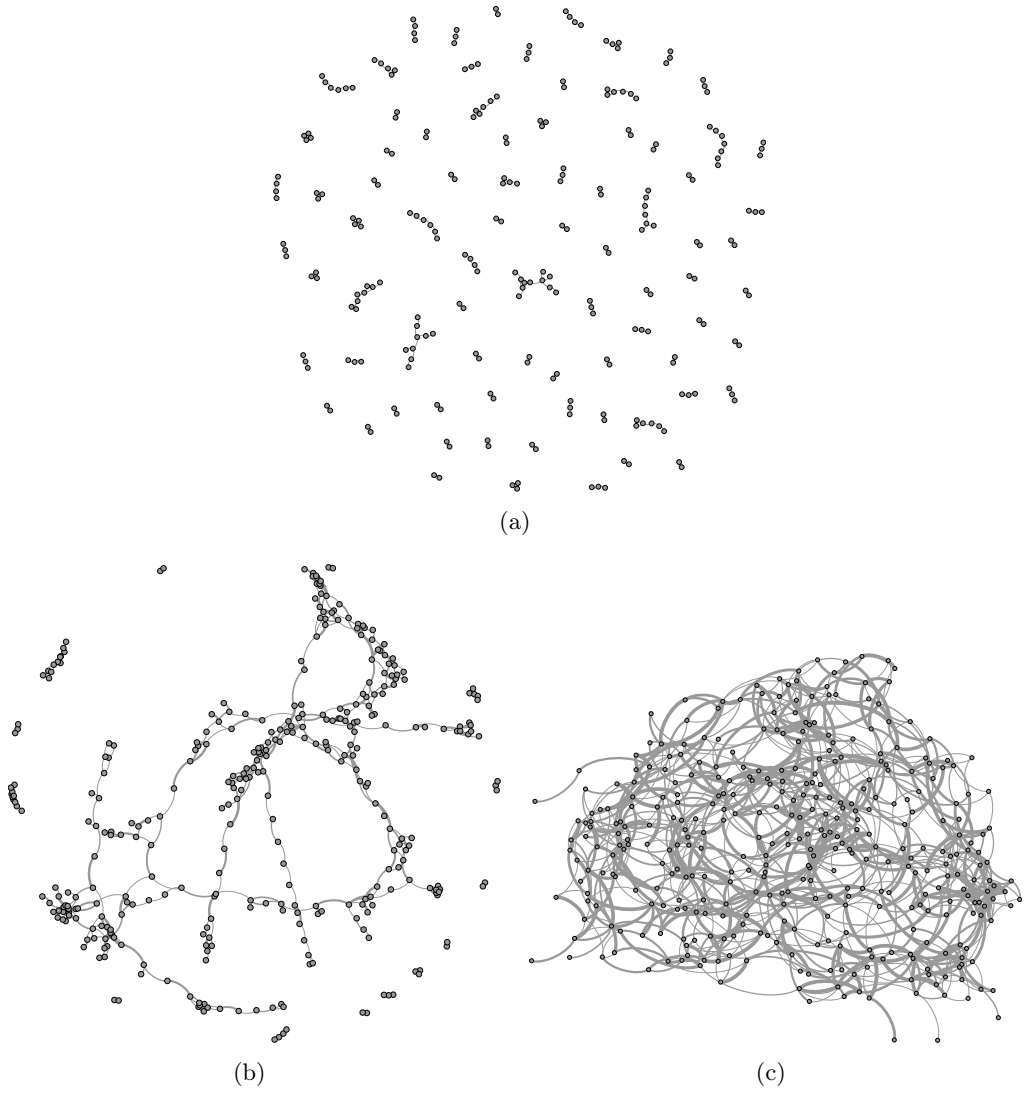
4.3.2 Distribution des agents dans l'espace

Dans le cas d'agents immobiles, si nous supposons qu'il n'y a aucune superposition des agents, le nombre de visiteurs par cellule est 0 ou 1.
Puisque la distribution des agents sur la grille est aléatoire (voir Section 4.2), le nombre moyen de visiteurs par cellule correspond à la densité d'agents δ dans le monde.
Cependant, quand les agents se déplacent dans un espace spatio-temporel, la situation devient plus complexe. En effet, le nombre total d'*agents-visiteurs* dans un intervalle de temps pour une cellule peut, évidemment, être supérieur à 0 et plus.

Pour une cellule c_i, nous définissons V_i comme l'ensemble des agents a_k qui visitent c_i, c.-à-d. $V_i = \{a_k^i\}$. $\#V_i$ est donc le nombre de *trajectoires-polygones* qui passent par la cellule c_i. Ainsi, certaines cellules peuvent ne pas avoir du tout de visiteurs, alors que d'autres peuvent en avoir un nombre élevé.

Dans le cas particulier où la mobilité est identique pour tous les agents, les résultats obtenus par simulation permettent d'établir que $\#V_i$ suit approximativement une *loi de poisson*.
Par exemple, sur la Figure 4.7, nous montrons la distribution du nombre de visiteurs par cellule quand la densité est fixée à 10% et que le fTL est égal à 10 (courbe bleue), 20 (courbe rouge), 100 (courbe verte) et 360 (courbe violette).

Dans ces quatre cas, nous pouvons observer que, quelle que soit la mobilité, les distributions peuvent être approchées par des *lois de poisson*. Ainsi, le nombre de visiteurs $\#V$ sur les zones géographiques suit une loi de poisson avec une queue connaissant une décroissance rapide. Nous observons en effet un pic atteint à $\#V = \langle V \rangle$, puis une décroissance exponentielle avec $\#V$.



(a)

(b) (c)

FIGURE 4.6 – Exemples de *mSPNs* obtenus avec $\delta = 0.15$ quand (a) $fTL = 3$, (b) $fTL = 10$ et (c) $fTL = 20$

Cependant, en comparant les résultats obtenus entre les différentes mobilités, nous observons que la courbe s'aplatit quand la mobilité augmente : la valeur du pic diminue alors que $\langle V \rangle$ augmente. Typiquement, si les agents sont sédentaires ($fTL = 20$), de nombreuses cellules ont peu de visiteurs, alors qu'à l'inverse, dans le cas d'agents voyageurs ($fTL = 360$), les cellules ont globalement un nombre élevé de visites.

Ce résultat constitue un argument fort pour démontrer la validité du modèle de mobilité *ER*. En effet, des études menées en géographie ont montré que les statistiques sur les régions peuvent souvent être décrites par une loi de Poisson quand les individus sont indépendants et que les régions ont des probabilités équivalentes d'être visitées [Simon 2006, Vaillant 2011]. Ainsi, bien que notre modèle repose sur des règles de déplacement très simples, nous observons qu'il est tout de même en mesure de reproduire les résultats observés dans des situations réelles.

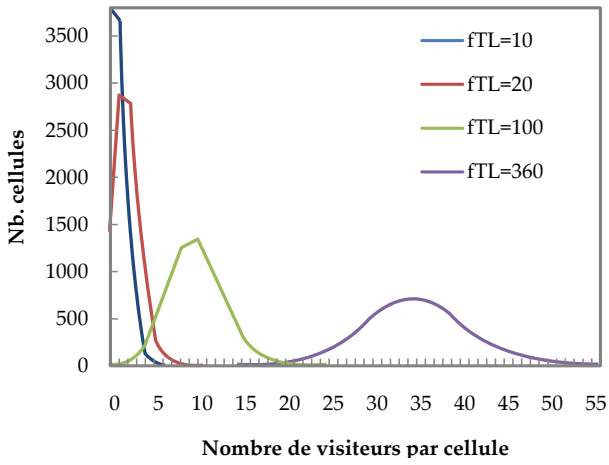


FIGURE 4.7 – Distribution du nombre de visiteurs par cellule quand $\delta = 0.1$ avec $fTL = 10$ (bleu), $fTL = 20$ (rouge), $fTL = 100$ (vert), $fTL = 360$ (violet)

Pour compléter cette étude, nous agrégeons les données obtenues en simulation pour comprendre comment évolue le nombre de visiteurs moyen selon la densité. Sur la Figure 4.8, nous comparons donc pour différents fTL, l'évolution de la moyenne du nombre de visiteurs par cellule, $\underset{i}{moy.}\ \#V_i$, en fonction de la densité δ, avec $\underset{i}{moy.}\ \#V_i = \frac{\sum \#V_i}{L_1 \times L_2}$.

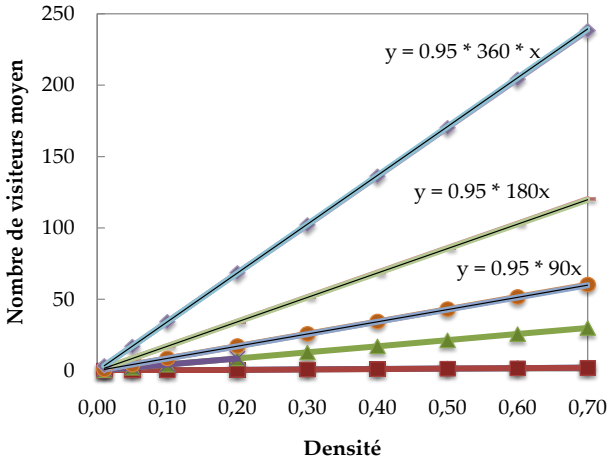


FIGURE 4.8 – Nombre de visiteurs moyen par cellule selon la densité
De haut en bas : $fTL = 360, 180, 90, 45, 3$

Sur cette figure, nous observons une corrélation linéaire entre ces deux variables, puisque le nombre de visiteurs moyen croît linéairement avec la densité. Ces résultats nous ont permis de mettre en évidence l'équation 4.5, qui suggère que le nombre moyen de visites $\underset{i}{mean}\ \#V_i$ est proportionnel au produit du fTL par la densité δ

$$\underset{i}{moy.}\ \#V_i \approx 0.95 \times fTL \times \delta \tag{4.5}$$

Ainsi, nous en déduisons que le nombre moyen de visiteurs est proche de k si $\delta = \frac{k}{0.95 \times fTL}$.

Cet autre résultat fournit un argument supplémentaire de la pertinence du modèle *ER*. La dépendance linéaire entre le nombre de visiteurs moyen et à la fois la mobilité et la densité est en effet attendu, puisque avec un fTL constant pour tous les agents, chaque agent visite fTL cellules. Le nombre de visiteurs sur une cellule devrait donc être égale à $fTL \times \delta$. L'écart de 5% observé expérimentalement peut être expliqué par les co-occurrences d'agents sur les cellules.
De la même façon, on peut établir que l'écart-type est approché par $\sqrt{0.95 \times fTL \times \delta}$.
Nous obtenons finalement :

$$P(\#V = k) \approx \frac{(-0.95 \times fTL \times \delta)^k \times e^{-0.95 \times fTL \times \delta}}{k!} \tag{4.6}$$

Enfin sur la Figure 4.9, nous montrons comment évolue la densité δ des agents selon la mobilité pour différents nombres moyens de visiteurs par cellule. Du haut vers le bas, nous pouvons observer les quatre hyperboles qui correspondent respectivement aux valeurs $\#V = 1$, $\#V = 2$, $\#V = 3$ et $\#V = 4$.

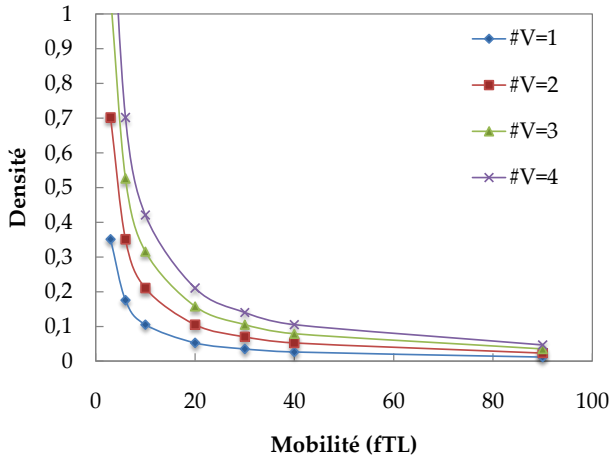


FIGURE 4.9 – Densité δ vs. Mobility fTL pour un nombre moyen de visiteurs donné

4.3.3 Répartition des contacts par agent

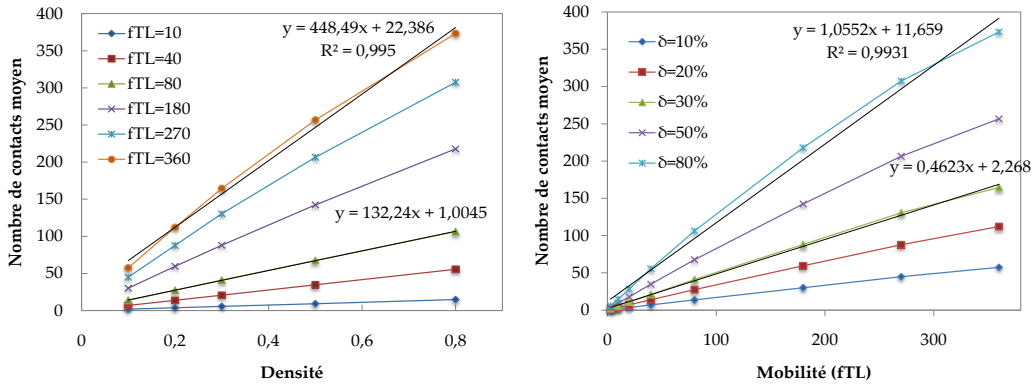
Pour chaque agent, nous considérons son nombre total de contacts dans le temps. Ce nombre correspond au *degré* du noeud correspondant dans le $mSPN_T$. Nous avions déjà pu observer sur la Figure 4.6 que la mobilité impactait directement la connectivité du réseau. Dans cette deuxième étude menée sur le modèle *ER*, nous cherchons à comprendre plus formellement l'influence de la mobilité sur le nombre de contacts.

A un instant donné, un agent a_j est dans le voisinage de l'agent a_i, si a_j est dans le disque circulaire de rayon 1 centré en a_i. Dans une telle situation, nous supposons qu'il y a un *contact social* fortuit entre les agents a_i et a_j.
Ainsi, pour chaque agent a_i, nous notons C_i l'ensemble de tous les agents a_j, avec $j \neq i$, dans son voisinage au cours du temps, c.-à-d. $C_i = \{a_j^i\}$. Le nombre total de contacts de l'agent a_i dans le temps est donc le cardinal de C_i, que l'on note $\#C_i$.
Précisons tout de même qu'un contact social nécessite une forme de *coïncidence spatio-temporelle*, dans le sens où le contact ne survient que si les deux agents sont au même endroit, au même moment.

Dans le cas d'agents immobiles, le nombre de contacts sociaux moyen est donné par l'Équation 4.7. Comme précédemment, ce résultat a été obtenu par des simulations menées sur 1000 agents et moyennées sur 100 exécutions.

$$\underset{i}{moy.}\ \#C_i \approx 0.4 \times \delta \tag{4.7}$$

Sur la Figure 4.10, nous montrons comment évolue le nombre moyen de contacts sociaux selon deux points de vue : (a) en fonction de la densité pour différentes valeurs de mobilité et (b) en fonction de la mobilité pour différentes densités.



(a) (b)

FIGURE 4.10 – Évolution du nombre moyen de contacts sociaux selon (a) la densité pour différents fTL et (b) la mobilité pour différents δ

Nous pouvons observer que si les agents sont en mouvement avec la même mobilité dans le modèle *ER*, le nombre moyen de contacts sociaux s'avère être approximativement proportionnel à la densité des agents comme le montre la Figure 4.10(a).
Inversement, pour une densité donnée, le nombre de contacts sociaux moyen s'avère être proportionnel à la mobilité. On observe en effet sur la Figure 4.10(b) que plus la mobilité augmente et plus le nombre moyen de contacts est élevé.

Enfin, si la mobilité est identique pour tous les agents, les résultats obtenus par simulation permettent d'établir que $\#C_i$ suit une *loi normale* avec la même moyenne et écart-type.

4.4 Mobilité et phénomènes de diffusion

Le modèle *ER* et les contacts de proximité ont été définis et construits dans le but de comprendre comment un processus de diffusion évolue quand les agents sont en mouvement et que le réseau de proximité sous-jacent supportant la diffusion est fortement perturbé par des contacts apparaissant ou disparaissant à chaque instant. Dans cette section, nous détaillons les simulations conduites pour comprendre ce phénomène et présentons les résultats expérimentaux obtenus.
Comme dans le chapitre précédent, le contexte épidémique sert ici de référence puisque les modèles épidémiques sont assez génériques pour étendre nos résultats à d'autres types de diffusion.

En épidémiologie, la diffusion est le passage d'une maladie infectieuse, d'un individu infecté à un individu non-contaminé. Le terme de diffusion fait donc ici référence à la transmission de micro-organismes d'une personne à l'autre, par contacts physiques directs ou par contacts indirects.
Les contacts de proximité définis dans le modèle *ER* font référence à la catégorie des contacts directs. L'intérêt pour ce type de contacts repose sur leur implication dans le processus de transmission réel. En effet, comme nous l'avons expliqué dans la section précédente, il est pertinent de s'intéresser aux contacts de proximité puisque la *proximité spatio-temporelle* fournit un cadre adéquat pour la formation de liens sociaux favorisant la transmission de maladies, d'informations, de rumeurs, d'opinions, etc.

Cependant, nous avons également pu observer dans la section précédente que ces contacts dépendent à la fois de la densité et de la mobilité des agents. La Figure 4.6 a par exemple permis d'observer que la mobilité impactait directement la connectivité du réseau, une condition pourtant nécessaire à l'évolution du phénomène de diffusion.
Une des premières étapes de cette étude menée sur l'impact de la mobilité sur le processus de diffusion à donc consisté à s'intéresser à la structure du réseau et à sa capacité à permettre une diffusion ou pas. En d'autres termes, nous cherchons à comprendre si, malgré les forts changements topologiques survenant à chaque instant sur le réseau, sa structure permet tout de même à un maximum d'agents d'être affectés par le phénomène. Nos résultats expérimentaux montrent que la mobilité des agents est un paramètre décisif qui influence considérablement la densité d'agents nécessaire pour permettre à un phénomène de se propager.

Ainsi, nous proposons de prendre en compte (i) la théorie de la percolation et (ii) les modèles de diffusion épidémiques pour évaluer l'impact de ces paramètres. Cette section est organisée selon ces deux parties. Dans la Section 4.4.1, nous démontrons l'existence d'un seuil de connectivité nécessaire pour garantir la diffusion dans le réseau. Puis dans la Section 4.4.2, nous utilisons le modèle de diffusion *SIR*, pour comprendre l'effet de la mobilité sur le processus de propagation en s'intéressant notamment à l'évolution des courbes d'incidence et à leurs principales caractéristiques. Ces résultats permettent de mettre en évidence les configurations dans lesquelles la diffusion peut être maximisée tout en minimisant le nombre d'agents.

4.4.1 Étude des seuils de percolation

Le paradigme de percolation est utilisé pour l'étude des phénomènes de diffusion dans l'espace. Il permet par exemple d'identifier les seuils épidémiques pour l'invasion, séparant ainsi les configurations dites "*non-invasives*", dans lesquelles seul un faible pourcentage des individus est affecté par le phénomène, des configurations "*invasives*" ou la grande majorité de la population est affectée. Les seuils d'invasion pour les systèmes *hôte-parasite* montrent souvent des transitions très marquées vers les configurations invasives. Typiquement, ces seuils définissent les valeurs critiques de paramètres, à partir desquelles un agent donné est susceptible d'être infecté.

Cependant, les études classiques menées sur les seuils de percolation s'intéressent principalement aux situations dans lesquelles les agents sont statiques. Quand la mobilité est introduite, la situation devient plus complexe. En effet, dans le cas statique, la probabilité d'invasion est contrôlée par un seul paramètre : la proximité dans l'espace qui dépend uniquement de la densité des agents. Avec le modèle de mobilité *ER*, le seuil critique, s'il existe, dépend à la fois de la densité et de la mobilité.

Ainsi, du point de vue de la percolation, l'étude est centrée sur la structure, et non sur la variabilité liée au processus de diffusion lui-même. Les liens sociaux entre les agents

sont non-déterministes et la densité correspond à la probabilité d'avoir un lien entre deux agents. Contrairement au modèle de diffusion classique, si deux agents sont liés à un instant t, le phénomène se propage nécessairement, autrement dit la probabilité qu'un agent infecté transmette à un non-infecté est de 1. Dans ce contexte, les paramètres de percolation sont donc la densité δ des agents et leur mobilité fTL.
Dans nos expériences, nous supposons que tous les agents ont la même mobilité et nous étudions comment les seuils de percolation sont influencés à la fois par la densité des agents, mais également par leur mobilité. Pour ce faire, nous étendons le paradigme de la percolation au cas particulier de la mobilité. Nous supposons que chaque individu peut être dans deux états distincts : *Susceptible* ou *Infecté*. Dans le contexte de la diffusion d'une information ces états seraient définis comme *non-informé* ou *informé*.
Tous les agents sont donc initialement à l'état *Susceptible*, à l'exception d'un agent sélectionné aléatoirement qui est *Infecté*.

Nous posons la valeur de seuil critique δ_c, comme la valeur de densité permettant l'infection d'au moins 50% des agents. En supposant que tous les agents aient la même mobilité, nous avons mené des expériences pour déterminer le seuil de percolation δ_c selon différentes valeurs de fTL. La Figure 4.11 présente ces résultats.

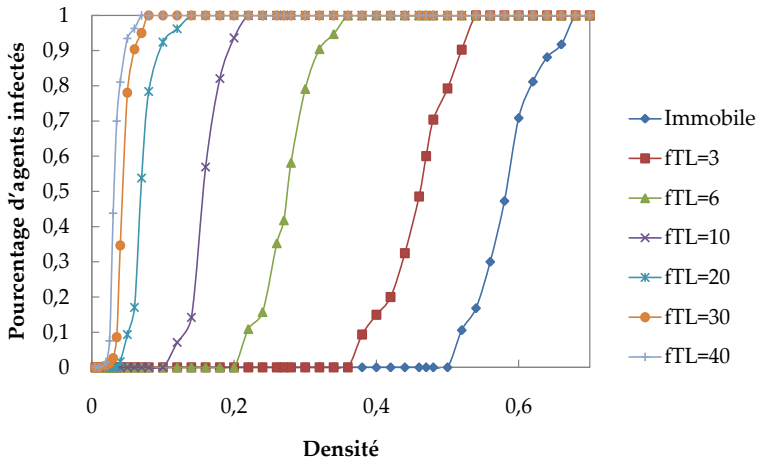


FIGURE 4.11 – Pourcentage d'agents infectés selon la densité pour différentes valeurs de mobilité : immobile, $fTL = 3, 6, 10, 20, 30, 40$

Comme attendu, nous observons que plus la mobilité augmente et plus le seuil de percolation δ_c est obtenu avec des densités faibles. Par exemple, pour les agents immobiles, nous observons $\delta_c = 60\%$, alors que pour $fTL = 3$, nous avons $\delta_c = 50\%$. La courbe représentant l'évolution du δ_c selon la mobilité semble ainsi décroître de façon monotone à partir de 60% (obtenu pour les immobiles), jusqu'à se rapprocher asymptotiquement de 0, comme le montre la Figure 4.11.
Notons d'ailleurs que la décroissance est importante. Par exemple, le seuil de percolation est de $\delta_c = 60\%$ pour des agents immobiles, alors qu'il passe à $\delta_c = 5\%$ pour des agents ayant une mobilité relativement faible ($fTL = 30$).
Par conséquent, nous observons que l'impact de la mobilité sur la capacité à diffuser (rumeur, maladie, etc.) est significatif. La mobilité des agents tend donc à amplifier la capacité du réseau à diffuser.

Toutefois, nous observons que le seuil de percolation δ_c décroît avec la mobilité en

suivant approximativement une *loi de puissance*. Plus précisément, la Figure 4.12 montre que la fonction δ_c selon la mobilité fTL obéit plus ou moins à la forme suivante :

$$\delta_c(fTL) \approx 1.625 \times fTL^{-1.043} \tag{4.8}$$

Cette relation permet (i) d'approcher le seuil de percolation si la mobilité est connue, ou (ii) inversement, pour une densité donnée, de calculer la mobilité nécessaire pour permettre la percolation (voir Figure 4.12 courbe bleue).
Intuitivement, avec un fTL mixte, on peut penser que seuls quelques voyageurs (agents avec un fTL élevé) devraient être nécessaires pour garantir la percolation.

Rappelons que la valeur de densité $\frac{k}{0.95 \times fTL}$ correspond au cas de k visiteurs par cellule en moyenne (voir Équation 4.5). Ainsi, comme on peut l'observer sur la Figure 4.12, quelle que soit la mobilité, le seuil de percolation δ_c est plus grand que $\frac{1}{0.95 \times fTL}$ et plus petit que $\frac{2}{0.95 \times fTL}$. Cela signifie que le réseau *percole* quand le nombre de visiteurs moyen par cellule est compris entre 1 et 2.
Plus précisément, en utilisant l'Équation 4.8 nous pouvons estimer à 1.54 le nombre de visiteurs moyen par cellule pour garantir la percolation.

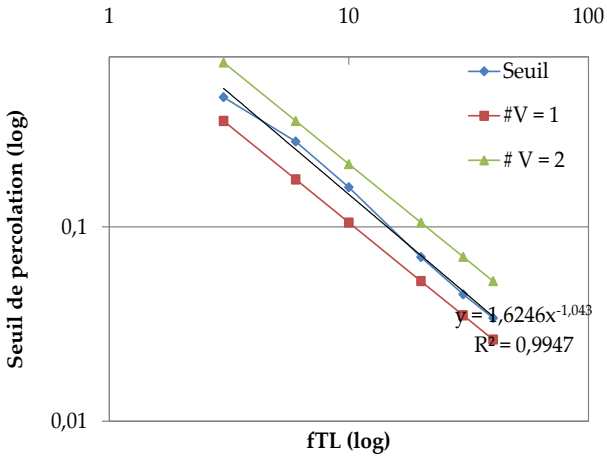


FIGURE 4.12 – Evolution de la densité (log.) selon la mobilité (log.) quand $\#V = 1$ (courbe rouge), seuil de percolation (courbe bleue), et $\#V = 2$ (courbe verte)

4.4.2 Mobilité et diffusion

Traditionnellement, les phénomènes de diffusion sont étudiés sur des réseaux statiques, dans lesquels ni les agents, ni les liens ne subissent d'évolution. Dans le chapitre précédent (voir Section 3.1) nous avons pu voir que même dans ce cas, il est difficile d'obtenir des données réelles sur l'évolution de ces phénomènes. Des simulations sont souvent utilisées pour analyser et comprendre ces processus.
Ce point de vue "*statique*" est très restrictif, puisqu'il ne permet pas de prendre en compte la réalité des comportements humains, qui conduisent souvent à la formation ou à la disparition de liens dans le réseau.

Le modèle *ER* intègre cette dimension, à travers une dynamique du réseau induite par la mobilité des agents. En effet, quand la mobilité est introduite, la dynamique est une com-

posante inhérente au réseau, puisque des contacts de proximité entre les agents se créent et s'évaporent de façon récurrente. Le réseau de contacts instantanés sous-jacent fournit ainsi un support très dynamique pour la diffusion de divers phénomènes tels que des rumeurs, des maladies, ou plus généralement des informations.
Dans un premier temps, nous avons montré comment la mobilité des agents impactait directement la structure du réseau de contacts sous-jacent (voir Section 4.3). Dans la Section 4.4.1, nous nous sommes intéressés à la structure de ce réseau et avons montré que pour que la diffusion y soit possible, un seuil de connectivité devait être atteint. Dans cette partie, nous cherchons donc à comprendre comment la dynamique, induite ici par la mobilité, influence le processus de diffusion.

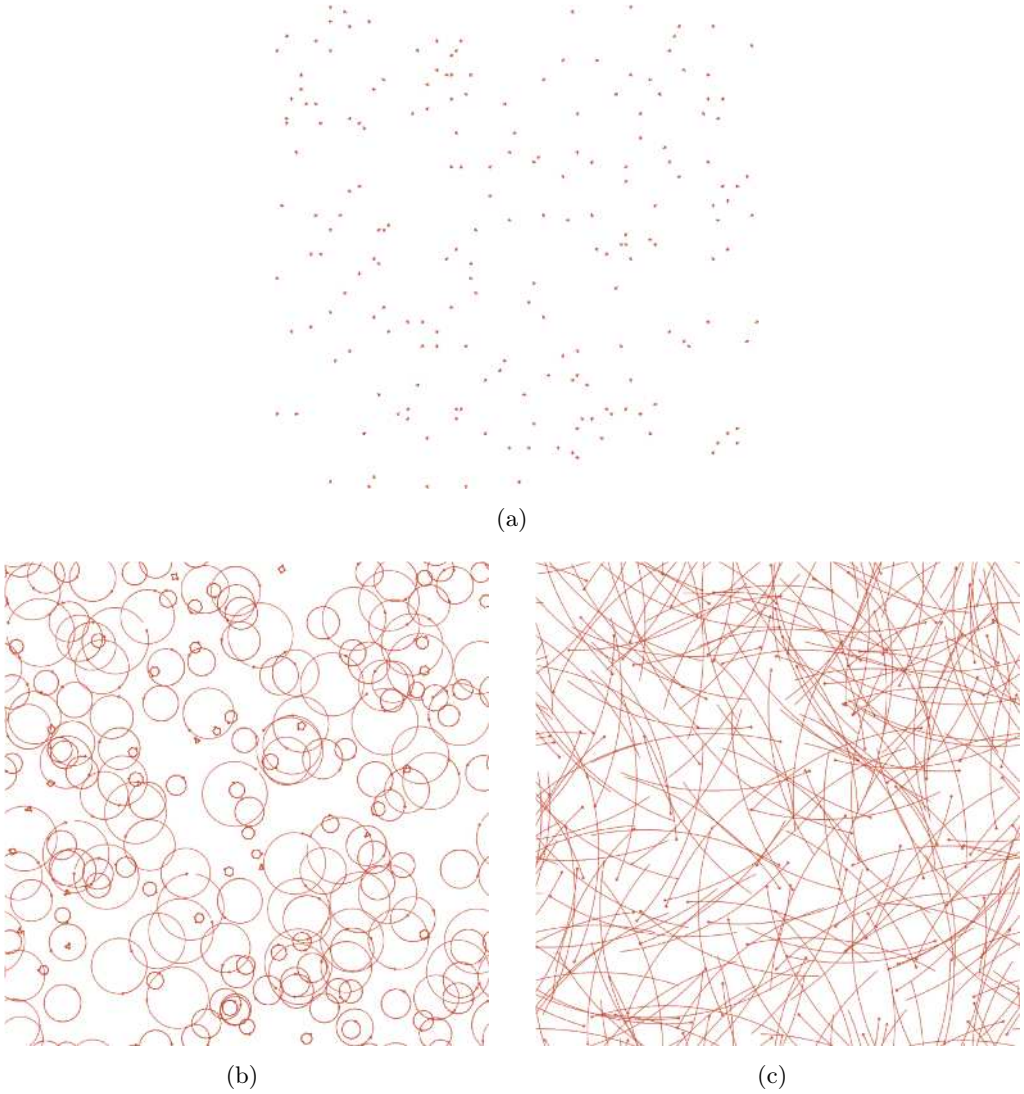
Pour ce faire, nous utilisons le modèle classique de diffusion épidémique *SIR*, que nous étendons au cas de la mobilité de la façon suivante :
Soit $G = \langle G_{t_0}, G_{t_1}, ..., G_{t_m} \rangle$ la séquence de réseaux de contacts de proximité instantanés dans laquelle chaque $G_{t_j} = (V, E_{t_j})$ représente le réseau de contacts entre agents à l'instant t_j avec $t_j \in T$ et $\forall j \in [0..m]$, $t_j < t_{j+1}$, comme défini dans la Section 4.3.
V est l'ensemble des agents et E_{t_j} est un sous-ensemble de $V \times V$ de paires d'agents connectés du fait de leur proximité dans l'espace à l'instant t_j.
Nous notons $F_{t_j} : V \to \{S, I, R\}$ la fonction qui renvoie l'état de l'agent $v \in V$ à l'instant t_j. Ainsi, à chaque instant $t_j \in T$, nous définissons $I_{t_j} = \{v \in V \ ; \ F_{t_j}(v) = I\}$, comme l'ensemble des agents infectés du réseau et $N^v_{t_j} = \{v' \in V \ ; \ (v, v') \in E_{tj} \ \ et \ \ F_{t_j}(v') = I\}$, comme l'ensemble des voisins infectés de l'agent v.
Soit α la probabilité de transmission par contact, la maladie se propage à un agent susceptible v à l'instant t_j avec la probabilité $1 - (1 - \alpha)^{\#N^v_{t_j}}$.
Chaque agent infecté a une probabilité β de guérir. Une fois qu'un agent est dans l'état R, il le reste jusqu'à la fin de la simulation et ne peut plus transmettre la maladie.

Ainsi, du point de vue de la diffusion, deux types de paramètre doivent être considérés : (i) les deux paramètres du modèle *SIR* (α et β) et (ii) les deux paramètres liés au modèle *ER* (fTL et δ).
L'objectif de notre étude est donc double. D'une part, nous cherchons à comprendre comment les paramètres des modèles (*SIR* et *ER*) influencent la propagation et d'autre part, comment le processus de diffusion se comporte selon la mobilité. Dans cette étude, nous distinguons trois catégories d'agents, définis selon leur facteur de mobilité :

- ***Imm.*** sont des agents immobiles, obtenus en fixant $fTL = 0$. Ces agents restent à la même position durant toute la durée de la simulation (voir Figure 4.13(a)).
- ***Séd.*** sont des agents sédentaires, obtenus en fixant leur fTL dans l'intervalle $[3, 45]$, c.-à-d. $\mu_i \in [0, 0.125]$. Ces agents explorent uniquement une très petite portion de leur environnement (voir Figure 4.13(b)).
- ***Voy.*** sont des agents voyageurs, obtenus en fixant leur fTL dans l'intervalle $[315, 360]$, c.-à-d. $\mu_i \in [0.875, 1]$. Ce sont des agents qui explorent une très grande partie de l'espace (voir Figure 4.13(c)).

Comme précédemment, un seul agent est initialement infecté. Pour comparer l'effet des différents types de mobilité, chaque test est composé uniquement d'agents de la même catégorie (*Imm.*, *Séd.* ou *Voy.*) et les résultats ont été moyennés sur 100 exécutions.

Dans une première approche, nous commençons par nous intéresser aux courbes d'incidence, qui constituent souvent le premier indicateur du comportement global du processus au cours du temps. Sur la Figure 4.14, nous comparons les courbes d'incidence obtenues avec les différents types de mobilité pour $\delta = 0.7$ et dans deux configurations choisies arbitrairement : (a) $\alpha = 0.325$, $\beta = 0.2$ et (b) $\alpha = 0.775$, $\beta = 0.05$. La densité a été fixée à 70%



(a)

(b)

(c)

FIGURE 4.13 – Trajectoires selon les catégories d'agents avec $\delta = 0.02$ à $t = 40$ Agent (a) immobiles (*Imm.*), (b) sédentaires (*Séd.*), (c) voyageurs (*Voy.*)

dans cette première approche, car nous avons montré dans la section précédente que cette valeur garantissait la percolation quelle que soit la mobilité considérée.

Comme attendu, le modèle *SIR* fournit des courbes d'incidence en forme de *cloche*, sur lesquelles nous pouvons observer que le phénomène se propage jusqu'à atteindre un pic épidémique, avant de connaitre une phase de décroissance. Dans le chapitre précédent, nous avions déjà observé que trois caractéristiques de la diffusion pouvaient être considérées : (i) la valeur du pic VP, (ii) son temps d'apparition TP et (iii) l'aire sous la courbe d'incidence AUC, qui représente la couverture globale du phénomène.
Dans cette première approche, nous pouvons donc observer l'effet de la mobilité des agents sur ces trois caractéristiques. Comme attendu, toutes les formes de mobilité fournissent un pic épidémique dans ces deux configurations, puisque la densité utilisée permet la percolation. Toutefois, la virulence de la diffusion semble croître avec la mobilité des agents dans

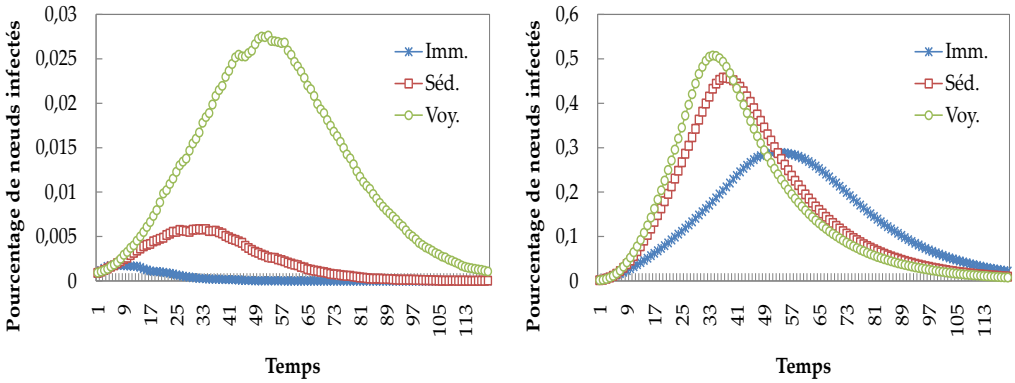


FIGURE 4.14 – Courbes d'incidence selon les différents types de mobilité avec $\delta = 0.7$ (a) $\alpha = 0.325$, $\beta = 0.2$ et (b) $\alpha = 0.775$, $\beta = 0.05$

les deux configurations. Typiquement, les agents voyageurs sont ceux qui fournissent le pic épidémique le plus élevé, suivi des sédentaires, puis des immobiles. En revanche, quand on compare le temps d'apparition du pic, nous observons que les paramètres α et β semblent perturber fortement les tendances. En effet, alors que dans la configuration (a) le pic survient d'abord pour les immobiles, ensuite pour les sédentaires et enfin pour les voyageurs, la tendance est complètement inversée dans la configuration (b). En ce qui concerne l'aire sous la courbe, il est facile d'observer les variations entre les configurations (a) et (b). Pour les agents voyageurs par exemple, nous observons que l'AUC tend à diminuer quand la mobilité est faible, laissant ainsi supposer que le phénomène s'étend moins dans le temps.

Pour mieux comprendre comment évolue le pic selon ces différents paramètres, nous commençons par étudier, pour chaque type de mobilité, l'évolution de la valeur du pic selon α et β. La Figure 4.15 présente ces résultats sous forme de schémas en $3D$ pour une densité de 70% (toujours pour garantir la percolation). La Figure 4.15(a) concerne les agents immobiles, (b) les sédentaires et (c) les voyageurs.

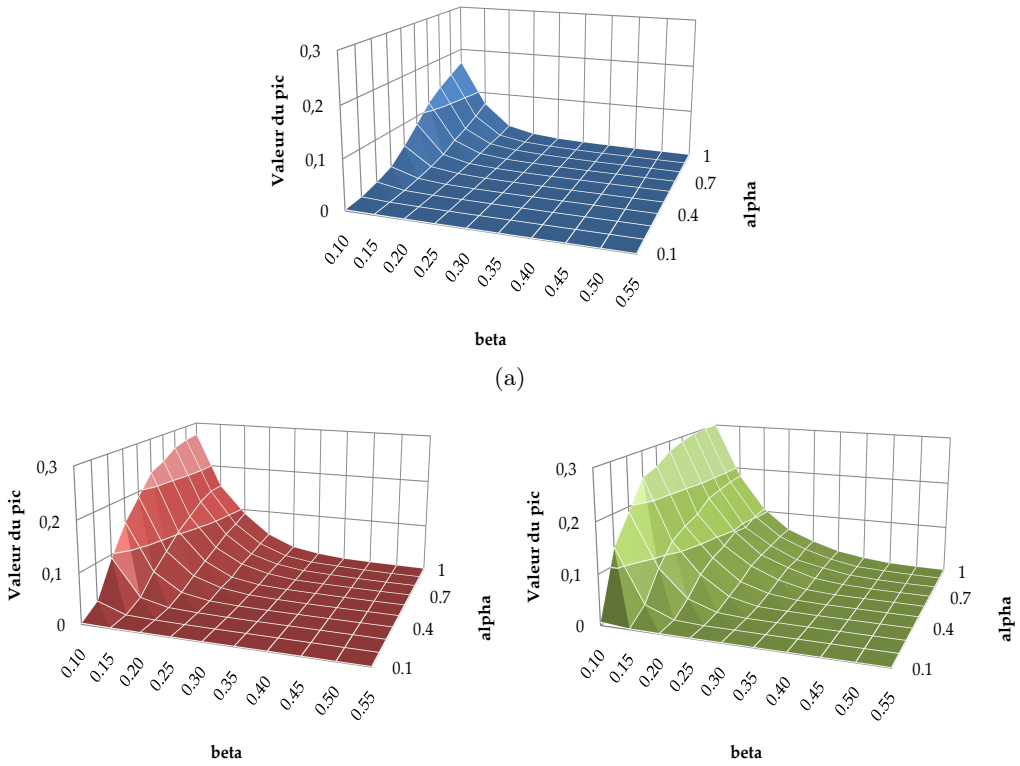
Sur cette figure, nous pouvons observer que quel que soit le type de mobilité, la valeur du pic croît quand α augmente et que β diminue. Par exemple, pour les agents voyageurs, $VP \approx 30\%$ pour $\alpha = 1$ et $\beta = 0.1$, alors que $VP \approx 10\%$ pour $\alpha = 0.7$ et $\beta = 0.2$.
Cependant, nous pouvons également observer les fortes différences obtenues entre les différents types de mobilité. Par exemple, pour $\alpha = 1$ et $\beta = 0.1$, $VP \approx 15\%$ pour les agents immobiles, contre 25% pour les sédentaires et 30% pour les voyageurs. Ces résultats confirment les observations faites précédemment (voir Figure 4.14) sur la valeur du pic : plus les agents sont en mouvement et plus la diffusion est virulente.

Ainsi, si nous considérons uniquement les valeurs de α et de β qui maximisent VP (dans nos tests $\alpha = 1$ et $\beta = 0.1$ (voir Figure 4.15)), nous pouvons comparer les effets de la mobilité sur (1) VP, (2) TP et (3) AUC. Sur la Figure 4.16, nous comparons l'évolution des ces caractéristiques pour une densité de 70%.

En ce qui concerne la valeur du pic, nous observons quels que soient les paramètres utilisés, le classement suivant :

$$VP_{Imm.} \leq VP_{Séd.} \leq VP_{Voy.} \tag{4.9}$$

En effet, pour toutes les variations de α et β, nous observons que le pic est toujours plus élevé pour les agents mobiles, puisque les valeurs de pic obtenues pour *Voy.* et *Séd.* sont



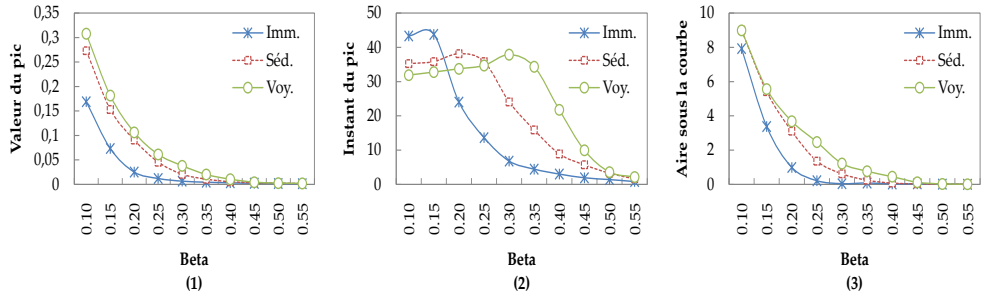
(a)

(b) (c)

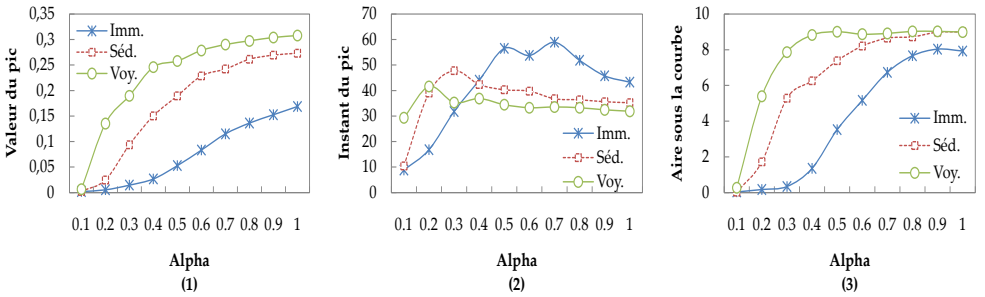
FIGURE 4.15 – Évolution de la valeur du pic selon α et β pour $\delta = 0.7$
Agents (a) immobiles, (b) sédentaires et (c) voyageurs

toujours supérieures à celles obtenues pour *Imm.*.
Cependant, nous observons également que les agents mobiles permettent systématiquement une plus large diffusion que les agents sédentaires. Ces résultats sont cohérents avec ceux obtenus dans la Section 4.3, qui montraient que le nombre de contacts moyen dépendait de la mobilité des agents. Nous avions en effet montré que plus les agents sont mobiles et plus ils sont susceptibles de favoriser la propagation entre eux, le nombre de contacts étant plus élevé.
Dans cette première étude, nous avions considéré que les agents avaient la même mobilité. Or dans l'étude que nous menons sur l'impact de la mobilité sur le processus de diffusion, les agents sont catégorisés selon qu'ils soient *immobiles*, *sédentaires* ou *voyageurs*. Ainsi, pour expliquer les résultats obtenus pour VP, il est utile de comprendre comment évolue le nombre de contacts selon les classes d'agents.
Sur la Figure 4.17, nous montrons la distribution des contacts selon le type de mobilité. Les contacts à considérer sont évidemment les contacts agrégés sur toute la période de simulation, c.-à-d. le degré dans le $mSPN_T$ (voir Section 4.3).

Ainsi, nous observons que le nombre de contacts de proximité (min, max, et moyen) croît significativement avec la mobilité des agents. Typiquement, sur la Figure 4.17(a), nous pouvons observer que les agents immobiles ont un nombre de contacts dans l'intervalle $[1, 5]$. De même, les agents sédentaires (resp. voyageurs) ont un nombre de contacts dans l'intervalle $[26, 226]$ (resp. $[376, 476]$) comme le montre la Figure 4.17(b).

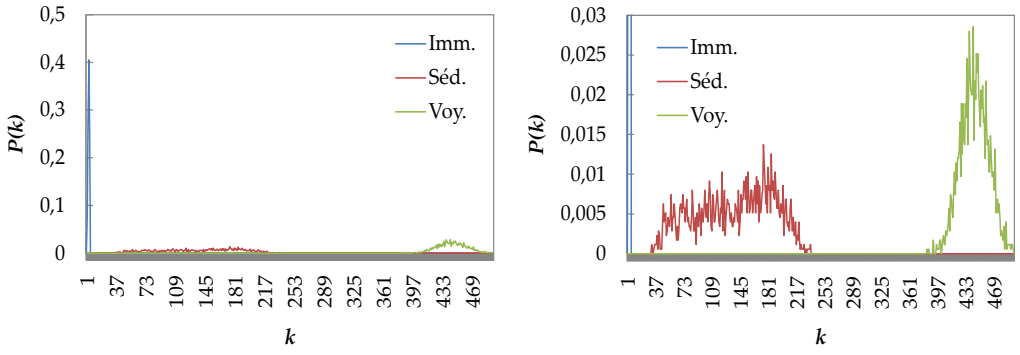


(a)



(b)

FIGURE 4.16 – Caractéristiques de la diffusion selon le type de mobilité avec $\delta = 0.7$ (1) VP, (2) TP et (3) AUC quand (a) $\alpha = 1$ et (b) $\beta = 0.1$



(a) (b)

FIGURE 4.17 – Distribution du degré du $mSPN_T$ selon le type de mobilité (a) pour tous les types de mobilité et (b) agrandissement sur *Séd.* et *Voy.*

Cela confirme les tendances observées pour la valeur du pic. Quand la mobilité des agents augmente, le nombre de contacts augmente également, créant ainsi plus d'opportunités de propagation pour le phénomène, ce qui se traduit par un accroissement du pic de diffusion comme observé sur les Figures 4.16(1).

En ce qui concerne le temps d'apparition du pic, il semble n'y avoir aucun ordre d'apparition selon la mobilité. Par exemple, pour $\alpha = 1$ et $\beta \in [0.1..0.15]$, le pic obtenu par les agents voyageurs survient le premier, suivi de celui obtenu par les agents sédentaires, puis celui des agents immobiles. En revanche, pour $\alpha = 1$ et $\beta \in [0.3..0.5]$ (voir Figure 4.16(a)(2)), nous

observons que le pic obtenu avec les agents immobiles survient le premier dans le temps, alors que ceux obtenus par les agents sédentaires et voyageurs arrivent respectivement en second et en troisième. Ce phénomène avait déjà pu être observé sur les Figure 4.14(a) et (b), et peut également être observé quand β est fixé (voir Figure 4.16(b)(2)).
Nous observons que ce phénomène d'inversion est lié aux paramètres α et β, et donc d'une certaine façon à la valeur du pic. Effet, quand les paramètres α et β favorisent la diffusion (c.-à-d. α élevé et β faible), l'ordre d'apparition du pic est le suivant : $Voy-Séd-Imm.$ En revanche, quand la diffusion est perturbée (α faible et β élevé), et donc que le pic est relativement faible, nous observons l'ordre inverse : $Imm-Séd-Voy.$ Ceci témoigne d'un phénomène de diffusion qui affecte très vite un nombre maximum, mais relativement faible, d'individus avant de connaître une phase de régression.
Ces résultats montrent ainsi que, bien que les pics obtenus avec les catégories de mobilité *Imm.* et *Séd.* soient les moins élevés, ils générèrent, sous certaines conditions, des configurations qui permettent au phénomène de diffusion d'atteindre très rapidement un nombre maximum d'individus infectés.

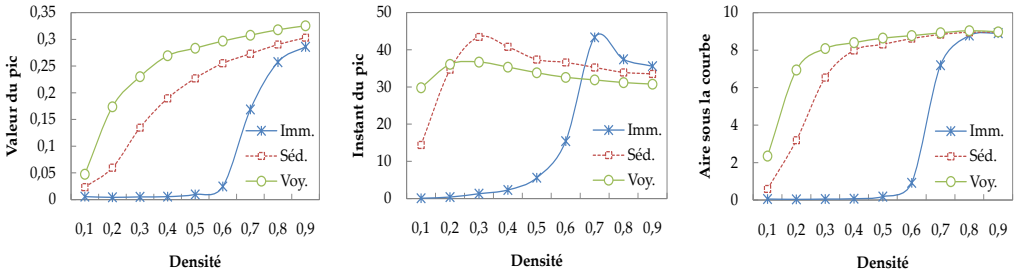
Des observations intéressantes peuvent être faites concernant l'évolution de l'aire sous la courbe (voir Figures 4.16(3)). En effet, comme pour la valeur du pic, quelles que soient les variations de α et de β, nous observons le classement suivant :

$$AUC_{Imm.} \leq AUC_{Séd.} \leq AUC_{Voy.} \tag{4.10}$$

De plus, pour des valeurs élevées de α, l'aire semble atteindre un état de stabilité. Par exemple, pour $\beta = 0.1$ et $\alpha \in [0.4..1]$, l'AUC obtenu avec les agents voyageurs reste relativement stable dans l'intervalle $[8.4, 8.9]$. De même, avec $\beta = 0.1$, nous observons une stabilité pour les agents sédentaires quand $\alpha \in [0.6..1]$ et pour les agents immobiles quand $\alpha \in [0.8..1]$.
Toutefois, il est important de préciser que la stabilité observée sur l'AUC n'affecte pas la croissance du pic (voir Figures 4.16(1)). Cela est caractéristique d'une courbe d'incidence qui croît en se resserrant sur elle-même. Le phénomène de diffusion est donc plus virulent, mais s'étend beaucoup moins dans le temps.

Les résultats présentés dans cette étude ont été obtenus avec une densité constante $density = 70\%$. Pour compléter cette étude, nous analysons l'évolution de ces trois caractéristiques (VP, TP et AUC) quand la densité d'agents évolue. Pour cela, nous considérons les paramètres qui maximisent le pic épidémique dans nos expériences ($\alpha = 1$ et $\beta = 0.1$) et nous comparons les variations selon la densité. Nous vérifions ainsi que les tendances observées soient bien confirmées pour toutes les densités.
La Figure 4.18 présente selon la densité, l'évolution de (a) la valeur du pic, (b) son temps d'apparition et (c) l'aire sous la courbe.

Ces résultats confirment ceux obtenus dans l'étude menée sur les seuils de percolation (voir Section 4.4.1, et montrent que quand les agents sont mobiles, la propagation est possible avec des valeurs de densité relativement faibles (c.-à-d. $\delta \geq 10\%$). En revanche, dans le cas d'agents immobiles, nous observons que la diffusion n'est possible qu'à partir d'un seuil $\delta \geq 70\%$. Ce résultat est également cohérent avec l'étude menée sur les seuils de percolation, puisque c'est précisément le seuil qui avait été identifié pour les agents immobiles (voir Figure 4.11).
Finalement, que quelle que soit la densité, si la percolation dans le système est garantie, ces résultats synthétiques démontrent l'impact de la mobilité, et donc de la dynamique du réseau, sur le processus de diffusion.



(a) (b) (c)

FIGURE 4.18 – Caractéristiques de la diffusion selon la densité ($\alpha = 1$ et $\beta = 0.1$) (a) Valeur du pic, (b) temps d'apparition et (c) aire sous la courbe

4.5 L'outil de simulation *ER-Net*

Le modèle *ER* a été implémenté avec *NetLogo* au sein d'un outil graphique que nous appelons *ER-Net* (***ER*** based on ***Net****Logo*). Notre objectif avec *ER-Net* est de fournir un outil de simulation simple et efficace pour étudier l'évolution des processus de diffusion sur des agents en mouvement.

Bien que notre outil ait été limité au modèle de diffusion épidémique *SIR*, nous rappelons que ce modèle est suffisamment générique pour être appliqué à d'autres formes de diffusion (rumeurs, informations, etc.) qui peuvent également être véhiculées à travers les contacts de proximité émergeant de la mobilité des individus. Comme le montre la Figure 4.19, l'interface de *ER-Net* se compose de trois panneaux principaux.

1. Le panneau de configuration situé à gauche de la fenêtre, permet à l'utilisateur de calibrer l'application avant le lancement de la simulation, et d'interagir avec l'affichage durant la simulation. Typiquement, on commence par y définir les dimensions $L_1 \times L_2$ de la zone de simulation (bouton "*Settings*" de la barre d'outils) et la densité d'agents sur cette zone (curseur de défilement "*density*"). Le nombre d'agents est ainsi automatiquement calculé et affiché dans la boite "*nb-agents*". Une fois les agents et la zone paramétrés, une classe peut être affectée à l'ensemble des agents présents sur la zone : (i) Immobiles, (ii) Sédentaires, (iii) Voyageurs, (vi) Mixtes ou (v) Fixes.
(i) Une population immobile fait référence à des agents qui restent à une même place durant toute la durée de la simulation. Le réseau de proximité est donc statique et est basé uniquement sur la proximité lors de la phase d'initialisation des agents. Ce type d'agents est obtenu en activant le bouton "*agents-immobiles*".
(ii) Une population sédentaire correspond à des agents qui parcourent de courtes distances. Leur fTL est automatiquement défini entre 3 et 45.
(iii) Une population de voyageurs représente des agents qui parcourent, eux, de grandes distances. Leur fTL est défini entre 315 et 360.
(vi) Une population mixte est une population qui regroupe tous types d'agents en mouvement. On y trouve à la fois des sédentaires et des voyageurs. Leur fTL est uniformément choisi entre 3 et 360.
(v) Une population fixe, est une population qui ne contient que des agents ayant la même mobilité, c'est-à-dire que le fTL est le même pour tous les agents. Pour obtenir ce type de population, le curseur de défilement "fTl" doit être positionné sur la valeur souhaitée.
Une fois la population définie, le seuil de distance minimal $pDistance$, à partir duquel le contact est établi entre deux agents peut être paramétré. Nous rappelons quand dans nos

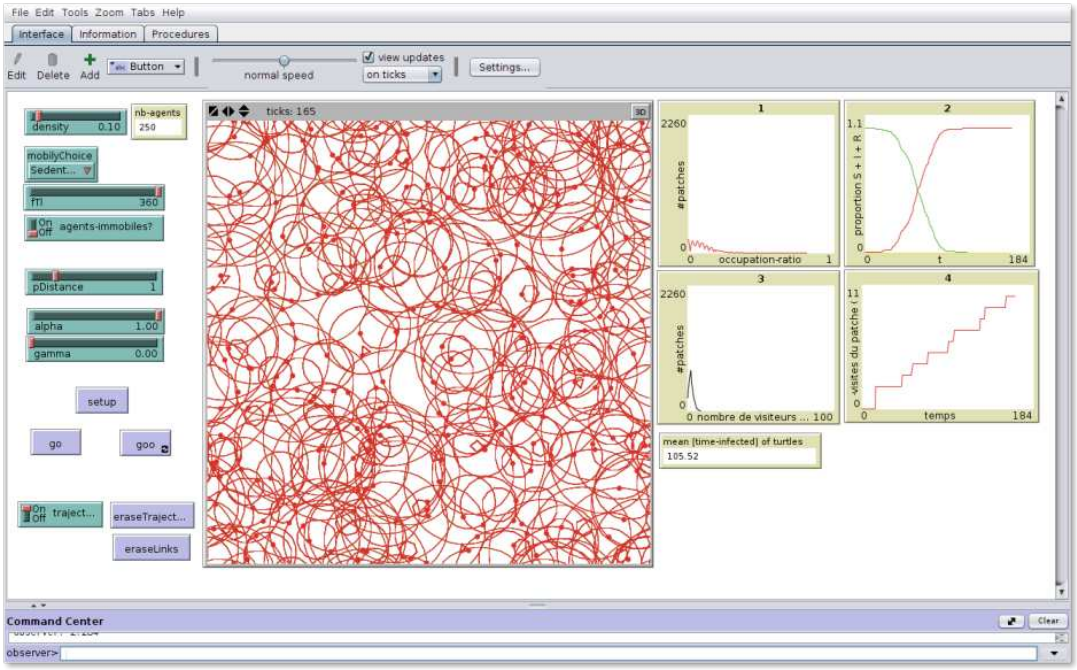
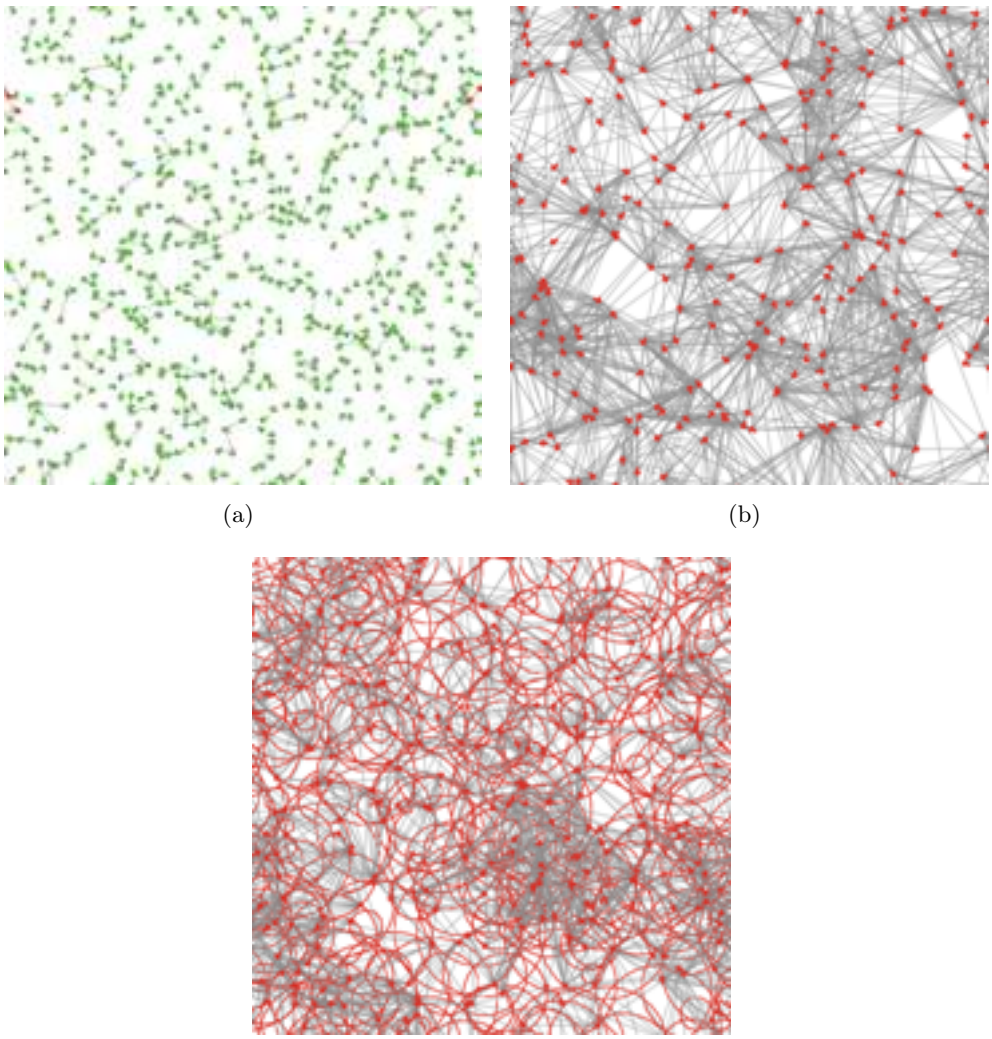


FIGURE 4.19 – Capture de l'interface de *ER-Net*

tests, ce seuil était fixé à 1.
Enfin, les paramètres du modèle SIR peuvent être spécifiés : la probabilité de transmission par contact α et la probabilité de guérison β.
Une fois l'outil entièrement paramétré, le bouton "*setup*" permet d'initialiser et d'afficher les agents sur le panneau central. Le bouton "*go*" fait la simulation avancer uniquement d'une unité de temps, alors que le bouton "*goo*" lance la simulation sur une période indéfinie. La simulation s'arrête lorsque l'utilisateur clique de nouveau sur le bouton "*goo*".

2. Le panneau de visualisation placé au centre de la fenêtre, permet à l'utilisateur de suivre le déroulement de la simulation. On peut y voir les agents, leur état (vert pour *susceptible*, rouge pour *infecté* et jaune pour *guéri*), leurs trajectoires et le réseau de contacts proximité. La Figure 4.20 montre, par exemple, quelques vues obtenues avec *ER-Net*.
Le panneau de configuration permet d'interagir avec l'affichage. Typiquement, les trajectoires peuvent être masquées ou effacées de la vue (voir Figure 4.20(a)). Par défaut, le réseau de contacts de proximité est masqué, mais il peut être affiché en utilisant le panneau de configuration (voir Figures 4.20).

3. Le panneau de résultats permet de suivre en temps réel l'évolution de certains indicateurs calculés au cours de la simulation. En effet, le langage *NetLogo* permet de maintenir, à chaque itération, des courbes ou des mesures, effectuées à partir d'un certain nombre de critères sur l'environnement.
À chaque itération, nous nous intéressons (i) au taux d'occupation moyen par cellule, (ii) au pourcentage d'individus susceptibles, infectés et guéris, et au temps d'infection moyen, (iii) au nombre de visiteurs par cellule et (iv) au nombre de visites sur une cellule donnée au cours du temps. Toutes ces informations peuvent ensuite être sauvegardées dans des fichiers textes pour des analyses plus approfondies.



(a) (b)

(c)

FIGURE 4.20 – Exemple de visualisation obtenus avec *ER-Net*
(a) Contacts de proximité instantanés, (b) $mSPN_T$ et (c) trajectoires et $mSPN_T$

Enfin, à chaque simulation, un fichier texte contenant les contacts de proximité survenant dans le réseau à chaque itération est automatiquement généré. Ce fichier à la forme :

$$<Agent\ i> <Agent\ j> <itération\ k>$$

Nous rappelons que c'est l'ensemble de ces liens qui constitue le $mSPN_T$.
Ce fichier peut ainsi être lu par l'outil *DynSpread* présenté dans le chapitre précédent à la Section 3.5, pour simuler des dynamiques de réseaux issus de comportements plus réalistes. Il peut également être réécrit au format UCINET [Borgatti 2002b] en utilisant *DynSpread*, de façon à être lu par d'autres outils d'analyse de réseau tels que UCINET [Borgatti 2002b], NetDraw [Borgatti 2002a] ou Gephi [Bastian 2009].

A long terme, nous espérons que *ER-Net* pourra être utilisé pour intégrer toutes les évolutions qui seront apportées au modèle *ER*. Une évolution simple serait par exemple de

complexifier le modèle, en permettant aux agents de revenir à leur point de départ sans nécessairement suivre des trajectoires circulaires.

Plus généralement, la mobilité des individus est aujourd'hui identifiée comme étant l'un des principaux facteurs favorisant la propagation à grande échelle de maladies infectieuses [(WHO) 2007]. Un tel outil pourrait ainsi permettre de mieux comprendre les comportements liés à la mobilité des êtres humains, et leurs effets à la fois sur le réseau de contacts sous-jacent, mais également sur les processus de diffusion. À terme, un tel outil pourrait permettre l'élaboration de stratégies d'intervention adaptées tenant compte de cette mobilité, et plus généralement de la dynamique du réseau.

4.6 Conclusion

Dans ce chapitre, nous avons pu observer que différents facteurs pouvaient être à l'origine de modifications au sein de la structure du réseau social. L'un de ces facteurs concerne la mobilité géographique des agents, qui est une dimension transversale à toute pratique sociale. La mobilité géographique est en effet à l'origine de contacts de proximité dynamiques ayant la capacité de permettre la diffusion de divers types de phénomènes tels qu'une maladie, une rumeur, ou plus généralement une information.

Dans ce travail, nous avons abordé le problème de la diffusion dans des réseaux sociaux dynamiques, en nous intéressant au cas spécifique où la dynamique du réseau est induite par la mobilité des individus. Ce travail s'inscrit donc dans l'axe de la modélisation pour la compréhension des phénomènes sociaux. Les contributions de ce chapitre peuvent être résumées comme suit :

(i) Nous avons proposé le modèle de mobilité humaine ***Eternal-Return*** (***ER***), un système multi-agents qui modélise le comportement d'agents qui se déplacent dans un espace géographique et explorent des zones plus ou moins étendues de leur environnement. Contrairement aux modèles traditionnels, basés sur un ensemble de règles stochastiques, le modèle que nous proposons s'inspire des découvertes faites très récemment à partir de traces réelles et intègre deux aspects fondamentaux des déplacements humains : la périodicité et l'hétérogénéité. Nous avons ensuite montré comment, à travers le modèle, la mobilité des agents induisait un réseau social basé sur des contacts de proximité dynamiques. Enfin, plusieurs études ont été menées pour montrer que le modèle est en mesure de reproduire certains motifs et propriétés observés dans la réalité.

(ii) Le modèle *ER* a ensuite été utilisé pour comprendre **l'impact de la dynamique du réseau sur le processus de propagation**, lorsque la dynamique est induite par la mobilité des individus. Ces résultats ont ainsi pu mettre en lumière l'effet du facteur mobilité sur le processus. Du point de vue de la percolation, nous avons pu observer que la connexité de la structure n'est garantie qu'à partir d'un certain seuil de densité qui varie selon la mobilité des individus. Les résultats obtenus ont ainsi mis en évidence cette relation en montrant notamment comment de faibles niveaux de mobilité suffisent à réduire considérablement le seuil de percolation. Du point de vue de la diffusion, nous avons pu observer que lorsque la percolation est assurée, la taille de l'épidémie tend à croître avec le niveau de mobilité alors que son apparition est plus précoce.

(iii) Le modèle *ER* a été implémenté au sein de **l'outil graphique *ER-Net***, qui vise à étudier comment un processus de diffusion prend place et évolue lorsque des agents sont en mouvement. Cet outil présente trois intérêts majeurs pour le domaine de la recherche.
Il peut être utilisé pour générer des traces sur le déplacement des individus. Ces traces peuvent ensuite être utilisées pour valider ou adapter des protocoles mis en place pour les réseaux ad hoc.

Il permet d'obtenir des réseaux de contacts dynamiques, basés sur la proximité géographique. Cette information sur l'évolution des liens est particulièrement utile pour les travaux menés sur la dynamique des réseaux, tels que l'extraction de communautés dans des réseaux en évolution ou la prédiction de liens.
L'outil permet de simuler et d'étudier l'évolution de processus de diffusion lorsque les agents sont en mouvement.
Pour les professionnels de santé par exemple, connaissant la densité et le niveau de mobilité des individus, l'impact de la diffusion d'une maladie infectieuse dans une zone géographique définie peut être évalué.

Ainsi d'une façon générale, ce chapitre contribue à une meilleure compréhension des phénomènes de diffusion prenant place dans la réalité. Typiquement, la relation mise en évidence entre le seuil de percolation et la mobilité des individus permet dans des cas réels, de déduire la mobilité minimale obligatoire pour qu'une diffusion soit possible quand la densité est fixée, ou inversement, la densité minimale obligatoire quand la mobilité est fixée.
Dans le domaine du marketing par exemple, ces résultats peuvent être utilisés pour déterminer les seuils de mobilité ou de densité à garantir pour maximiser la diffusion d'informations sur un produit ou un événement dans une zone géographie ciblée.

Chapitre 5

Fouille de données sociales : vers une analyse conceptuelle des réseaux sociaux

Sommaire

Comme nous l'avons présenté dans le Chapitre 2, les méthodes d'extraction de connaissances ont été appliquées à l'analyse des réseaux sociaux [Getoor 2005, Blondel 2008, Zhou 2009]. Ces méthodes, qui mettent en oeuvre des techniques standards de fouille de données (classification, prédiction, clustering, recherche de motifs fréquents), ne considèrent généralement que la structure topologique du réseau, en ignorant les attributs des noeuds. C'est le cas par exemple des méthodes d'extraction de motifs fréquents. La plupart des approches exploitent uniquement la structure topologique, pour extraire des sous-réseaux qui se retrouvent fréquemment dans un ensemble de réseaux, ou dans un unique réseau beaucoup plus large.

Pourtant, considérer ces deux types d'informations semble être une question fondamentale si l'on veut pouvoir tirer pleinement parti de toutes les informations disponibles sur les réseaux. En effet, alors que les liens décrivent naturellement la nature des relations entre les

individus, les attributs sont, eux, souvent porteurs d'informations pertinentes sur le rôle, la position ou l'influence d'un noeud dans le réseau. Dans le cas de la diffusion de maladies par exemple, bien que le lien social entre deux individus soit le principal vecteur de transmission, on peut supposer que des propriétés, telles que l'âge ou les origines, peuvent influencer de façon assez significative la probabilité qu'aura un individu de contracter la maladie, et donc également le processus global de diffusion.

Devant la multiplication des jeux de données disponibles, dans lesquels, en plus des liens, on dispose d'informations sur les noeuds, de nouvelles méthodes se sont intéressées à la dimension portée par les attributs des noeuds [Zhou 2009, Riadh 2009]. De nouveaux algorithmes, ou adaptations d'algorithmes existants, ont en effet été proposés pour considérer les attributs comme une source d'information complémentaire dans l'étude de problèmes variés, tels que la visualisation [Snasel 2008], la classification [Kuznetsov 2009] ou l'identification de communautés [Gaume 2010].

Pourtant, bien que des efforts importants aient été réalisés pour proposer des méthodes d'extraction de connaissance hybrides, capables de prendre en compte à la fois des informations sur la structure du réseau et les attributs des noeuds, des questions importantes et inédites se posent et sont encore peu étudiées :

1. Quels sont les groupes de noeuds les plus connectés du réseau ?
2. Quelles sont les caractéristiques les plus fréquemment retrouvées en connexion sur l'ensemble du réseau ?
3. Si un nouveau lien est créé, quels types d'agents sont susceptibles d'être connectés ?

Dans ce chapitre, notre objectif est d'aborder le problème de la recherche de motifs fréquents dans les réseaux sociaux sous un angle nouveau. En effet, contrairement aux méthodes classiques, qui s'intéressent principalement à la structure du réseau à travers la recherche de sous-graphes fréquents, nous proposons ici une vision d'ensemble, qui considère non pas des sous-structures, mais des groupes de noeuds.
Toutefois, les groupes ciblés ici ne sont pas les groupes traditionnellement extraits à partir de méthodes de clustering, à savoir des ensembles de noeuds fortement connectés (autrement dit des communautés), mais plutôt des groupes partageant des attributs communs, un type de motifs similaire à ceux recherchés par l'analyse de concepts formels.

Dans ce travail, nous proposons donc une approche hybride, qui combine à la fois les techniques d'analyse de concepts formels dans les réseaux et les notions couramment utilisées en recherche de motifs fréquents. Plus précisément, notre approche consiste : (i) à utiliser la similarité des attributs comme le critère le plus important pour le regroupement des noeuds au sein de *clusters conceptuels*, puis (ii) à conserver les liens du réseau les plus fréquents entre ces clusters. Nous appelons ces motifs des *liens conceptuels*.
L'originalité de notre approche réside dans cette capacité à exploiter ces deux types d'information : **les attributs** pour la génération des groupes et **la structure** pour l'extraction des liens conceptuels.

L'intérêt d'une telle approche est double. D'une part, les motifs extraits fournissent une connaissance pertinente sur les types de noeuds les plus connectés du réseau. D'autre part, nous montrons que l'approche proposée permet de représenter le réseau de façon plus sémantique, en synthétisant la connaissance acquise dans une structure de graphe que nous appelons *vue conceptuelle*.
Comme l'illustre la Figure 5.1, ce travail s'inscrit dans l'axe de l'exploitation des données, et aborde plus particulièrement le problème de l'extraction de modèles à partir de données réelles ou simulées. La vue conceptuelle est en effet un modèle capable de décrire les *liens forts* au sein d'un réseau social.

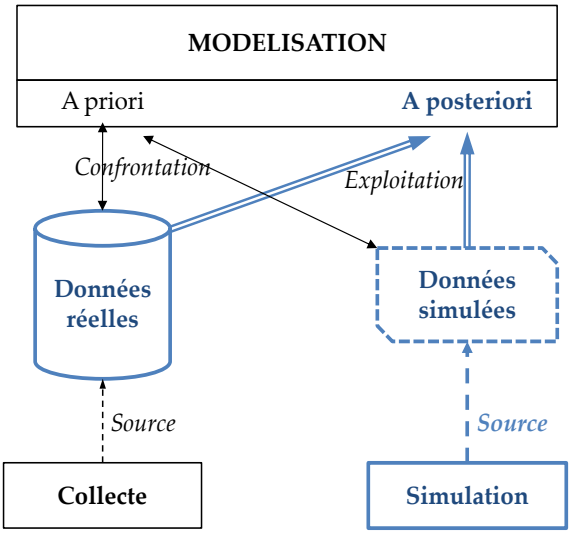


FIGURE 5.1 – Des données sociales aux modèles : extraction de modèles à partir de données réelles ou simulées

Ainsi, pour répondre aux questions posées, nous commençons par définir formellement la notion de liens conceptuels. Nous discutons ensuite des difficultés liées à la recherche des liens conceptuels, et présentons l'algorithme d'extraction *MFCL-Min*, qui vise à extraire uniquement les *liens conceptuels fréquents maximaux*. Enfin, la pertinence de la méthode, aussi bien d'un point de vue qualitatif que d'un point de vue quantitatif, est démontrée à travers diverses expériences qui montrent les bonnes performances de la solution dans plusieurs configurations.

Ce chapitre est organisé comme suit. La Section 5.1 passe en revue les principaux travaux liés à la recherche de motifs dans les réseaux sociaux. Dans la Section 5.2, nous présentons formellement les notions de *lien et de vue conceptuels*. La Section 5.3 est consacrée à *MFCL-Min*, l'algorithme d'extraction. La Section 5.4 détaille la génération de vues conceptuelles à partir des motifs extraits. Dans la Section 5.5 nous démontrons l'efficacité de la solution proposée à travers des résultats expérimentaux. Enfin, la Section 5.6 présente l'outil graphique qui implémente cette solution et la Section 5.7 conclut ce chapitre.

5.1 Extraction de motifs dans les données sociales

Les méthodes récentes d'analyse des réseaux sociaux appliquent généralement les concepts de fouille de données traditionnels aux réseaux. Ces approches, qui font partie du domaine assez nouveau de la *fouille de réseaux sociaux* (ou également *fouille de liens*), font référence selon Getoor [Getoor 2005] à toutes les "*techniques de fouille de données qui considèrent explicitement les liens lors de la construction de modèles descriptifs ou prédictifs à partir de données de type réseau*". Ainsi, comme en fouille de données traditionnelle, la fouille de réseaux sociaux recouvre plusieurs catégories de tâches telles que la détection de groupes, la classification, la prédiction de liens, etc.

Le problème que nous abordons dans ce travail partage des concepts avec la recherche de motifs fréquents dans les réseaux, le clustering basé sur les liens et également l'analyse conceptuelle des réseaux sociaux. Cette section présente ces différentes tâches et montre

en quoi les motifs qui nous intéresse diffèrent de ceux traditionnellement extraits par les méthodes standards.

5.1.1 Motifs fréquents

Dans le domaine de l'analyse des réseaux, les méthodes recherchant des sous-graphes sont certainement la famille de méthodes se rapprochant le plus d'une tâche classique d'extraction de motifs fréquents. En effet, la définition la plus largement répandue d'un "*motif*" dans le contexte des réseaux, est celle du "*sous-graphe*" [Cheng 2010]. Ainsi, le problème de l'extraction de motifs fréquents dans les réseaux sociaux consiste généralement à rechercher des sous-réseaux fréquemment retrouvés dans un ensemble de réseaux [Inokuchi 2000] ou dans un unique et large réseau [Kuramochi 2005] en fonction d'un seuil minimum de support.

L'approche traditionnelle consiste à associer des étiquettes aux noeuds et aux liens. En utilisant une telle représentation du réseau, le problème est réduit à la recherche d'ensembles d'étiquettes retrouvées suffisamment fréquemment.
Un exemple classique est celui des réseaux qui peuvent être obtenus à partir de paniers de produits. Dans de tels réseaux, les noeuds correspondent aux produits et deux produits sont liés s'ils appartiennent au même panier. L'ensemble des produits d'un même panier forme ainsi un graphe complet. Une fois que ce réseau a été généré pour chaque panier, les sous-graphes retrouvés fréquemment dans l'ensemble des réseaux représentent un motif fréquent au sens traditionnel.

Les principaux algorithmes d'extraction de sous-graphes fréquents peuvent être classifiés en deux familles [Cheng 2010] :
(i) Les approches basées sur Apriori font référence aux techniques qui exploitent les principes de l'algorithme *Apriori* [Agrawal 1994] pour retrouver les sous-réseaux fréquents. Le processus d'extraction se déroule souvent en deux étapes. (1) Une étape de génération des sous-graphes candidats, qui se base sur les sous-graphes fréquents de taille n pour générer les candidats de taille $n+1$. (2) Une phase d'évaluation qui évalue la fréquence de chaque candidat en utilisant la propriété d'isomorphisme dans les graphes.
Les algorithmes les plus connus de cette catégorie sont par exemple l'algorithme *AGM* proposé par Inokuchi *et al.* [Inokuchi 2000] pour minimiser à la fois le stockage et le temps de calcul, ou l'algorithme *FSG* introduit par Kuramochi et Karypis [Kuramochi 2001] conçu pour être robuste à la taille des réseaux.
(ii) Les approches basées sur la croissance de motifs regroupent des méthodes qui recherchent des sous-réseaux fréquents de taille $n+1$, en étendant des sous-réseaux fréquents de taille n par l'ajout de liens dans toutes les directions possibles [Nijssen 2005]. La principale difficulté de ce type d'approche est que des structures identiques peuvent être générées à différentes itérations. Ce problème est par exemple étudié par Yan et Han [Yan 2002], qui proposent l'algorithme *gSpan* pour limiter l'espace de recherche.

5.1.2 Clustering basé sur les liens

Le clustering basé sur les liens (Link-Based Clustering), plus connu comme étant le problème de "*détection de communautés*", fait référence à une catégorie de méthodes qui effectuent des regroupements de noeuds en s'appuyant uniquement sur les connexions entre les objets. Ces méthodes ont pour objectif de partitionner le réseau en plusieurs composantes, au sein desquelles les noeuds sont fortement connectés [Steinhaeuser 2010, Yoon 2011].

Plus formellement, les algorithmes tentent d'identifier les groupes qui maximisent le nombre de liens intra-communautaires et minimisent le nombre de liens inter-

communautaires. Le principal défi de ces approches réside dans la définition de distances adaptées car les mesures de distance traditionnellement utilisées en fouille de données ne peuvent pas être appliquées directement. Les mesures de similarité utilisées se basent souvent sur des indicateurs tels que le degré ou le coefficient de clustering pour calculer la similarité entre des couples de noeuds, ou des groupes, afin d'identifier les communautés.

Plusieurs classifications de ces méthodes ont été proposées [Fortunato 2009]. On retrouve comme en data mining classique les deux types d'algorithmes standards. La plus courante consiste à distinguer deux grandes familles de méthodes. **(i) Les algorithmes agrégatifs**, dans lesquels les groupes sont fusionnés à chaque itération si leur score de similarité est suffisamment élevé. **(ii) Les algorithmes séparatifs**, dans lesquels les groupes sont éclatés à chaque itération en supprimant les liens entre les noeuds possédant de faible niveau de similarité.

Des travaux très récents se sont intéressés à l'exploitation des attributs des noeuds lors de la recherche et la construction des communautés [Wang 2003]. Ces techniques peuvent être vues comme des approches hybrides à la frontière entre les méthodes de clustering basées sur les liens, c'est-à-dire les méthodes de recherche de communautés, et les méthodes de clustering traditionnelles basées uniquement sur les attributs.
Comme le mentionnent Zhou *et al.* dans [Zhou 2009], la définition d'un "*groupe*" a été adaptée à ce nouveau type d'approches. En effet, plutôt que de rechercher uniquement des groupes de noeuds densément connectés, ces nouvelles approches considèrent un cluster comme "*un sous-graphe densément connecté, au sein duquel les noeuds possèdent des valeurs d'attributs homogènes*" [Zhou 2009]. Typiquement, ce nouveau type d'algorithmes vise à partitionner le réseau en recherchant un équilibre entre la similarité de la structure et la similarité des attributs, l'objectif étant que les noeuds possédant des attributs communs soient regroupés au sein d'une même partition [Tian 2008, Yoon 2011].

5.1.3 Analyse conceptuelle

Certains travaux très récents tentent d'appliquer les principes de l'analyse de concepts formels (FCA) aux réseaux sociaux. Ces travaux ont pour objectif de combiner les techniques d'analyse et de visualisation de façon à aborder le monde en termes d'objets et d'attributs [Le Grand 2009].

L'analyse de concepts formels est une approche mathématique d'analyse des données introduite en 1984 par Rudolf Wille [Wille 1984], qui vise à extraire de la connaissance de données structurées. Le principe de base consiste à regrouper des objets en classes selon les propriétés qu'ils partagent. Le couple "*(Objets,Propriétés partagées)*" est appelé *un concept* et l'ensemble des concepts forme un treillis de Galois, également appelé treillis de concepts [Liquiere 2006, Kuznetsov 2009].

Dans le domaine des réseaux sociaux, les travaux récents menés sur l'analyse conceptuelle ont considéré les noeuds du réseau comme les objets à regrouper et les noeuds avec lesquels ils sont connectés (leurs voisins) comme leurs propriétés [Snasel 2008, Riadh 2009]. L'analyse conceptuelle du réseau social revient alors à extraire du réseau les groupes de noeuds qui ont des contacts en commun. Il est important de préciser que les noeuds du réseau initial peuvent être identifiés dans plusieurs concepts puisque les clusters extraits peuvent se chevaucher. En ce sens, l'analyse conceptuelle des réseaux sociaux peut être vue comme une approche qui combine à la fois la recherche de motifs (l'ensemble des concepts) et le clustering (les noeuds au sein d'un concept). Évidemment, ce type de méthodes est utilisé pour aborder les problèmes de clustering dans les réseaux sociaux [Liquiere 2006], mais également les problèmes liés à la visualisation [Snasel 2008], la classification [Kuznetsov 2009] ou l'identification d'éléments particuliers [Riadh 2009].

Par exemple, Snasel *et al.* [Snasel 2009] s'intéressent au problème de la visualisation et proposent une solution qui réduit les liens entre les objets de façon à optimiser la génération du treillis de concepts. Dans [Gaume 2010], Gaume *et al.* proposent une nouvelle approche, qui combine des techniques de clustering conceptuel et d'identification de groupes, pour extraire des communautés dans les réseaux bipartis.
Un autre problème concerne la taille du treillis de concepts, qui croît exponentiellement avec la taille du réseau, ce qui peut rapidement le rendre ininterprétable. Pour éviter la génération et la manipulation du treillis, Le Grand *et al.* [Le Grand 2009] proposent par exemple plusieurs mesures pour caractériser un noeud. La solution peut ensuite être étendue à l'ensemble du jeu de données de façon à caractériser le réseau en termes d'homogénéité/hétérogénéité et identifier les éléments les plus significatifs.

5.1.4 Notion de liens conceptuels

L'approche que nous proposons dans ce travail est une approche hybride, qui combine à la fois des techniques d'analyse de concepts formels appliquées aux réseaux sociaux et des techniques de recherche de motifs fréquents.
Dans notre approche, les noeuds sont d'abord regroupés sur la base d'une similarité des attributs. Les liens fréquents entre ces groupes sont ensuite extraits. Ces motifs fournissent une information sur les groupes de noeuds, et donc les propriétés, les plus connectés du réseau.

Sur la Figure 5.2, nous comparons les liens conceptuels aux motifs obtenus par les principales méthodes d'extraction, à partir d'un réseau de référence 5.2(a).
(b) Les méthodes de clustering traditionnelles recherchent uniquement les groupes de noeuds qui partagent des propriétés communes. Dans ce type d'approche, la structure du réseau n'est pas du tout prise en compte.
Par exemple, le cluster contenant les noeuds 1, 2, 3, 4 et 5 est extrait, car les noeuds partagent tous la propriété x.
(c) Les méthodes de clustering basées sur les liens recherchent les groupes de noeuds densément connectés, c'est-à-dire les communautés. Seuls les liens du réseau sont considérés. Ainsi, les noeuds à l'intérieur d'une communauté sont fortement connectés, mais peuvent avoir des propriétés très différentes.
Par exemple, la communauté contenant les noeuds 1 et 2 et celle contenant les noeuds 3, 4 et 5, sont extraites du réseau initial puisqu'elles ne possèdent aucun lien entre elles.
(d) Les méthodes de clustering hybrides, qui prennent en compte les deux informations disponibles sur le réseau (structure et propriétés), extraient les groupes de noeuds fortement connectés, au sein desquels les noeuds possèdent des attributs homogènes. Ce type d'approche fournit un partitionnement plus sémantique du réseau, mais ne permet pas d'identifier les types de noeuds les plus connectés.
Par exemple, la communauté formée par les noeuds 6, 7 et 10 est à la fois densément connectée et partage la propriété commune y.
(e) Les méthodes de recherche de sous-graphes fréquents s'intéressent aux sous-structures retrouvées le plus fréquemment dans l'ensemble du réseau. Aucun regroupement de noeuds n'est effectué dans ce type d'approche.
Par exemple, le sous-graphe fréquent composé des étiquettes x et y est extrait du réseau initial.
(f) Les méthodes récentes d'analyse conceptuelle permettent d'extraire des groupes de noeuds ainsi que les voisins qu'ils partagent. Bien que ces approches effectuent un regroupement des noeuds en se basant sur les liens et les propriétés, elles ne s'intéressent

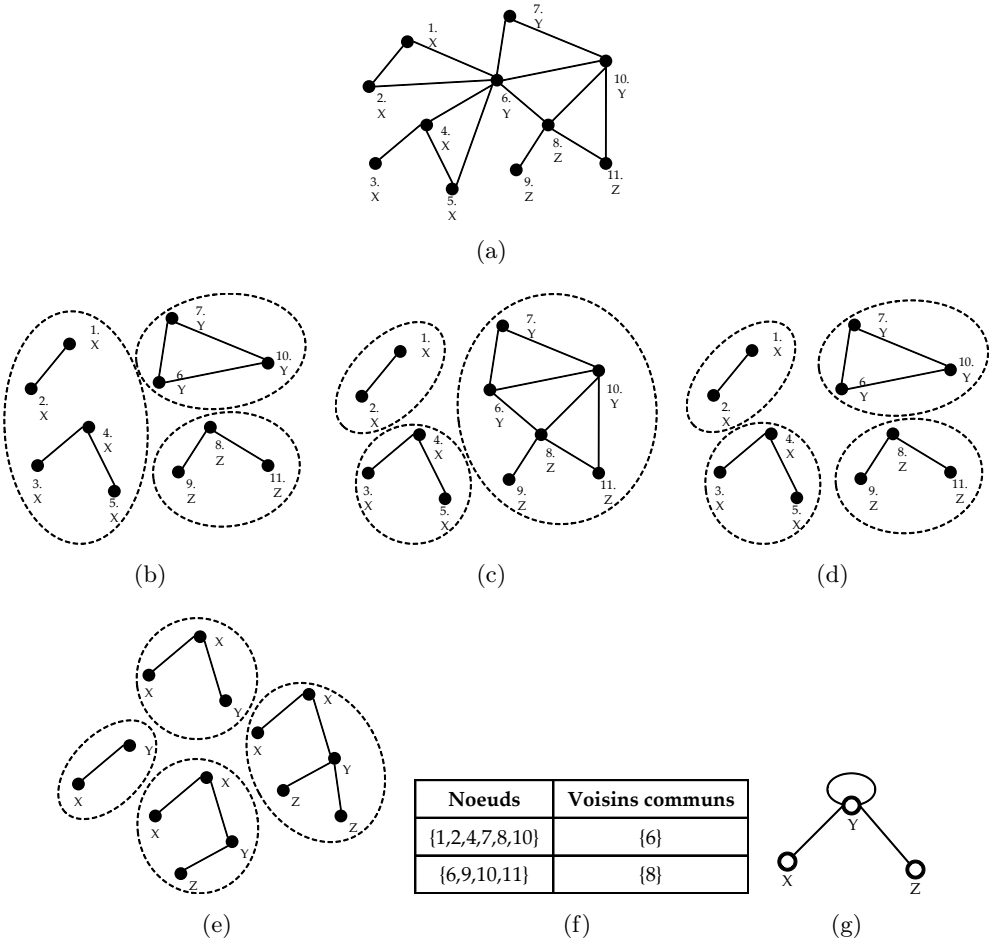


FIGURE 5.2 – Comparaisons des différents motifs selon les méthodes d'extraction (a) Réseau de référence, (b) Clust. classique, (c) Clust. liens, (d) Clust. hybride, (e) Sous-graphes fréquents, (f) Analyse conceptuelle, (g) Liens conceptuels

pas à la fréquence des motifs sur l'ensemble du réseau, mais plutôt à des caractéristiques partagées par des groupes de noeuds.
Par exemple, les noeuds 6, 9, 10 et 11 ont pour voisin commun le noeud 8.
(g) Enfin, l'extraction de liens conceptuels que nous proposons, permet d'obtenir des informations sur les groupes de noeuds les plus connectés du réseau. Les groupes étant définis par des ensembles de noeuds possédant des attributs communs.
Par exemple, les noeuds qui vérifient la propriété "X", sont fréquemment connectés à ceux qui vérifient la propriété "Y" ; les "Y" étant eux-mêmes fréquemment liés entre eux et à ceux qui vérifient la propriété "Z".

5.2 Liens et Vues conceptuels : définitions

Dans ce travail, nous proposons une nouvelle vision de la recherche de motifs fréquents dans les réseaux sociaux en redéfinissant la notion de *motif*. Les motifs recherchés, que nous appelons des "*liens conceptuels*", associent à la fois des informations sur la structure

du réseau et les attributs des noeuds. Plus précisément, un *lien conceptuel* est une notion abstraite qui représente un *ensemble de liens entre deux groupes de noeuds, dans lesquels les noeuds de chaque groupe partagent des attributs communs.* Quand ce type de motifs se retrouve fréquemment dans l'ensemble du réseau, il constitue un motif fréquent au sens traditionnel du terme et nous l'appelons "*lien conceptuel fréquent*".

Plus formellement, soit $G = (V, E)$ un réseau dans lequel V est l'ensemble des noeuds et E l'ensemble des liens avec $E \subseteq V \times V$.
V est défini comme une relation $R(A_1, ..., A_p)$ où chaque A_i est un attribut.
Chaque noeud $v \in V$ est défini par un tuple $(a_1, ..., a_p)$ où $\forall k \in [1..p],\ v[A_k] = a_k$.
a_k étant la valeur de l'attribut A_k dans v et $p = |R|$ le nombre d'attributs.

Un item est une expression logique $A = x$, où A est un attribut et x une valeur. L'item vide est noté $\emptyset$. Un itemset est une conjonction d'items par exemple :

$$A_1 = x \quad \text{et} \quad A_2 = y \quad \text{et} \quad A_3 = z$$

Quand un itemset est une conjonction de k items non-vides, il est appelé un k-itemset.

Chaque noeud du réseau peut donc être associé à un itemset. Nous notons I_V l'ensemble de tous les itemsets construits à partir de V.
Soit $G = (V, E)$ un *réseau dirigé uniparti*, pour tout itemset $m \in I_V$, nous notons V_m l'ensemble des noeuds de V qui satisfont m, c.-à-d.

$$V_m = \{v \in V \ ;\ v \text{ satisfait } m\} \tag{5.1}$$

Soient m et sm deux itemsets. sm est appelé *sous-itemset* de m et on note $sm \subseteq m$ si par exemple, $m = xyz$ et $sm = xy$.
Si sm est un sous-itemsets de m, on a $V_m \subseteq V_{sm}$.
m est appelé *sur-itemset* de sm.
Tout itemset est un sous-itemset de lui-même.

Soient $G = (V, E)$ un réseau, I_v l'ensemble des itemsets construits à partir de V et $m \in I_V$ un itemset. Nous définissons :

- Le *m-ensemble de liens de G du coté gauche*, LE_m, comme l'ensemble des liens de E qui partent de noeuds vérifiant m, c.-à-d.

$$LE_m = \{e \in E \ ;\ e = (a, b) \quad a \in V_m\} \tag{5.2}$$

- Le *m-ensemble de liens de G du coté droit*, RE_m, comme l'ensemble des liens de E qui arrivent à des noeuds vérifiant m, c.-à-d.

$$RE_m = \{e \in E \ ;\ e = (a, b) \quad b \in V_m\} \tag{5.3}$$

Définition 1. Lien conceptuel. Soient deux itemsets m_1 et m_2 appartenant à I_V. Nous définissons l'ensemble des liens connectant des noeuds de V_{m_1} (c.-à-d. noeuds vérifiant m_1) à des noeuds de V_{m_2} (c.-à-d. noeuds vérifiant m_2) comme le *lien conceptuel* (m_1, m_2) de G.

$$(m_1, m_2) = \{e \in E \ ;\ e = (a, b) \quad a \in V_{m_1} \text{ et } b \in V_{m_2}\} \tag{5.4}$$

Par exemple, si m_1 est l'itemset cd et que m_2 est l'itemset efj, le *lien conceptuel* $(m_1, m_2) = (cd, efj)$ correspond à tous les liens de E entre les noeuds de V qui satisfont

la propriété cd et les noeuds de V qui satisfont la propriété efj.

Notons L_V l'ensemble de tous les liens conceptuels qui peuvent être définis sur $G = (V, E)$. Pour tout élément (m_1, m_2) de L_V, on a :

$$(m_1, m_2) = LE_{m_1} \cap RE_{m_2} \tag{5.5}$$

Définition 2. Support d'un lien conceptuel. Nous appelons *support* de tout élément $l = (m_1, m_2)$ de L_V, le pourcentage de liens dans E qui appartiennent à l.

$$supp(l) = \frac{|\{e \in E \ ; \ e = (a, b) \quad a \in V_{m_1} \ et \ b \in V_{m_2}\}|}{|E|} \tag{5.6}$$

Remarquons que pour tout itemset m et tout lien conceptuel l, si $l = (\emptyset, m)$ ou $l = (m, \emptyset)$, alors $supp(l) = 0$.

Définition 3. Lien conceptuel fréquent. Un lien conceptuel l de L_V est dit *fréquent* si son support est supérieur à un seuil minimum de support β.

$$supp(l) > \beta \tag{5.7}$$

Nous notons FL_V l'ensemble des liens conceptuels fréquents (FCL) de $G = (V, E)$ selon un seuil de support donné β.

$$FL_V = \bigcup_{m_1 \in I_V, m_2 \in I_V} \{ \ (m_1, m_2) \in I_V^2 \ ; \ \frac{|(m_1, m_2)|}{|E|} > \beta \ \} \tag{5.8}$$

Propriété 1. Propriété de fréquence. Si le lien conceptuel (m_1, m_2) est fréquent, alors les ensembles LE_{m_1} et RE_{m_2} satisfont la propriété suivante :

$$|LE_{m_1}| > \beta \times |E| \quad et \quad |RE_{m_2}| > \beta \times |E| \tag{5.9}$$

Preuve. Si (m_1, m_2) est un lien conceptuel fréquent, alors

$$\begin{aligned} \frac{|(m_1, m_2)|}{|E|} > \beta & \\ & \Rightarrow \frac{|LE_{m_1} \cap RE_{m_2}|}{|E|} > \beta \\ & \Rightarrow |LE_{m_1} \cap RE_{m_2}| > \beta \times |E| \\ & \Rightarrow |LE_{m_1}| > \beta \times |E| \quad \text{et} \quad |RE_{m_2}| > \beta \times |E| \end{aligned}$$

La relation $\subseteq$ qui définie un ordre partiel sur I_V peut être étendue à L_V. Ainsi, comme pour la recherche classique d'itemsets fréquents [Agrawal 1994], certains des motifs extraits peuvent être inclus dans d'autres. De plus, comme nous le montrons ci-dessous, $(L_V, \subseteq)$ induit un treillis de concepts dans lequel il est possible d'extraire les liens conceptuels fréquents dits *maximaux* ($MFCL$).
Avant de présenter plus formellement la notion de $MFCL$, nous introduisons la notion de *sous-lien conceptuel*.

Définition 4. Sous et Sur-lien conceptuel. Soient les itemsets sm_1, sm_2, m_1 et m_2 tels que $sm_1 \subseteq m_1$ et $sm_2 \subseteq m_2$.
Le lien conceptuel (sm_1, sm_2) est appelé un *sous-lien conceptuel* de (m_1, m_2).
De même, (m_1, m_2) est appelé un *sur-lien conceptuel* de (sm_1, sm_2).

Par simplicité, nous notons $(sm_1, sm_2) \subseteq (m_1, m_2)$.

Propriété 2. Propriété de fermeture descendante. Si un lien conceptuel est fréquent, alors tous ses sous-liens conceptuels sont également fréquents.
Si un lien conceptuel est non-fréquent, alors tous ses sur-liens conceptuels sont également non-fréquents.

Preuve. Soient sm_1 et sm_2 respectivement des sous-itemsets de m_1 et m_2. Les propriétés $V_{m_1} \subseteq V_{sm_1}$ et $V_{m_2} \subseteq V_{sm_2}$ sont conservées.
Ce qui entraine $|(m_1, m_2)| \leq |(sm_1, sm_2)|$.

Définition 5. Lien conceptuel fréquent maximal. Soit β un seuil minimum de support donné, nous appelons *lien conceptuel fréquent maximal*, tout lien conceptuel fréquent l, tel qu'il n'existe aucun sur-lien conceptuel l' de l qui soit également fréquent :
l est maximal si et seulement si $\nexists l' \in FL_V$ tel que $l \subset l'$.

Nous notons $FL_{V_{max}}$ l'ensemble des *FCLs* maximaux de FL_V

$$FL_{V_{max}} = \bigcup_{m_1 \in I_V, m_2 \in I_V} \{ (m_1, m_2) \in FL_V \ ; \ (m_1, m_2) \ \ maximal \ \} \tag{5.10}$$

L'ensemble des *FCLs* fournit une *vue conceptuelle* du réseau, dans le sens où ces motifs apportent une connaissance sur les groupes de noeuds partageant des propriétés *internes* communes (*les concepts*) et étant les plus connectés du réseau social.
Ainsi, l'ensemble des *MFCLs* fournit une vue conceptuelle, une sorte de synthèse du réseau social à travers les motifs extraits.
Typiquement, si les liens (b, a) et (ab, a) sont des *FCLs* et si (ab, a) est un lien conceptuel fréquent maximal, alors le *MFCL* (ab, a) représente le concept social : "*parmi les liens de G, le taux de ceux qui connectent des noeuds vérifiant les propriétés a et b, à des noeuds vérifiant la propriété a sont supérieurs à un seuil* β".

Selon la théorie de l'analyse de concepts formels [Ganter 2005], nous devrions écrire ce lien conceptuel comme suit : $((\{v_1, v_2\}, ab), (\{v_3, v_4, v_5\}, a)$, où $(\{v_1, v_2\}, ab)$ et $(\{v_3, v_4, v_5\}, a)$ sont les deux concepts impliqués dans le lien conceptuel, chacun étant défini par son *intention* ab ou a (les propriétés) et son extension $\{v_1, v_2\}$ ou $\{v_3, v_4, v_5\}$ (l'ensemble des *noeuds-objets* qui satisfont les propriétés).
En ce sens, le lien conceptuel $((\{v_1, v_2\}, ab), (\{v_3, v_4, v_5\}, a))$ est un élément de $P(V)^2 \times I_V^2$.

Définition 6. Relation binaire. Nous définissons la relation binaire $\subseteq$ sur $P(V)^2 \times I_V^2$ comme suit :
Soient deux liens conceptuels $((x, y), (m, p))$ et $((x', y'), (m', p'))$ dans $P(V)^2 \times I_V^2$, on a :
$((x, y), (m, p)) \subseteq ((x', y'), (m', p'))$ si et seulement si $(x', y') \subseteq (x, y)$ et $(m, p) \subseteq (m', p')$.

Propriété 3. Treillis de concepts sociaux. $(P(V)^2 \times I_V^2, \subseteq)$ est un treillis de concepts. Nous l'appelons le *treillis de concepts sociaux* défini sur L_V.

Preuve. Les propriétés standards des treillis de concepts classiques $(P(V) \times I_V, \subseteq)$ sont conservées.

Définition 7. Vue conceptuelle du réseau social. Soient $G = (V, E)$ un réseau social, β un seuil de support minimum et $FL_{V_{max}}$ l'ensemble des liens conceptuels fréquents maximaux de FL_V.
FL_V fournit une *vue conceptuelle* du réseau, dans laquelle seuls les liens pertinents entre

les groupes de noeuds les plus connectés sont représentés.
$FL_{V_{max}}$ fournit, lui, une *vue conceptuelle synthétique* de G, puisque seuls les liens conceptuels maximaux y sont représentés.

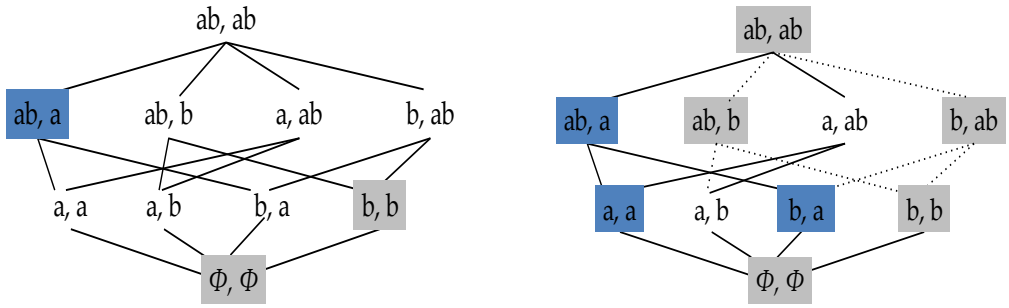


FIGURE 5.3 – Exemple de treillis de concepts sociaux
(a) (ab, b) est fréquent et (b, b) est non-fréquent, (b) exemple de branches élaguées

Sur la base de la propriété de fermeture descendante (voir Propriété 2), l'extraction des *MFCLs* peut être optimisée en élaguant certaines branches du treillis de concepts sociaux, comme le montre la Figure 5.3.
Typiquement, lors d'une recherche descendante dans le treillis, quand un *MFCL* comme (ab, a) est trouvé, tous ses sous-liens conceptuels peuvent être élagués puisqu'ils ne sont pas des *MFCLs*.
De même, lors d'une recherche ascendante dans le treillis, quand un lien conceptuel non-fréquent comme (b, b) est trouvé, tous ses sur-liens conceptuels peuvent alors être élagués, puisqu'ils ne peuvent pas être fréquents.

5.3 Extraction des liens conceptuels fréquents maximaux

La recherche des liens conceptuels fréquents maximaux dans un réseau social peut être coûteuse si l'espace de recherche est de taille importante. Les problèmes analogues liés au comptage ou à la recherche de motifs fréquents dans les jeux de données traditionnels sont connus comme étant $\#P$ et NP-difficiles. C'est la raison pour laquelle dans ce travail, nous proposons l'algorithme d'extraction *MFCL-Min* (Maximal Frequent Conceptual Links Mining), qui effectue une recherche *ascendante* des *MFCLs* au sein du treillis de concepts sociaux, tout en réduisant progressivement l'espace de recherche. Notre algorithme ne nécessite aucune connaissance a priori sur le réseau.

Dans un premier temps, la Section 5.3.1 détaille notre approche pour le cas général des réseaux orientés et unipartis. Puis dans la Section 5.3.2, nous montrons que notre approche est adaptable à tout type de réseaux et nous discutons la complexité de notre solution et les optimisations qui peuvent y être apportées. Enfin, la Section 5.3.3 présente quelques mesures permettant d'évaluer la pertinence d'un lien conceptuel avec plus d'intérêt.

5.3.1 L'algorithme *MFCL-Min*

Dans le domaine des réseaux, il est généralement admis que le nombre de liens est un paramètre clé dans les phases de calcul et de traitement. Ainsi, concevoir un algorithme qui extrait l'ensemble des *MFCLs* d'un réseau social peut nécessiter des calculs dont le coût n'est pas envisageable. Dans une approche naïve, l'extraction des *MFCLs* s'effectue en deux étapes. (i) Tous les itemsets possibles sont générés à partir des attributs des noeuds. (ii) La fréquence de chaque paire d'itemsets, c'est-à-dire d'un lien conceptuel, est ensuite évaluée. Deux étapes qui rendent très vite une telle approche inadaptée à un grand jeu de données.

L'Algorithme que nous proposons, *MFCL-Min* (voir Algorithme 6), inspiré des principes de Apriori, effectue une recherche ascendante des liens conceptuels fréquents maximaux et applique les propriétés 1 et 2 (voir Section 5.2) pour restreindre l'espace de recherche uniquement aux sur-liens conceptuels potentiellement impliqués dans des *MFCLs*. Plus généralement, les *MFCLs* impliquant des t-itemsets sont recherchés exclusivement à travers les sur-liens conceptuels des *FCLs* impliquant des $(t-1)$-itemsets.

A l'état initial $t = 1$, *MFCL-Min* construit les ensembles LI_1 et RI_1, des structures permettant de stocker respectivement l'ensemble des 1-itemsets m_1 et m_2 potentiellement impliqués de chaque côté des liens conceptuels fréquents selon la *Propriété de fréquence 1* (voir Lignes 6-7 Algorithme 6).
Les *FCLs* sont ensuite recherchés parmi ces itemsets et stockés dans une liste temporaire L, chargée de collecter à chaque itération t, l'ensemble des *FCLs* impliquant uniquement des t-itemsets. Ainsi, à $t = 1$, seuls les *FCLs* reliant des 1-itemsets y sont stockés. Durant cette première phase, les itemsets impliqués du *côté-gauche* et ceux impliqués du *côté-droit* sont stockés dans des piles séparées pour les traitements futurs. Tous les *FCLs* extraits lors de cette première étape sont considérés momentanément comme maximaux et sont donc ajoutés à $FL_{V_{max}}$ (voir Lignes 8-17).

Ensuite à chaque itération t, les itemsets candidats du *côté-gauche* et du *côté-droit* sont générés et stockés respectivement dans les piles LI_{cand} et RI_{cand} (voir Lignes 20-21). De chaque côté, ces itemsets candidats sont obtenus à partir de l'union
(i) des sur-itemsets formés à partir des $(t-1)$-itemsets impliqués du *côté-gauche* (ou du droit) des *FCLs* de L, c'est-à-dire qu'ils sont les t-itemsets potentiellement impliqués dans des liens conceptuels fréquents selon les propriétés 1 et 2, et
(ii) de tous les $(t-k)$-itemsets, avec $k \in [1..(t-1)]$, déjà impliqués dans des liens conceptuels fréquents (puisque les *MFCLs* n'incluent pas nécessairement des itemsets maximaux). C'est cette inclusion des itemsets déjà impliqués qui permet éventuellement de retrouver des liens conceptuels fréquents entre des itemsets de taille t et des itemsets de taille $(t-k)$.
Notons que les éléments des ensembles LI_{cand} et RI_{cand} sont triés du plus grand au plus petit (en termes de nombre d'items).
Une fois les candidats de chaque côté générés, la liste L est vidée et les *MFCLs* sont recherchés parmi les candidats (voir Lignes 22-31). La comparaison est effectuée uniquement si au moins un des deux candidats à une taille de t (dans le but de ne pas s'intéresser à des sous-liens conceptuels déjà traités) et si un sur-lien conceptuel n'est pas déjà présent dans L. En effet, comme les structures LI_{cand} et RI_{cand} sont ordonnées, les premiers liens conceptuels fréquents identifiés sont nécessairement les maximaux au regard de l'itération t. Ainsi, la comparaison est effectuée pour vérifier si un sur-lien conceptuel n'a pas déjà été ajouté à L, ou plus formellement si (m_1, m_2) est maximal. Si (m_1, m_2) est un *MFCL*, il est ajouté à L et $FL_{V_{max}}$ et tous ses sous-liens conceptuels sont supprimés de $FL_{V_{max}}$.
Ces opérations sont répétées jusqu'à ce que plus aucun *FCL* ne soit détecté, ou que toutes les combinaisons soient effectuées, c.-à-d. $t = |R|$, (voir Lignes 19-33).

algorithme 6 Algorithme *MFCL-Min* pour l'extraction des *MFCLs*

Précondition : $G = (V, E)$: **Réseau**, et $\beta \in [0..1]$: **Seuil de support minimum**

1. $FL_{V_{max}}$: **Ensemble des *MFCLs*** $\leftarrow \emptyset$
2. LI_{cand} : **Pile des itemsets candidats au *côté-gauche*** $\leftarrow \emptyset$
3. RI_{cand} : **Pile des itemsets candidats au *côté-droit*** $\leftarrow \emptyset$
4. L : **Liste tampon de *FCLs*** $\leftarrow \emptyset$
5. t : **Itération** $\leftarrow 1$

   %*Génération des liens conceptuels fréquents impliquant des* 1*-itemsets*
6. $LI_1 \leftarrow$ Générer 1-itemsets m à partir de V tels que $|LE_m| > \beta \times |E|$
7. $RI_1 \leftarrow$ Générer 1-itemsets m à partir de V tels que $|RE_m| > \beta \times |E|$
8. **pour tout** itemset $m_1 \in LI_1$ **faire**
9. **pour tout** itemset $m_2 \in RI_1$ **faire**
10. **si** $|(m_1, m_2)| > \beta \times |E|$ **alors**
11. ajouter m_1 à LI_{cand}
12. ajouter m_2 à RI_{cand}
13. ajouter (m_1, m_2) à L
14. ajouter (m_1, m_2) à $FL_{V_{max}}$
15. **fin si**
16. **fin pour**
17. **fin pour**

    %*Génération des autres liens conceptuels fréquents*
18. $t \leftarrow t + 1$
19. **tant que** $L \neq \emptyset$ et *toutesCombinaisons() = faux* **faire**
20. $LI_{cand} \leftarrow$ {Jointure de tous les $(t-1)$-itemsets distincts m impliqués du *côté-gauche* de L et partageant $(t-2)$ items, tels que $|LE_m| > \beta \times |E|$} $\bigcup$ LI_{cand}
21. $RI_{cand} \leftarrow$ {Jointure de tous les $(t-1)$-itemsets distincts m impliqués du *côté-droit* de L et partageant $(t-2)$ items, tels que $|RE_m| > \beta \times |E|$} $\bigcup$ RI_{cand}
22. $L \leftarrow \emptyset$
23. **pour tout** itemset $m_1 \in LI_{cand}$ **faire**
24. **pour tout** itemset $m_2 \in RI_{cand}$ **faire**
25. **si** $((|m_1| = t$ ou $|m_2| = t)$ et $\nexists l \in L$ tel que $(m_1, m_2) \subset l$ et $|(m_1, m_2)| > \beta \times |E|)$ **alors**
26. ajouter (m_1, m_2) à L
27. supprimer tous les *FCLs* $q \in FL_{V_{max}}$ tels que $q \subset (m_1, m_2)$
28. ajouter (m_1, m_2) à $FL_{V_{max}}$
29. **fin si**
30. **fin pour**
31. **fin pour**
32. $t \leftarrow t + 1$
33. **fin tant que**
34. **retour** $FL_{V_{max}}$

5.3.2 Discussion

Les réseaux du monde réel sont souvent de natures diverses : orientés, non-orientés, unipartis, multipartis. Une caractéristique intéressante de tout algorithme qui vise à analyser les réseaux sociaux est sa capacité à s'adapter à tout type de réseaux. C'est le cas de l'Algorithme *MFCL-Min*.

Il est courant de représenter un réseau non-dirigé comme un réseau dirigé dans lequel les

liens sont décrits dans les deux directions. De cette façon, *MFCL-Min* peut être appliqué directement. Cependant, remarquons que la propriété de fréquence définie sur les liens conceptuels devient symétrique : si (m_1, m_2) est un *FCL*, alors (m_2, m_1) est également un *FCL*. Ainsi, comme le montre l'Algorithme 7, *MFCL-Min* peut être simplifié pour éviter les comparaisons symétriques, en utilisant un seul ensemble I_1 au lieu des deux ensembles LI_1 et RI_1 utilisés pour le stockage des 1-itemset initiaux.

algorithme 7 Adaptation des lignes 6-17 de *MFCL-Min* pour réseaux non-dirigés

Précondition : $G = (V, E)$: **Réseau**, et $\beta \in [0..1]$: **Seuil de support minimum**
1. m_1, m_2 : **itemset**
2. $I_1 \leftarrow$ Générer 1-itemsets m à partir de V tel que $|LE_m| > \beta \times |E|$
3. **pour tout** m_1 de I_1 allant de l'indice 0 à $I_V.taille$ **faire**
4. **pour tout** m_2 de I_1 allant de l'indice de m_1 à $I_V.taille$ **faire**
5. %*Les lignes 10-15 de l'Algorithme MFCL-Min restent inchangées*
6. **fin pour**
7. **fin pour**

En ce qui concerne les réseaux multipartis, l'algorithme peut également être appliqué avec des adaptations légères pour éviter les comparaisons inutiles qui sont effectuées, puisque les 1-itemsets initiaux sont générés à partir de l'ensemble des noeuds du réseau (voir Lignes 6-7 de l'Algorithme 6). Si nous disposons d'informations sur les catégories de noeud impliquées de chaque côté des liens, la génération des ensembles LI_1 et RI_1 peut être effectuée selon la catégorie. Dans le cas d'un réseau biparti par exemple, avec $G = (V_A, V_B, E)$ et $E \subseteq V_A \times V_B$, les 1-itemsets initiaux doivent être générés séparément selon qu'ils soient de type A ou de type B. Les Lignes 6-7 de l'Algorithme 6 peuvent ainsi être remplacées par :

- $LI_1 \leftarrow$ Générer 1-itemsets m à partir de V_A tels que $|LE_m| > \beta \times |E|$
- $RI_1 \leftarrow$ Générer 1-itemsets m à partir de V_B tels que $|RE_m| > \beta \times |E|$

Concernant la complexité de l'algorithme *MFCL-Min*, la phase de calcul la plus importante vient de l'évaluation du support des liens conceptuels (m_1, m_2) à chaque itération. Un moyen simple et efficace d'implémenter cette tâche et d'accélérer le processus consiste à utiliser une structure de noeud qui stocke à la fois les liens entrant et les liens sortant. Ainsi, plutôt que de parcourir l'ensemble des liens du réseau, la recherche peut être restreinte aux noeuds comme nous le détaillons sur l'Algorithme 8.

algorithme 8 Optimisation de la génération des ensembles (m_1, m_2)

Précondition : $G = (V, E)$: **Réseau**, m_1 : **itemset**, m_2 : **itemset**
1. LE_{m_1} : **Ensemble de liens** $\leftarrow \emptyset$
2. RE_{m_2} : **Ensemble de liens** $\leftarrow \emptyset$
3. **pour tout** noeud $v \in V$ **faire**
4. **si** v satisfait m_1 **alors**
5. Ajouter tous les liens sortant de v à LE_{m_1}
6. **fin si**
7. **si** v satisfait m_2 **alors**
8. Ajouter tous les liens entrant de v à RE_{m_2}
9. **fin si**
10. **fin pour**
11. $(m_1, m_2) \leftarrow LE_{m_1} \cap RE_{m_2}$

Deux paramètres sont impliqués dans le nombre de comparaisons à effectuer : le nombre d'attributs $|R|$ et la taille du réseau, caractérisée par le nombre de noeuds $|V|$ et le nombre de liens $|E|$.

Prenons le cas d'un réseau dont les noeuds ne contiennent que des attributs binaires. Comme nous l'avons expliqué précédemment, le nombre de comparaisons $\Theta(AN)$ effectuées par une approche d'extraction naïve (AN) peut être approchée par :

$$\Theta(AN) = 2^{|R|} \times 2^{|R|} \times |E| \tag{5.11}$$

Pour comparer cette approche à *MFCL-Min*, considérons dans un premier temps le cas extrême d'un réseau complet (*Configuration C1*). Dans une telle configuration, tous les liens conceptuels sont fréquents et donc, la totalité du treillis de concepts sociaux doit être explorée pour extraire les *MFCLs*. Dans cette configuration empirique, le nombre de comparaisons effectuées par *MFCL-Min*, noté $\Theta(C1)$, est borné par :

$$\Theta(C1) \leq |R| \times |R| + \sum_{k=1}^{|R|} C_{|R|}^{k} \times C_{|R|}^{k} \times |V| \tag{5.12}$$

Considérons maintenant la configuration extrême duale, dans laquelle aucun lien conceptuel n'est fréquent (*Configuration C2*). Lors de la recherche des éventuels *FCLs* impliquant des 1-itemsets (voir Lignes 8-17 de l'Algorithme 6), *MFCL-Min* est en mesure de détecter qu'aucun lien conceptuel n'est fréquent. Dans une telle configuration, le nombre de comparaisons nécessaires, $\Theta(C2)$, est borné par :

$$\Theta(C2) \leq |R| \times |R| \tag{5.13}$$

Prenons l'exemple d'un réseau complet $G = (V, E)$, avec $|V| = 10000$ et $|E| = |V| \times (|V| - 1)$. La Figure 5.4 montre, en fonction du nombre d'attributs $|R|$, comment évoluent (a) le nombre d'opération de comparaisons en logarithme et (b) le gain associé en nombre de comparaisons par rapport à l'approche naïve.

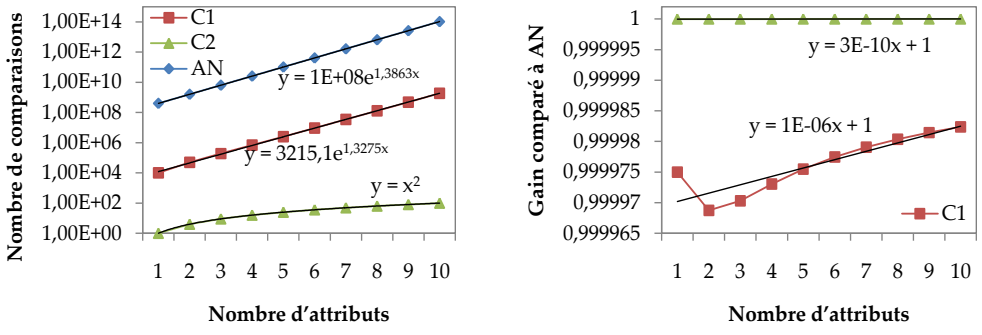


FIGURE 5.4 – Estimation du nombre de comparaisons pour différentes configurations (a) Log. du nombre de comparaisons et (b) Gain comparé à l'approche naïve

La croissance linéaire du logarithme observée sur la Figure 5.4(a), confirme que le nombre de comparaisons pour une configuration de type $C1$ croît exponentiellement avec le nombre d'attributs $|R|$.

Toutefois, nous pouvons également observer sur la Figure 5.4(b) que pour les configurations $C1$ et $C2$, le gain sur le nombre de comparaisons reste significatif comparé à l'approche naïve, puisqu'il est systématiquement supérieur à 99%.

5.3.3 Mesures d'intérêt sur les liens conceptuels

Nous pouvons nous interroger sur la façon d'évaluer la pertinence d'un lien conceptuel fréquent sous un autre point de vue que son support. En fouille de données, le problème des mesures d'intérêt pour évaluer et ordonner les modèles extraits, comme les règles d'association, a été abordé de différente manière [Collard 2007, Brisson 2008].
Par exemple, la mesure de support que nous proposons (voir Équation 5.6) ne donne aucune indication sur des propriétés plus précises telles que le pourcentage de liens partant de noeuds vérifiant m_1 et arrivant à des noeuds vérifiant m_2, ou même la proportion de noeuds appartenant à V_{m_2} qui reçoivent une connexion de noeuds dans V_{m_1}.
Suivant l'exemple des travaux menés en fouille de données classique, d'autres mesures peuvent être utilisées pour mettre en évidence des propriétés intéressantes des *FCLs*.

Soient deux seuils de pertinence $\gamma_{\leftarrow}$ et $\gamma_{\rightarrow}$, nous proposons trois mesures d'intérêt qui possèdent une sémantique particulière et qui apportent chacune, une information spécifique sur le lien conceptuel fréquent.

La première est la **mesure de dépendance** $m_1 \dashrightarrow m_2$, qui évalue la proportion de liens qui partent de noeuds dans V_{m_1} (extrémité gauche) et qui connectent des noeuds de V_{m_2} (extrémité droite).
Ainsi, un *FCL* (m_1, m_2) est dit pertinent selon la mesure de dépendance $m_1 \dashrightarrow m_2$ si :

$$\frac{|\ \{e \in E\ ;\ e = (a,b) \quad a \in V_{m_1}\ et\ b \in V_{m_2}\}\ |}{|\ \{e \in E\ ;\ e = (a,b) \quad a \in V_{m_1}\}\ |} > \gamma_{\rightarrow} \tag{5.14}$$

Symétriquement, nous pouvons mesurer la **dépendance** $m_1 \dashleftarrow m_2$. En d'autres termes, le pourcentage de liens qui connectent des noeuds dans V_{m_2} (extrémité droite) et qui ont pour origine des noeuds dans V_{m_1} (extrémité gauche). Cette mesure est particulièrement intéressante sur des réseaux bipartis tels que des réseaux utilisateur-musique ou utilisateur-livre, puisque le point d'intérêt peut être focalisé sur les noeuds d'arrivée plutôt que sur les noeuds de départ.
Un *FCL* (m_1, m_2) est dit pertinent selon $m_1 \dashleftarrow m_2$ si :

$$\frac{|\ \{e \in E\ ;\ e = (a,b) \quad a \in V_{m_1}\ et\ b \in V_{m_2}\}\ |}{|\ \{e \in E\ ;\ e = (a,b) \quad b \in V_{m_2}\}\ |} > \gamma_{\leftarrow} \tag{5.15}$$

Cependant, les mesures (5.14) et (5.15) peuvent suggérer un lien de causalité qui n'est pas statistiquement vrai. Nous proposons donc un **test de dépendance symétrique**, basé sur le même principe que le *lift*.
Un *FCL* (m_1, m_2) est dit pertinent selon le test de dépendance symétrique si :

$$\frac{|E| \times |\ \{e \in E\ ;\ e = (a,b),\quad a \in V_{m_1}\ et\ b \in V_{m_2}\}\ |}{|\ \{e \in E\ ;\ e = (a,b) \quad a \in V_{m_1}\}\ | \times |\ \{e \in E\ ;\ e = (a,b) \quad b \in V_{m_2}\}\ |} > 1 \tag{5.16}$$

Enfin, nous définissons **la précision d'un lien conceptuel** comme étant le plus grand nombre d'attributs contenu dans les concepts impliqués.
Soit $l = (m_1, m_2)$ un lien conceptuel, la précision $\sigma(l)$ est obtenue par :

$$\sigma(l) = max(|m_1|, |m_2|) \tag{5.17}$$

5.4 Génération de vues conceptuelles

Avec les liens conceptuels extraits du réseau social G, la connaissance acquise peut être synthétisée à travers une structure de graphe. Nous appelons cette structure la *vue*

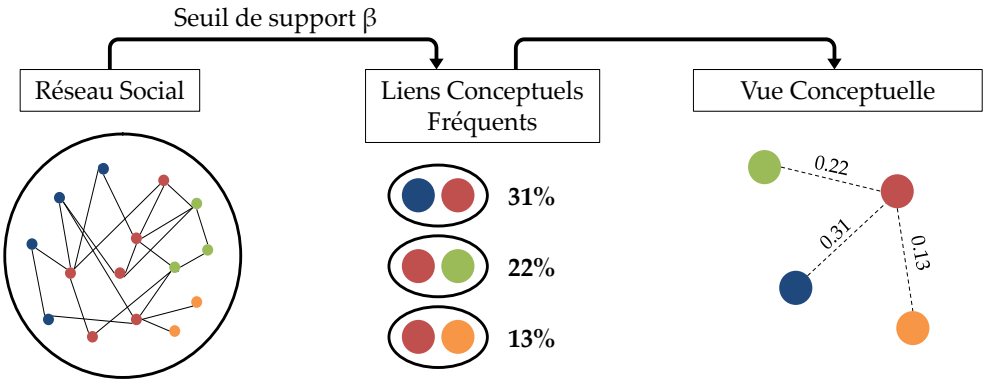


FIGURE 5.5 – Différentes étapes de la génération des vues conceptuelles

conceptuelle de G, dans le sens où elle résume l'ensemble des liens entre les groupes de noeuds, et par conséquent les ensembles de propriétés, les plus connectés dans le réseau. La vue conceptuelle est une *représentation sémantique* du réseau, dans laquelle un noeud représente une propriété présente dans G et une liaison décrit la notion sémantique "*être fréquemment connecté dans le réseau initial*".

Ainsi dans une vue conceptuelle, les noeuds correspondent aux itemsets alors que les connexions représentent un lien conceptuel fréquent.
Plus formellement, soient $G = (V, E)$ un réseau social, I_V l'ensemble des itemsets construits à partir de V et β un seuil de support minimum. La vue conceptuelle $CView$ de G est une structure $CView = (V_{concept}, E_{concept})$ définie comme suit :
$V_{concept}$ est l'ensemble des meta-noeuds V_m tel que $\exists m' \in I_V$ avec $(m, m') \in FL_V$ ou $(m', m) \in FL_V$.
$E_{concept}$ est l'ensemble des liens $(x, y) \in V_{concept} \times V_{concept}$ tel que $x = V_{m_1}$, $y = V_{m_2}$ et $(m_1, m_2) \in FL_V$.
Un noeud dans V peut ainsi être associé à plusieurs meta-noeuds dans $V_{concept}$. Un lien dans $E_{concept}$ représente alors une relation sémantique entre deux groupes de noeuds dans V, qui sont chacun définis par une ensemble de propriétés. Ainsi, tout lien $(x, y) \in E_{concept}$ signifie qu'il existe un lien conceptuel fréquent entre les noeuds appartenant à V_x et ceux appartenant à V_y. Toutefois, certains noeuds de V_x, ou de V_y, peuvent ne pas participer à ce lien conceptuel.
Pour un seuil de support minimum β donné, l'ensemble des liens conceptuels fréquents maximaux, $FL_{V_{max}}$, fournit une vue conceptuelle synthétique, puisque seuls les liens conceptuels maximaux y sont représentés.

Notons que la vue conceptuelle conserve les mêmes caractéristiques que le réseau initial G. Si G est non-orienté (resp. orienté), $CView$ est également non-orienté (resp. orienté). De même, si G est biparti, $CView$ l'est également.

Comme le montre la Figure 5.5, le processus de génération des vues conceptuelles synthétiques s'effectue en deux étapes consécutives :

1. Les *MFCLs* sont extraits à partir du réseau initial G
2. La vue conceptuelle synthétique est générée à partir de $FL_{V_{max}}$.

L'algorithme 9, *CView-MFCL* (Conceptual View Based on Maximal Frequent Conceptual Links), détaille la procédure de génération de la vue conceptuelle.

Il est important de préciser que la vue conceptuelle n'est pas une représentation directe ou allégée du réseau initial. En effet, il n'existe aucune correspondance simple entre la vue

algorithme 9 *CView-MFCL* pour la génération de vue conceptuelle synthétique

Précondition : $FL_{V_{max}}$: **Ensemble des liens conceptuels fréquents maximaux**

1. $V_{concept}$: **Ensemble d'itemsets (meta-noeud)** $\leftarrow \emptyset$
2. $E_{concept}$: **Lien conceptuel** $\leftarrow \emptyset$
3. $CView \leftarrow (V_{concept}, E_{concept})$: **Vue conceptuelle**
4. **pour tout** MFCL $l = (m_1, m_2) \in FL_{V_{max}}$ **faire**
5. **si** $m_1 \notin V_{concept}$ **alors**
6. Ajouter m_1 à $V_{concept}$
7. **fin si**
8. **si** $m_2 \notin V_{concept}$ **alors**
9. Ajouter m_2 à $V_{concept}$
10. **fin si**
11. Ajouter $e = (m_1, m_2)$ à $E_{concept}$
12. **fin pour**
13. **retour** $CView$

conceptuelle et le réseau initial, puisque les noeuds du réseau initial peuvent être représentés par plusieurs meta-noeuds de la vue conceptuelle, et inversement, les meta-noeuds de la vue conceptuelle englobe un groupe de noeuds du réseau initial. En ce sens, la structure résultante n'est qu'une représentation sémantique, car elle peut être considérée comme une forme de représentation des connaissances acquises sur les connexions entre les groupes de noeuds les plus connectés du réseau social.

5.5 Résultats expérimentaux

Différents types d'expériences ont été menées pour évaluer l'efficacité de cette approche à la fois d'un point qualitatif, en nous intéressant aux motifs extraits et à leur aspect pratique, mais également d'un point de vue quantitatif en étudiant les performances de la méthode d'extraction proposée dans différentes configurations. Dans cette section, nous détaillons ces expériences et présentons les résultats obtenus.

La Section 5.5.1 décrit le réseau utilisé et les pré-traitements effectués avant la recherche de liens conceptuels fréquents maximaux. Dans la Section 5.5.2, nous abordons la question de la qualité des motifs extraits, en nous intéressant à la pertinence des *MFCLs* et à l'évolution de leur précision selon différentes configurations. Dans la Section 5.5.3, nous analysons le processus d'extraction d'un point de vue quantitatif en étudiant comment évolue le nombre de motifs extraits, le temps nécessaire à leur extraction et le gain obtenu comparé à l'approche naïve. Enfin, dans la Section 5.5.4, nous analysons les vues conceptuelles associées en étudiant l'évolution de leurs principales caractéristiques selon différents seuils.

5.5.1 Environnement de test

Le jeu de données utilisé dans nos expériences est un réseau de contacts de proximité géographique obtenu avec EpiSims [Barrett 2008], un outil de simulation qui reproduit les déplacements d'individus dans la ville de Portland. Deux individus sont connectés dans ce réseau s'ils sont géographiquement proches ; autrement dit, s'ils fréquentent les mêmes lieux. Les principales caractéristiques de ce réseau sont décrites sur la Figure 5.6.

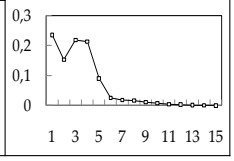
Comme le montre la Figure 5.6(a), le réseau contient 3000 individus et 4863 liaisons.

Les données ont été traitées de façon à ce que chaque noeud soit caractérisé par six attributs démographiques :

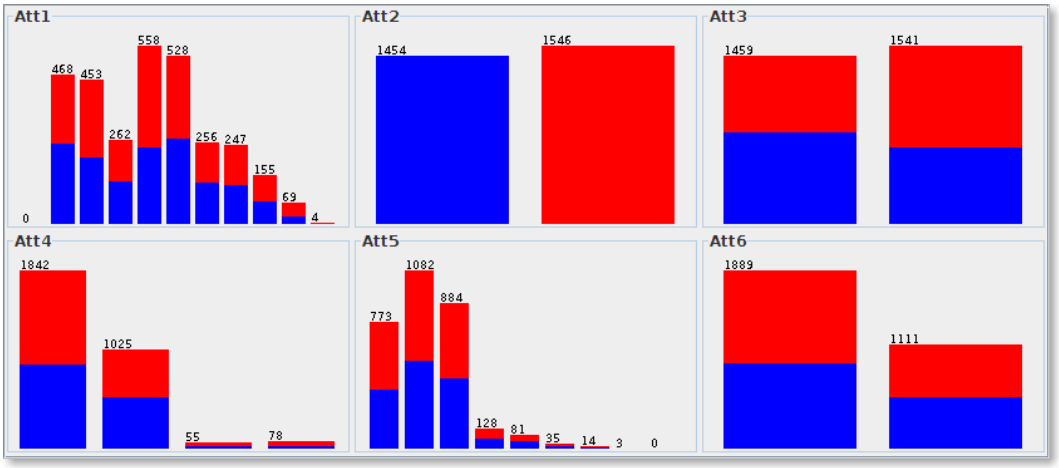
1. **La classe d'age**, obtenue par $\lfloor \frac{age}{10} \rfloor$
2. **Le sexe** (1-homme, 2-femme),
3. **Le statut-professionnel** (1-travailleur, 2-non-travailleur),
4. **Le type de relation avec le chef de famille** (1-conjoint, partenaire, ou chef de famille, 2-enfant, 3-parent adulte, 4-autre)
5. **La classe de contact**, obtenue par $\lfloor \frac{degr}{2} \rfloor$
6. **L'appartenance à une communauté** (1-coeff. clust. > 0.5, 2-*sinon*).

La répartition de ces attributs selon le sexe (*bleu*-homme et *rouge*-femme) est décrite sur la Figure 5.6(b).

General	Nature	Contacts de proximité
	Type	Non-Dirigé
	#nodes	3000
	#links	4683
	Density	0.00110413
cc	avg	0.63627

Degree	avg	3.215
	max	15
	Distribution	

(a)



(b)

FIGURE 5.6 – Principales caractéristiques du jeu de données utilisé
(a) Propriété du réseau (cc : coefficient de clustering) (b) répartition des attributs

La recherche de liens conceptuels fréquents sur ce réseau permet de mettre en évidence les corrélations entre les contacts entretenus par les individus et leurs propriétés. De tels motifs apportent des informations sur la façon dont les individus se rencontrent et échangent les uns avec les autres, permettant par exemple la transmission d'informations à travers ces liens sociaux, implicitement basés sur leurs caractéristiques communes et les contacts de proximité qu'ils maintiennent.

Dans nos expériences, nous avons fait varier la taille du réseau en extrayant des sous-graphes du réseau global. Les différentes tailles de réseau utilisées sont les suivantes $(|V|, |E|) = \{(500\ ,\ 806), (1000{,}1750), (1500{,}2685), (2000\ ,\ 3304), (2500{,}3988), (3000{,}4683)\}$. Par simplicité, nous ferons référence à la taille réseau dans ce qui suit, en nous référant uniquement à $|V|$. De même, nous faisons évoluer le nombre d'attributs $|R|$ en nous restreignant aux premiers attributs des noeuds. Par exemple, pour $|R| = 3$, seuls les attributs 1, 2 et 3 sont conservés.

MFCL-Min a été développé en JAVA et intégré à l'outil *GT-FCLMin* présenté dans la Section 5.6. Tous les tests ont été moyennés sur 100 exécutions et menés sur la configuration matérielle suivante : Intel Core 2 Duo P8600, 2.4Ghz, 3Go Ram, Linux Ubuntu 10.10 avec Java JDK 1.6.

5.5.2 Étude qualitative

Dans un premier temps, nous nous sommes intéressés à la pertinence des motifs extraits. Sur la Figure 5.7, nous commençons par comparer les *MFCLs* obtenus avec la configuration $|V| = 3000$ et $|R| = 4$ quand (a) $\beta = 0.10$ et (b) $\beta = 0.20$. Dans chaque motif, le caractère générique '$*$' signifie que l'attribut peut prendre n'importe quelle valeur.

MFCL	**Support**
$((4;*;1;*),(*;*;2;*))$	0.107
$((2;*;*;2),(*;*;2;2))$	0.105
$((*;1;1;*),(*;*;1;*))$	0.113
$((1;*;2;2),(*;1;*;*))$	0.102
$((*;1;1;1),(*;2;*;*))$	0.133

(a)

MFCL	**Support**
$((*;2;*;*),(*;*;1;*))$	0.231
$((*;1;*;*),(*;*;2;*))$	0.288
$((*;2;*;*),(*;1;*;*))$	0.297
$((*;*;2;*),(*;*;2;*))$	0.349
$((*;*;2;*),(*;*;1;*))$	0.213

(b)

FIGURE 5.7 – Exemples de liens conceptuels fréquents maximaux extraits avec la configuration $|V| = 3000$ et $|R| = 4$ quand (a) $\beta = 0.10$ et (b) $\beta = 0.20$

Sur cette figure, nous pouvons observer les types de noeuds les plus connectés du réseau qui sont mis en évidence. Par exemple, la première ligne du tableau de la Figure 5.7(a) indique que 10.7% des liens du réseau connectent un individu de 40 à 49 ans qui travaille à un individu qui ne travaille pas. Suivant le même principe de lecture, la première ligne du tableau de la Figure 5.7(b) montre que 23.1% des liens du réseau connectent une femme à un individu qui travaille. Rappelons que la sémantique de la relation est "*fréquenter les mêmes lieux*".

Bien que ces motifs soient pertinents, dans le sens où ils apportent une connaissance sur les liens fréquents entre des groupes de noeuds du réseau, nous observons qu'ils n'ont pas la même précision selon le seuil de support utilisé. En effet, les *MFCLs* extraits avec $\beta = 0.2$ sont naturellement plus généraux que ceux extraits avec $\beta = 0.10$, puisque les concepts impliqués couvrent une plus grande quantité de noeuds quand $\beta = 0.2$.
Ainsi, nous pouvons étudier comment se distribue cette précision selon le seuil de support minimum β. La Figure 5.8 présente ces résultats pour les configurations (a) $|V| = 500$ et (b) $|V| = 3000$.

Comme attendu, nous pouvons observer que les faibles valeurs de β fournissent les liens conceptuels fréquents maximaux les plus précis, c'est-à-dire ceux qui sont définis par le plus de propriétés. Dans ces deux configurations, les seuls liens conceptuels contenant 4 propriétés sont obtenus avec $\beta \in [0.03..0.09]$. En revanche, quand $\beta \in [0.27..0.3]$, seuls des *MFCLs* contenant 1 propriété sont extraits.

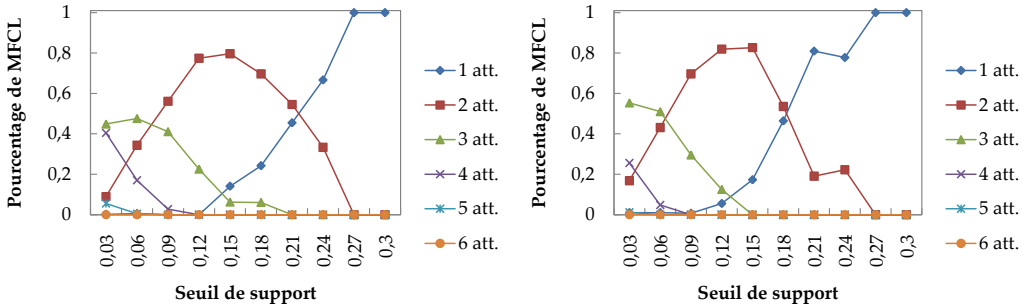


FIGURE 5.8 – Distribution de la précision des *MFCLs* selon le seuil de support (a) $|V| = 500$ et (b) $|V| = 3000$

Il est important de garder à l'esprit que quand le seuil de support augmente, seuls les groupes les plus connectés du réseau sont conservés. Cela est par exemple démontré par la Propriété 1. La réalité des comportements humains fait que ces groupes sont souvent ceux qui incluent une grande quantité de noeuds (c.-à-d. ceux qui sont décrits avec le moins d'items), ce qui peut expliquer ces résultats.

5.5.3 Étude quantitative

Dans l'étude quantitative, nous nous intéressons au nombre de *MFCLs* extraits, au temps d'exécution et au gain obtenu en comparaison de l'approche naïve.

Commençons par étudier comment évolue le nombre de motifs extraits selon différents seuils de support : (a) $\beta = 0.1$, (b) $\beta = 0.15$, (c) $\beta = 0.2$ et (d) $\beta = 0.25$. La Figure 5.9 présente ces résultats quand la taille du réseau et le nombre d'attributs évoluent.

Plusieurs observations intéressantes peuvent être faites. Tout d'abord, nous pouvons observer que pour toutes les configurations, le nombre de *MFCLs* croît avec le nombre de propriétés. Par exemple, pour $\beta = 0.1$ et $|V| = 3000$, le nombre de *MFCLs* extraits est d'environ 125 quand $|R| = 6$, contre 100 quand $|R| = 5$ (voir Figure 5.9(a)). Quand le nombre d'attributs augmente, le nombre de concepts potentiellement impliqués dans des *MFCLs* augmente également, générant ainsi une quantité plus importante de liens conceptuels.

Toutefois, nous observons que pour un nombre d'attributs donné, le nombre de *MFCLs* extraits reste relativement stable quand la taille du réseau varie. Ce résultat, par ailleurs très intéressant, peut s'expliquer par deux facteurs. (i) Le premier concerne la nature du jeu de données. En effet, de nombreux attributs sont binaires, et donc, lorsque l'on s'intéresse à un sous-ensemble des données, un sous-graphe dans notre contexte, la probabilité de retrouver les mêmes concepts avec différentes tailles de réseau est forte. (ii) Le second concerne les comportements humains en général. En effet, les facteurs sous-jacents qui créent ou influencent les comportements se retrouvent également à des échelles plus petites. En d'autres termes, cela signifie que si nous nous concentrons sur un sous-ensemble suffisamment pertinent, la distribution des données est telle, qu'il devient alors possible d'extraire une grande majorité des liens conceptuels fréquents.

Enfin, nous observons comme attendu que le nombre de *MFCLs* diminue quand le seuil β augmente. Par exemple, quand $|R| = 6$ et $|V| = 3000$, le nombre de *MFCLs*'élève à environ 125 pour $\beta = 0.1$ contre 50 pour $\beta = 0.15$ et 23 pour $\beta = 0.2$. C'est une propriété bien

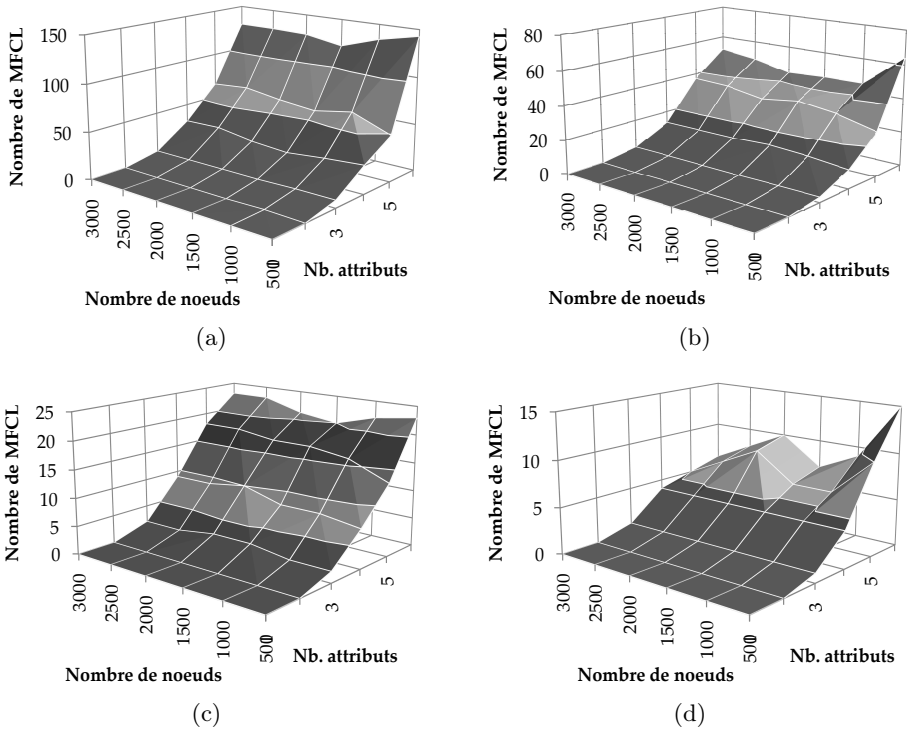
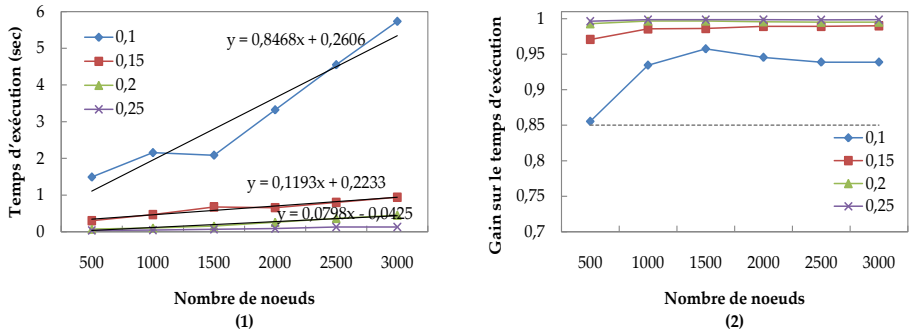


FIGURE 5.9 – Évolution du nombre de *MFCLs* selon $|V|$ et $|R|$
(a) $\beta = 0.1$, (b) $\beta = 0.15$, (c) $\beta = 0.2$ et (d) $\beta = 0.25$

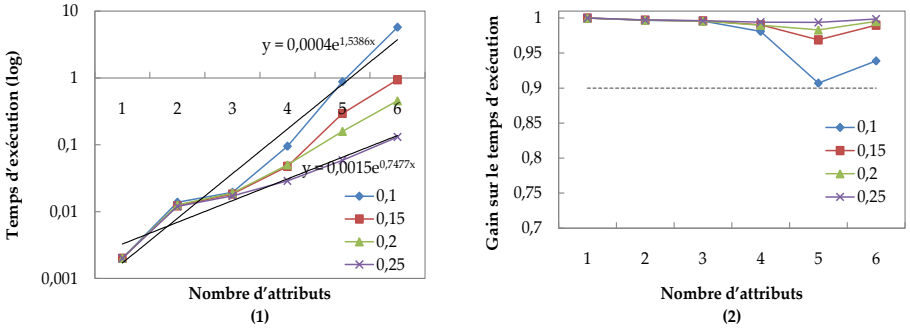
connue du domaine de la recherche de motifs fréquents, qui est due à la réduction de l'espace d'acceptabilité des solutions.

Pour compléter cette étude, nous nous sommes intéressés aux performances de la solution, en mesurant l'impact du nombre d'attributs et de la taille du réseau sur le temps d'exécution (en secondes) et le gain obtenu sur le temps d'exécution comparé à l'approche naïve. Sur la Figure 5.10, nous comparons les résultats obtenus avec différents seuils de support quand (a) la taille du réseau varie et $|R| = 6$ et (b) le nombre d'attributs varie et $|V| = 3000$.

En ce qui concerne le temps d'exécution (voir Figures 5.10(1)), nous observons globalement que le temps nécessaire à l'extraction des *MFCLs* croît quand β décroît. Par exemple, quand $|V| = 3000$, le temps de calcul est d'environ 5.8 sec. pour $\beta = 0.1$, alors qu'il est d'environ 1 sec. pour $\beta = 0.15$ (voir Figure 5.10(a)(1)). Cela est dû à l'algorithme *MFCL-Min*, qui est capable de réduire progressivement l'espace de recherche durant la phase d'extraction.
Cependant, deux observations doivent être faites : (i) le temps d'exécution croît linéairement avec la taille du réseau, (ii) alors qu'il croît exponentiellement avec le nombre d'attributs.
(i) En effet, comme le montre la Figure 5.10(a)(1), le temps requis par le processus d'extraction croît avec la taille du réseau quasi-linéairement (les équations associées sont affichées). Typiquement, quand $\beta = 0.1$, le temps de calcul peut être approché par $y = 0.8468 \times |V| + 0.2606$. Pour $\beta = 0.15$, il est approché par $y = 0.1193 \times |V| + 0.2233$. Cela peut s'expliquer à la fois par la nature du jeu de données, dans lequel beaucoup d'attributs sont binaires, mais également par l'optimisation apportée à *MFCL-Min* et présentée dans



(a)



(b)

FIGURE 5.10 – Performances de *MFCL-Min* selon temps d'exécution et gain (a) $|R| = 6$ et (b) $|V| = 3000$

l'Algorithme 8 qui évalue le support d'un lien conceptuel en parcourant uniquement les noeuds.
(ii) Cependant, comme on peut l'observer sur la Figure 5.10(b)(1), le *log.* du temps de calculs croît linéairement quand le nombre d'attributs augmente. Cela suggère que le temps d'exécution croît, lui, exponentiellement avec $|R|$. Par exemple, pour $\beta = 0.1$, le temps de calcul peut être approché par $y = 0.0004e^{1.5386\times|R|}$, contre $y = 0.0015e^{0.7477\times|R|}$ pour $\beta = 0.25$. D'une certaine façon, ces résultats confirment l'étude menée sur la complexité (voir Section 5.3) qui suggérait une croissance exponentielle du nombre de calculs effectués selon le nombre d'attributs.

Les résultats sur le gain obtenu comparé à l'approche naïve démontrent l'efficacité de la solution. Nous pouvons en effet observer que quelles que soient les variations opérées sur le nombre d'attributs ou la taille du réseau, le gain sur le temps de calcul est toujours supérieur à 85%, et ce, pour tous les seuils de support utilisés dans les expériences (voir Figures 5.10(2)). Là encore, ces résultats confirment l'étude menée sur la complexité et démontrent les bonnes performances de *MFCL-Min* dans l'extraction des liens conceptuels fréquents maximaux.

Il est important de préciser que les mêmes tendances ont été observées pour toutes les valeurs de $|R|$ et de V utilisées. Ainsi, en accord avec l'observation faite sur la croissance linéaire du temps d'exécution (voir Figure 5.10(a)(1)), nous avons étudié l'évolution de la pente décrivant la courbe du temps de calcul quand $|R|$ varie. Sur la Figure 5.11, nous montrons comment évolue cette pente (en logarithme).

Bien que nous ayons pu constater que le temps d'exécution croisse linéairement avec

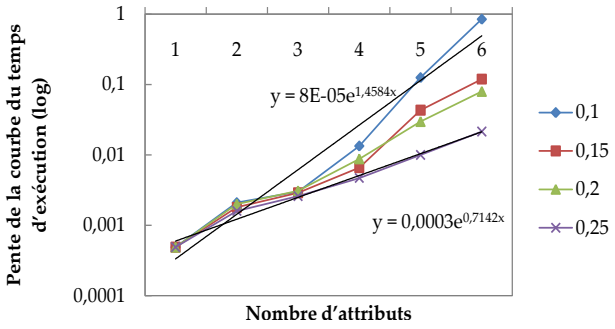


FIGURE 5.11 – Évolution de la pente (*log.*) de la courbe décrivant le temps de calcul

le nombre de noeuds, il est intéressant d'observer que la pente de la courbe du temps d'exécution croît, elle, de façon exponentielle avec le nombre d'attributs, comme le suggère la croissance plus ou moins linéaire du logarithme de la pente. Ce résultat vient confirmer les observations faites sur les Figures 5.10(1).
Le temps d'exécution y_β de *MFCL-Min* peut donc être approché par une fonction dépendant à la fois du nombre de noeuds $|V|$ et du nombre d'attributs $|R|$, c.-à-d. :

$$y_\beta \;=\; a \times e^{(b \times |R|)} \times |V| + c. \tag{5.18}$$

Par exemple, avec notre environnement de test, le temps de calcul pour $\beta = 0.25$ peut être approché par $y_{0.25} = 0.0003\ e^{0.7142 \times |R|} \times |V| + c$, c'est à dire environ 1 heure pour la configuration définie par $|V| = 10000$ et $|R| = 10$.
Bien évidemment, toutes ces estimations sont liées à la fois au jeu de données et à la configuration matérielle.

5.5.4 Vues conceptuelles : exemples et évolution

Pour conclure cette section sur les expériences, nous nous intéressons aux vues conceptuelles, en étudiant comment la structure de réseau sous-jacente évolue selon différents seuils de support β.

Nous montrons sur la Figure 5.12 quelques exemples de vues conceptuelles obtenues à partir (a) du réseau initial présenté sur la Figure 5.6, en utilisant les seuils (b) $\beta = 0.1$, (c) $\beta = 0.15$, (d) $\beta = 0.2$ et (e) $\beta = 0.25$. Sur les deux dernières représentations, le petit nombre de noeuds nous permet d'afficher les itemsets et les supports associés.

Les vues conceptuelles sont des représentations agrégées du réseau initial. Globalement, nous pouvons également observer que la taille du réseau décroît quand le seuil de support augmente. Sur la Figure 5.13, nous décrivons comment évoluent les principales caractéristiques des vues conceptuelles selon β : (a) nombre de noeuds et de liens, (b) densité et coeff. de clustering et (c) distribution du degré.

Comme attendu, le nombre de noeuds et de liens décroît quand le seuil augmente (voir Figure 5.13(a)). En effet, l'augmentation du seuil de support induit une réduction de l'espace d'acceptabilité des solutions ; moins de liens conceptuels sont donc ainsi identifiés comme fréquents.
Il est difficile d'expliquer l'allure générale de la courbe obtenue pour le coefficient de clustering (voir Figure 5.13(b)). Quand le support croît, seuls les groupes des noeuds qui possèdent de fortes connexions entre eux sont conservés. Il est possible qu'au sein d'un même groupe, les comportements conduisant à la formation de liens soient très similaires, renforçant ainsi

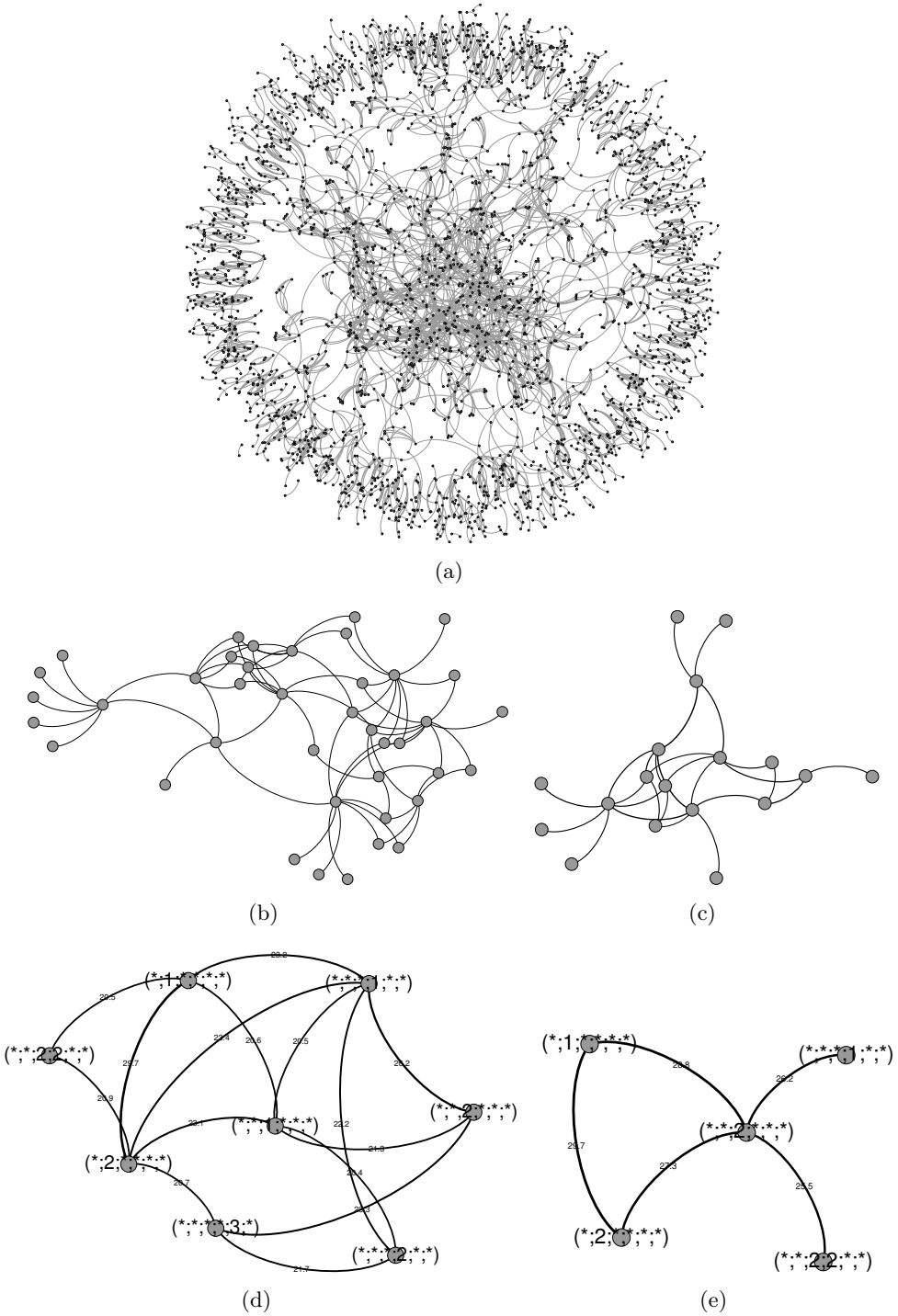


FIGURE 5.12 – Exemples de vues conceptuelles
(a) Réseau initial, (b) $\beta = 0.1$, (c) $\beta = 0.15$, (d) $\beta = 0.2$ et (e) $\beta = 0.25$

la *structure communautaire globale*. Ceci peut expliquer à la fois la croissance du coefficient de clustering, mais également celle de la densité.

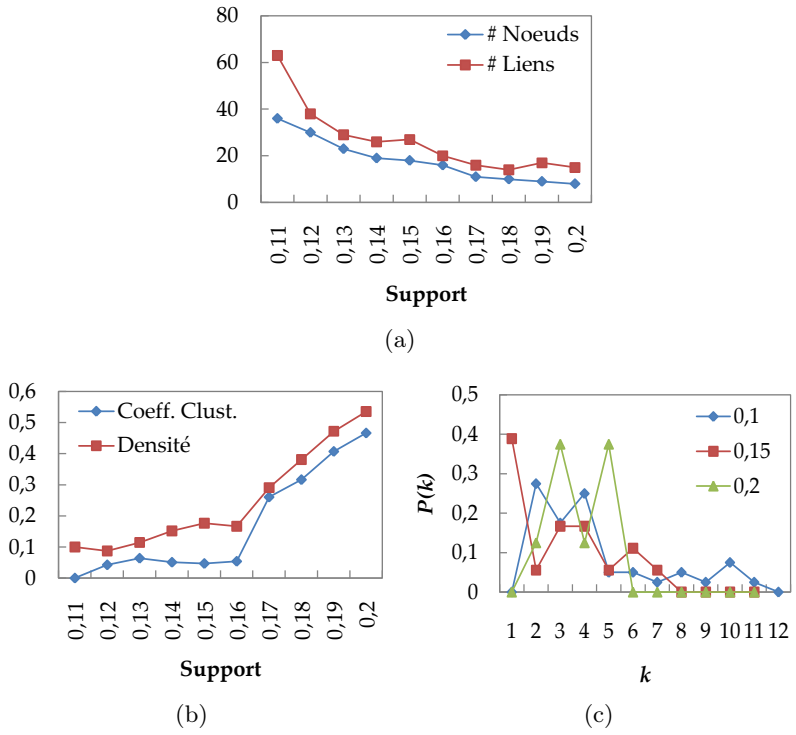


FIGURE 5.13 – Caractéristiques des vues conceptuelles selon le seuil de support (a) nombre de noeuds et de liens, (b) densité et coeff. de clustering, (c) distribution du degré

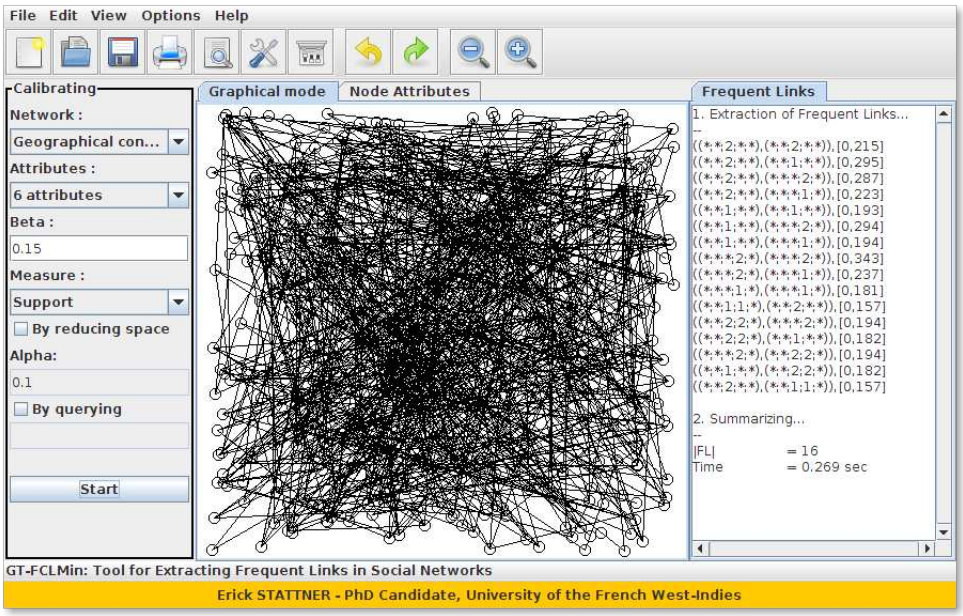
Enfin, en ce qui concerne la distribution des degrés (voir Figure 5.13(c)), nous observons globalement que les vues conceptuelles ont des propriétés fréquemment observées sur les réseaux du monde réel : une forte proportion de groupes possède un petit nombre de connexions, alors qu'un très faible pourcentage de groupes est très connecté.

5.6 L'outil *GT-FCLMin*

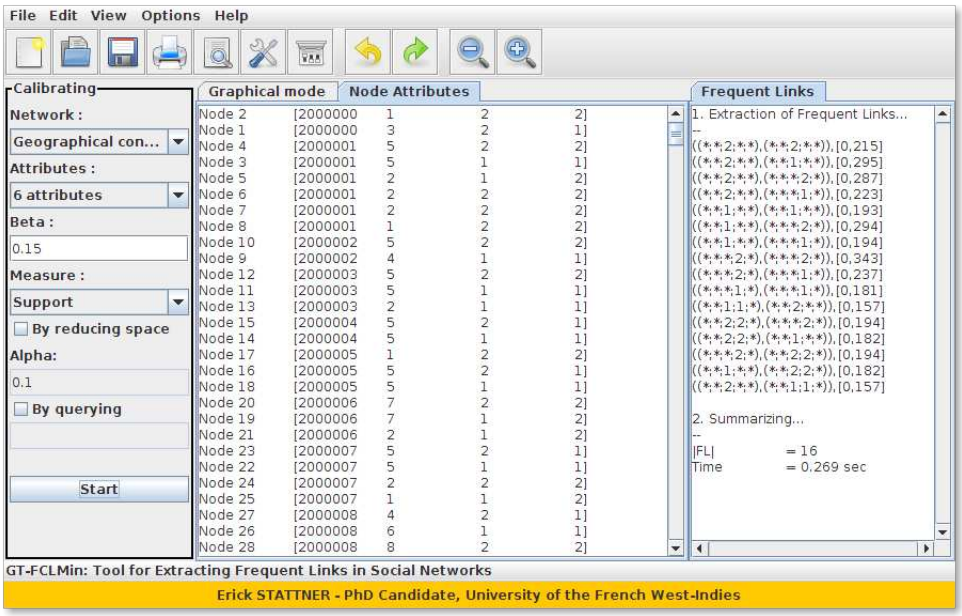
Dans le but de compléter la liste des outils actuellement disponibles pour l'extraction de connaissances au sein de réseaux sociaux, nous avons développé *GT-FCLMin*[1] (*Graphical Tool for Mining Frequent Conceptual Links*), un outil graphique dédié à l'analyse conceptuelle des réseaux sociaux. Notre objectif avec *GT-FCLMin* est de fournir un environnement d'analyse et de fouille de données simple, ergonomique et efficace, capable de prendre en compte tous les aspects de l'analyse conceptuelle des réseaux, et plus particulièrement l'étude et l'extraction des liens conceptuels et des vues associées. Tous les résultats présentés dans ce chapitre ont été obtenus avec *GT-FCLMin*.

Comme preuve de la flexibilité de l'outil, des évolutions majeures y ont déjà été apportées. Typiquement, dans sa première version [Stattner 2012b], *GT-FCLMin* s'intéressait à tous les liens conceptuels fréquents, ce qui générait inévitablement une redondance de certains des motifs extraits, qui pouvaient être contenus dans d'autres. Dans la version actuelle, l'outil est en mesure d'extraire les liens conceptuels fréquents maximaux [Stattner 2012c]

1. *GT-FCLMin* : http ://www.erickstattner.com/GT-FLMin



(a)



(b)

FIGURE 5.14 – Capture des deux principales vues de *GT-FCLMin*
Visualisation du réseau en (a) mode graphique et (b) mode textuel

et intègre la recherche des *MFCLs* basée sur la fréquence des itemsets [Stattner 2012h]. Il permet également à l'utilisateur de choisir entre plusieurs mesures d'intérêt et introduit un système de requête qui permet à l'utilisateur de rechercher les motifs liés uniquement à un groupe particulier de noeuds.

L'outil a été développé en JAVA et son interface graphique se compose de trois panneaux principaux comme le montre la Figure 5.14.

1. Le panneau de configuration, situé à gauche, permet à l'utilisateur d'effectuer les opérations de calibrage telles que le chargement du réseau (noeuds, liens et attributs), le réglage des différents paramètres ou le choix des éventuelles mesures d'intérêt à appliquer. Plus précisément, le réseau est chargé soit par la lecture d'un fichier au format UCINET [Borgatti 2002b] ou par le biais d'un simple fichier texte décrivant les liens du réseau sous la forme suivante :

$$<noeud\ 1> <noeud\ 2> <valeur>$$

Chaque ligne du fichier représente ainsi un lien entre deux noeuds. De la même façon, les attributs peuvent être lus à partir d'un fichier texte ayant la forme suivante :

$$<noeud\ 1> <attribut\ 1> <attribut\ 2> ... <attribut\ n>$$

Un filtre peut également être appliqué aux attributs, pour ne conserver qu'un certain nombre d'entre eux lors de l'analyse.
Une fois le réseau chargé, les paramètres liés à l'extraction des liens conceptuels peuvent être fixés : seuil de support minimum β et mesure d'intérêt à appliquer. L'outil permet également de rechercher des liens conceptuels fréquents entre des groupes eux-mêmes fréquents. Il est ainsi possible de réduire l'espace de recherche en fixant un seuil α définissant la taille minimale des groupes à considérer [Stattner 2012h].
Enfin, il peut être intéressant de rechercher les liens conceptuels liés uniquement à un groupe de noeuds. L'outil fournit ainsi un espace de requête permettant à l'utilisateur de saisir les groupes à cibler sous la forme d'un itemset. Typiquement, pour extraire tous les liens conceptuels impliquant les hommes de 40 ans, la requête peut être la suivante "$(40; 1; *; *; *; *)$", si le réseau est le même que celui utilisé dans nos expériences.
Une fois les opérations de calibrage effectuées, c'est ce panneau qui permet de lancer le processus d'extraction.

2. Le panneau de visualisation placé au centre, permet à l'utilisateur de visualiser et d'interagir avec l'application. Deux vues du réseau sont proposées : (i) Un mode graphique (voir Figure 5.14(a)), qui permet de visualiser le réseau en *2D*. Les noeuds affichés peuvent être déplacés à l'aide de la souris. (ii) Un mode textuel (voir Figure 5.14(b)) qui fournit une description plus détaillée du réseau : noeuds et propriétés, description des liens et diverses informations sur les caractéristiques du réseau (nombre de noeuds, de liens, degré min., degré moyen, degré max., coefficient de clustering, distribution du degré, etc.)
Le changement de vue s'effectue grâce aux onglets situés au-dessus du panneau.

3. Le panneau de résultats, situé à droite de la fenêtre, permet d'obtenir le résultat du processus d'extraction : liens conceptuels fréquents et support. Des informations résumant ce processus, telles que le nombre de liens conceptuels extraits et le temps d'exécution sont également fournis. En utilisant les options offertes par la barre de menu, ces informations peuvent ensuite être sauvegardées, ou utilisées pour générer et visualiser la vue conceptuelle. La vue conceptuelle peut, elle-même, être ensuite sauvegardée au format UCINET de façon à être manipulée par d'autres logiciels d'analyse de réseaux tels que UCINET [Borgatti 2002b], NetDraw [Borgatti 2002a] ou Gephi [Bastian 2009].

À terme, nous voulons intégrer à *GT-FCLMin* à la fois les méthodes d'analyse issues de l'analyse conceptuelle des réseaux sociaux [Riadh 2009, Le Grand 2009], mais également les méthodes issues de la fouille de liens [Fortunato 2009], l'objectif étant de mener l'analyse conceptuelle de pair avec l'analyse classique. Il serait par exemple intéressant d'appliquer

certaines techniques de fouille des réseaux sociaux, telles que la détection de communautés ou la prédiction de liens, aux vues conceptuelles.

D'un point de vue pratique, notre outil pourrait trouver des applications intéressantes dans divers domaines. Dans le domaine du marketing par exemple, appliquer *GT-FCLMin* à un réseau d'achats, dans lequel les individus sont connectés aux produits qu'ils achètent, pourrait permettre d'identifier les motifs *types d'utilisateur/produits* les plus fréquemment connectés.

5.7 Conclusion

Les techniques de data mining standard ont naturellement été appliquées et adaptées aux données sociales pour classifier, rechercher des motifs fréquents ou encore prédire l'apparition de liens. Pourtant, si des efforts ont été faits dans l'adaptation des méthodes existantes, nous constatons que la plupart de ces travaux exploitent uniquement la structure du réseau et ne permettent pas de répondre à certaines questions impliquant les propriétés des noeuds et portant sur les corrélations entre ces propriétés et l'appartenance d'un noeud à une communauté.

Dans ce chapitre, nous avons abordé la question de l'extraction de motifs fréquents en tenant compte à la fois des informations sur la structure du réseau et des informations disponibles sur les noeuds. Les contributions de ce chapitre peuvent être résumées comme suit :

(i) En étendant les notions issues de l'analyse de concepts formels, nous avons proposé une définition d'un motif, que nous appelons "***lien conceptuel***", qui combine ces deux types d'informations : structure du réseau et propriétés des noeuds. Nous avons montré que notre approche présente un intérêt à deux niveaux. D'une part, les motifs que nous proposons permettent d'obtenir une connaissance pertinente sur les connexions entre les groupes de noeuds (et donc les caractéristiques) les plus connectés du réseau. D'autre part, nous avons montré que la connaissance acquise peut être synthétisée à travers des "*vues conceptuelles*", des structures de graphe qui synthétisent la sémantique du réseau initial.

(ii) Nous avons proposé l'**algorithme *MFCL-Min*** pour l'extraction des liens conceptuels fréquents maximaux à partir de tout type de réseau social. L'algorithme effectue une recherche ascendante des motifs dans le treillis de concepts sociaux formés par les liens conceptuels, tout en réduisant progressivement l'espace de recherche. Les résultats expérimentaux obtenus à partir d'un réseau de contacts basé sur la fréquentation de lieux communs, ont permis d'observer d'une part la pertinence des motifs extraits et d'autre part les bonnes performances de l'algorithme comparé à une approche naïve.

(iii) Enfin, nous avons développé l'**outil graphique *GT-FCLMin*** qui implémente notre l'algorithme d'extraction *MFCL-Min* et fournit un environnement simple et ergonomique pour la recherche de liens conceptuels fréquents et la génération de vues conceptuelles. Cet outil peut être utilisé sur tout type de réseaux sociaux et intègre un ensemble de mesures d'intérêt pour évaluer la pertinence de la connaissance extraite.

Le travail présenté dans ce chapitre constitue une première étape dans l'exploitation des attributs des noeuds pour la recherche de motifs fréquents. D'une façon plus générale, nous pensons que la structure du réseau et les attributs des noeuds sont deux informations qui doivent systématique être prises en compte pour tirer pleinement parti de toutes les informations disponibles sur le réseau. En ce qui concerne notre approche, nous pouvons par exemple citer deux applications intéressantes.

Exemple 1. (Conception de stratégie marketing) Supposons que des données transactionnelles soient utilisées pour générer un réseau impliquant des utilisateurs et les produits

qu'ils achètent. La recherche de liens conceptuels dans un tel réseau peut s'avérer utile, car ils fournissent des informations pertinentes sur les liens fréquents entre les utilisateurs et les produits achetés, une connaissance qui peut ensuite être utilisée pour des campagnes de promotion ciblées.

Exemple 2. (Aménagement du territoire) Supposons qu'en utilisant les dispositifs de géo-localisation un réseau biparti entre les individus et les lieux qu'ils fréquentent soit construit. Les motifs extraits par notre approche pourraient fournir des informations précieuses pour un aménagement efficace du territoire tenant compte des visiteurs potentiels.

Chapitre 6

Réseaux de capteurs sans fil pour la collecte d'interactions sociales

Sommaire

Une problématique majeure des travaux menés sur les réseaux concerne la collecte des données représentant les interactions réelles. Du fait de la complexité de la tâche de collecte, de nombreux travaux se sont restreints à des jeux de données synthétiques. Dans les études décrites aux chapitres précédents par exemple, nous avons utilisé des jeux de données issus de modèles de génération, ou d'outils de simulation.
Dans le Chapitre 2, nous avons évoqué les perspectives nouvelles apportées par des dispositifs mobiles tels que les périphériques GPS, les téléphones mobiles ou les puces RFID, qui permettent le suivi en temps réel d'entités en mouvement et la collecte de données sociales autres que celles traditionnellement extraites à partir de sites d'échange et de partage.

Cependant, bien que les dispositifs mobiles présentent un grand nombre d'avantages dans la collecte d'interactions sociales, on observe que la mise en oeuvre de ce type de solutions soulève de nouvelles questions. C'est par exemple le cas de la collecte de données dans la nature, quand la végétation dense entrave le bon fonctionnement de ces dispositifs et impose des contraintes techniques et matérielles supplémentaires.

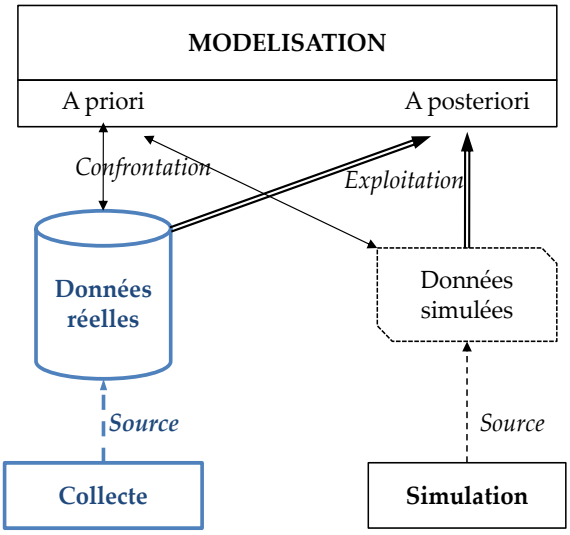


FIGURE 6.1 – Des données sociales aux modèles : collecte de données sociales réelles

Dans ce chapitre, nous abordons le problème de **la collecte de données sociales en situation réelle** (voir Figure 6.1). Nous proposons une alternative aux dispositifs mobiles à travers une architecture de collecte basée sur un réseau de capteurs sans fil équipés de périphérique pour collecter des signes d'interactions sociales au sein d'une population animale. Ce travail s'inscrit dans le cadre du projet CAP DOM [1], un programme de recherche qui vise à étudier les espèces d'oiseaux endémiques à la Martinique, et plus particulièrement une espèce en voie de disparition, le *Moqueur Gorge Blanche*, présent uniquement sur le site de la Caravelle (voir Figure 6.2).

Ce projet offre un cadre d'étude idéal pour proposer des méthodes automatiques de collecte de données sociales puisque :
(1) la collecte manuelle par des observateurs effectuant des relevés, méthode actuellement utilisée sur le site de la Caravelle, peut entrainer des modifications dans le comportement des animaux étudiés et donc biaiser les informations extraites. De plus, l'habitat des oiseaux est souvent hostile et rend l'observation difficile.
(2) les dispositifs mobiles, dont on pourrait équiper les oiseaux, sont exclus à cause de la végétation dense qui perturbe leur bon fonctionnement et de leur poids non-négligeable qui peut fortement handicaper ces oiseaux de petites tailles.

Dans la démarche d'amélioration des connaissances sur le mode de vie du Moqueur Gorge Blanche, un des objectifs est d'étudier les comportements sociaux de l'espèce. Les relations sociales maintenues, et déjà connues des scientifiques, sont les liens familiaux (couple, petits), les relations d'entraide entre individus de même famille, la recherche conjointe de nourriture ou la défense du territoire. La collecte et l'étude de données permettant de traduire ces liens s'inscrivent comme une étape importante pour la protection et la sauvegarde de l'espèce, et la surveillance des facteurs qui influencent leurs comportements et mettent l'espèce en danger.
Il s'agissait d'étudier la structure du réseau social maintenu par ces individus, ainsi que les groupes (ou communautés) qui pouvaient être identifiés.

Dans ce travail, nous avons proposé une architecture de collecte composée de capteurs

1. Site du projet : www.lifecapdom.org

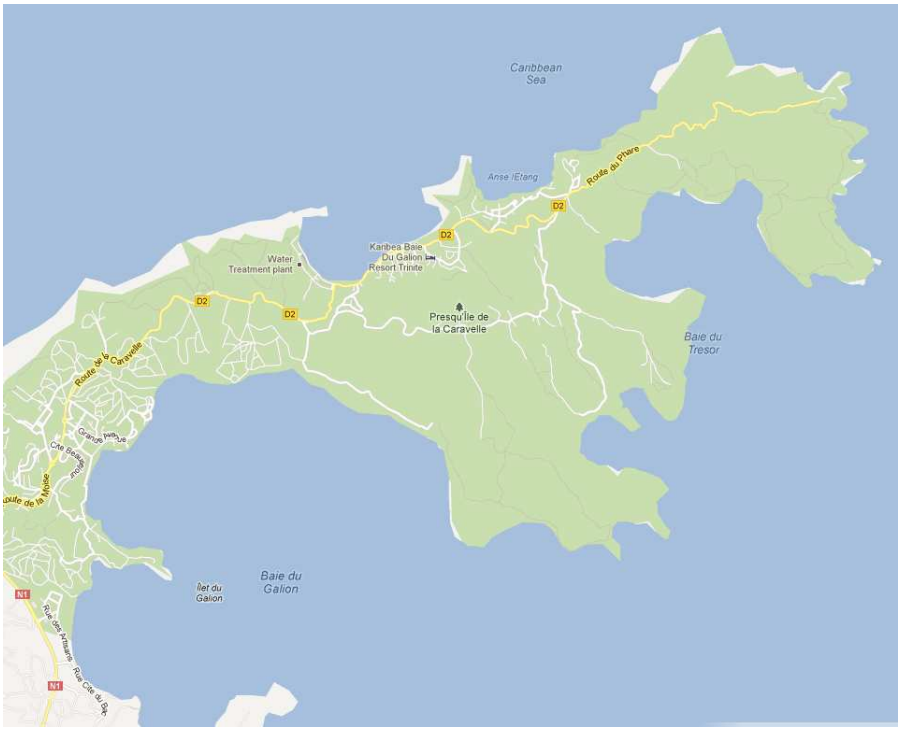


FIGURE 6.2 – Presqu'île de la Caravelle

sans fil fixes, dont l'objectif est de détecter des contacts de proximité entre les individus, d'en mesurer la fréquence et d'identifier les groupes familiaux. Comme les capteurs ne sont pas en mesure d'identifier avec précision la position des individus sur la zone d'étude, la construction du réseau social repose sur une discrétisation de l'espace en différentes régions et sur l'hypothèse que les individus peuvent être identifiés sans grande marge d'erreur sur la région sur laquelle ils sont présents.
Les capteurs permettent d'envisager différentes solutions de reconnaissance des individus, basées par exemple sur la capture de sons, d'images ou de vidéos. L'originalité de l'approche que nous avons choisie à été d'utiliser le chant des oiseaux pour identifier les individus et leur co-occurrence sur une région pour construire le graphe représentant le réseau social, pondéré par la fréquence des contacts de proximité spatiale. D'une part, le chant est utilisé comme témoin de la présence et de l'activité d'un individu sur une région. D'autre part, la co-occurrence de plusieurs individus sur une même région est utilisée pour la construction et le maintien des liens sociaux. Le réseau induit est analysé pour identifier les communautés, qui sont dans ce cas, les groupes familiaux.
Pour mener à bien ce travail, nous avons suivi la méthodologie proposée par Croft *et al.* [Croft 2008a], qui préconise d'organiser le processus de collecte en plusieurs étapes :

1. Formuler une question initiale
2. Déterminer une méthode pour identifier les individus
3. Définir une méthode d'enregistrement
4. Organiser les données
5. Construire et visualiser le réseau social

Dans l'état d'avancement actuel du projet, l'intérêt de ce travail est de fournir une

étude de faisabilité précédant une installation sur le terrain. Une évaluation de la solution proposée à partir de simulations s'avère essentielle avant un déploiement de matériels techniques onéreux. Dans cet objectif, la simulation présente en effet de nombreux avantages puisqu'elle permet d'isoler le système pour étudier les performances de la solution, comprendre comment évolue la qualité des données collectées selon différentes situations, comparer les effets de différents paramètres sur les performances, révéler des configurations qui maximisent les performances du système ou rechercher des compromis entre performances et coût d'installation. Cette phase préliminaire a fait l'objet des travaux présents dans ce chapitre.

Pour mesurer l'efficacité de l'approche et identifier les facteurs qui influencent les résultats avant un déploiement réel sur le terrain, nous avons conçu l'outil de simulation *Lypus*, qui reproduit un environnement virtuel sur lequel sont introduits des agents ayant des comportements similaires à ceux de l'espèce étudiée et un réseau de capteurs virtuel simulant le fonctionnement de la méthode de collecte proposée. Nos objectifs avec l'outil *Lypus* étaient, d'une part, de démontrer la faisabilité de l'approche, basée sur une discrétisation de l'espace, en vérifiant expérimentalement que cette approche est effectivement en mesure d'identifier les différentes familles d'individus, et, d'autre part, d'étudier la qualité des résultats dans différentes configurations de façon à optimiser le déploiement réel sur le terrain.

Ce chapitre est organisé comme suit. La Section 6.1 discute les principaux travaux menés sur la collecte d'interactions sociales chez les animaux et explique le choix effectué pour l'architecture de réseau de capteurs. La Section 6.2 est consacrée aux trois premiers points de la méthode de Croft *et al.* et la Section 6.3 s'intéresse aux deux derniers. La Section 6.4 présente *Lypus*, l'outil de simulation ainsi que les résultats obtenus. Enfin, la Section 6.5 conclut ce chapitre.

6.1 Collecte d'interactions sociales en milieu sauvage : un état de l'art

Collecter les interactions sociales entre des animaux nécessite de pouvoir mettre en place des stratégies de collecte définies en fonction du mode de vie connu ou soupçonné de l'espèce, du type d'habitat, de sa surface, du nombre d'individus à étudier, du type d'interactions envisagées, du temps et de la période d'étude, du coût de mise en place, etc. autant de paramètres qui rendent souvent difficiles de telles études sans une part d'automatisation.
Dans cette section, nous présentons les deux principales techniques de collecte utilisées dans le contexte des études menées sur les populations animales : (i) la méthode manuelle, qui se base sur l'observation et le relevé des comportements par des observateurs présents sur le terrain, et (ii) l'utilisation de dispositifs mobiles placés sur les individus et capables de collecter en temps réel des informations sur leur position.
L'analyse de ces deux approches nous permet de discuter les problématiques soulevées par l'utilisation de dispositifs fixes et de mettre en évidence les avantages et les faiblesses de chacune dans le contexte du projet CAP DOM.

La Section 6.1.1 est consacrée à la technique (i), actuellement utilisée sur la Presqu'île de la Caravelle. La Section 6.1.2 présente la technique (ii), la plus répandue dans les études menées sur les populations animales. La Section 6.1.3, s'intéresse au déploiement de dispositifs fixes. Enfin, la Section 6.1.4 détaille notre choix d'architecture.

6.1.1 Méthode manuelle d'observation

Les premières méthodes de collecte de données sur des populations animales consistaient à relever manuellement des informations sur les comportements des individus par le biais d'observateurs présents sur le terrain. Nous pouvons par exemple citer les travaux de Lusseau [Lusseau 2007], qui étudie les comportements d'une population de dauphins en Nouvelle-Zélande, ou ceux de Sade [Sade 1994], qui s'intéressent aux comportements des singes. Tous deux étudient les structures relationnelles de ces animaux à l'aide d'un réseau social construit sur l'observation des contacts sociaux entre les individus.
Cependant, dans certains cas, cette technique s'expose au problème de différenciation des individus. Il devient en effet difficile de garantir que deux individus observés à des instants différents sont les mêmes, ou pas. C'est à cet écueil que sont confrontés les spécialistes qui étudient le Moqueur Gorge Blanche.

Une solution proposée par Shorrocks et Croft [Shorrocks 2006] lors d'une étude menée sur des girafes consiste par exemple à différencier les individus en observant les marques naturelles présentent sur leur cou.
Cependant, dans le cas du Moqueur Gorge Blanche, l'identification visuelle d'un individu est rendue difficile pour différentes raisons : (1) La végétation dense, qui est un élément caractéristique de son habitat, rend difficile l'accès au site et la localisation des individus. (2) La législation très stricte sur cette espèce menacée exige qu'aucune modification ne soit apportée à son habitat (destruction de végétation, etc.). (3) La présence humaine elle-même peut perturber les animaux et modifier leur comportement, biaisant ainsi les informations collectées. (4) La durée des campagnes d'observation est limitée et dépend de la disponibilité des équipes. Or, il est important de pouvoir disposer de données suffisamment nombreuses et étalées sur de longues périodes.

6.1.2 Dispositifs mobiles

Les progrès techniques en matière de micro-technologies sont à l'origine de nouveaux types de dispositifs, capables de fournir en continu, et en temps réel, des informations spatio-temporelles. À partir du milieu des années 2000, la miniaturisation de plus en plus poussée de ces systèmes a permis leur utilisation dans la collecte automatique de données sociales. Des projets de recherche ont en effet su exploiter les nouvelles capacités offertes par ce que l'on nomme aujourd'hui les "*wearables sensors*" [Olguin 2008, Laibowitz 2006].

La technologie GPS est par exemple utilisée depuis plusieurs années pour suivre et identifier la position des animaux. Le problème d'identification des individus ne se pose pas puisque le capteur est porté par un individu et un seul.
Ce dispositif a pu être utilisé dans des situations où l'observation à grande échelle n'était pas envisageable. Par exemple pour étudier des animaux marins [Ryan 2004] ou en montagne pour suivre des élans [Rumble 2001]. Le réseau social construit à partir des données collectées est basé sur la proximité géographique des individus, un type de lien pertinent pour comprendre les structures communautaires.

Cependant, bien que le système GPS placé sur les animaux constitue actuellement la méthode la plus répandue, il n'était pas envisageable dans des contextes comme celui du Moqueur Gorge Blanche pour les raisons suivantes :

- Il ne fonctionne efficacement que dans des zones ouvertes et dégagées de tout objet pouvant obstruer les champs de vues du récepteur. Le fonctionnement sous un feuillage dense, qui constitue l'environnement du Moqueur Gorge Blanche, n'est pas possible.
- Il est particulièrement coûteux, car il faut capturer et équiper tous les individus.

- Il consomme énormément d'énergie.
- Il peut avoir un poids non-négligeable, ce qui ne permet pas son placement sur des oiseaux de petite taille, comme celui que l'on souhaite étudier.
- Ajoutons également qu'il ne permet d'étudier qu'un nombre limité de comportements, généralement liés aux interactions de proximité spatiale.

Ces inconvénients majeurs pour l'étude du Moqueur Gorge Blanche, nous ont donc orientés vers d'autres dispositifs de collecte que sont les capteurs fixes.

6.1.3 Capteurs fixes

Les capteurs fixes ont déjà été utilisés dans certaines études pour la collecte de données sociales. Chen *et al.* [Chen 2007a] ont par exemple mis en oeuvre des capteurs, équipés d'enregistreurs sonores et vidéos pour détecter et analyser les interactions sociales au sein d'une maison de retraite.
Selon les périphériques associés aux capteurs, ce type de dispositifs permet de détecter un nombre varié d'interactions à condition de savoir les identifier. Contrairement aux GPS, ces dispositifs sont fixes et ne nécessitent pas d'être installés sur les individus étudiés. Ils sont généralement installés au sol ou sur des équipements fixés au sol.

Cependant, l'utilisation de ce type de dispositifs nécessite de pouvoir identifier les individus et les interactions de façon précise. Typiquement, dans le cas d'enregistreur vidéo, des algorithmes d'analyse de séquence d'images doivent être mis en place pour permettre de reconnaître les individus ainsi que le type d'interaction. De plus, les données collectées sur les individus ne sont pas régulières dans le temps. Les individus ne sont détectés que quand ils sont dans le champ d'action des capteurs. Par exemple, le travail effectué par Chen *et al.* reste limité à un espace restreint et est facilité par la diversité des enregistreurs dont sont équipés les capteurs. Toutefois, sur des espaces beaucoup plus grands, il est nécessaire de placer un nombre important de capteurs de façon à détecter le maximum d'interactions. Ces dispositifs doivent également être en mesure d'identifier, de filtrer et de communiquer efficacement les informations qu'ils enregistrent à des systèmes centralisés, chargés de rassembler et de traiter les données.

6.1.4 Bilan

Dans le contexte de la collecte de données sociales sur le Moqueur Gorge Blanche, une méthode de collecte idéale doit pouvoir garantir certaines contraintes fonctionnelles majeures telles que le respect du comportement de l'animal et de son habitat, et donc la capacité à fonctionner sous un feuillage dense. Elle doit également pouvoir couvrir une zone de taille relativement importante, de l'ordre de plusieurs centaine de mètre, et être en mesure de détecter efficacement les individus et leurs interactions.
Sur le tableau de la Figure 6.3, nous comparons les trois familles de techniques face à un certain nombre de critères fonctionnels définis pour l'étude.

Les six premiers critères, incontournables pour le projet en cours, imposent d'exclure l'utilisation de dispositifs mobiles. En effet, bien que les dispositifs mobiles permettent d'effectuer la collecte sur de longue période et ne contribuent pas à la détérioration de l'habitat, ils nécessitent la capture et l'équipement de ces oiseaux de petite taille, ce qui peut d'une part engendrer des coûts de mise en place élevés et d'autre part handicaper les animaux dans leurs déplacements. De plus, le fonctionnement sous feuillage dense n'est pas garanti, ce qui peut nuire à la qualité des données collectées dans le contexte des études menées sur le Moqueur Gorge Blanche.

Contraintes fonctionnelles	Méthode manuelle	Dispositif mobile	Dispositif fixe
1. Durée d'étude (période)	Courte	Longue	Longue
2. Détérioration de l'habitat	Oui	Non	Non
3. Obligation de capturer	Non	Oui	Non
4. Coût financier	Moyen	Lourd	Moyen
5. Modification du comportement	Oui	Oui	Non
6. Fonctionnement sous feuillage dense	Non	Non	Oui
7. Identification des individus	Difficile	Facile	Moyen
8. Type d'interactions	Multiples	Limités	Multiples
9. Détection des interactions	Difficile	Facile	Moyen
10. Données régulières	Non	Oui	Non
11. Surface d'étude	Limitée	Grande	Variable

FIGURE 6.3 – Comparatif des trois techniques au regard des exigences fonctionnelles

Ainsi, les difficultés majeures liées à l'utilisation de dispositifs mobiles dans l'habitat du Moqueur Gorge Blanche nous ont orientés vers l'utilisation de capteurs fixes. En plus de leur conformité face aux exigences fonctionnelles majeures du projet (les six premières), un des avantages des capteurs est celui de pouvoir collecter plusieurs sources de données (sons, images ou vidéos), permettant ainsi d'envisager divers types de méthodes pour la reconnaissance des individus et de leurs interactions.
Cependant, les capteurs présentent l'inconvénient de ne pas fournir de données à intervalles de temps régulier puisque les individus ne sont détectés que quand ils sont dans le champ de détection d'un capteur. Ainsi, la quantité de données dépend essentiellement de la surface de la zone d'étude couverte par les dispositifs.

6.2 Stratégie de collecte

Toute étude du comportement social passe nécessairement dans un premier temps par une méthode de collecte efficace des interactions entre les individus. Une fois ces informations collectées, elles sont ensuite traitées puis analysées pour extraire des connaissances. Cette section est consacrée à la stratégie de collecte que nous proposons et répond aux trois premiers points de la méthode préconisée par Croft *et al.* [Croft 2008a] : (1) la question initiale, (2) la méthode d'identification des individus et (3) le type d'interactions à étudier.

Dans la Section 6.2.1 nous exposons les objectifs et discutons du type d'interactions à collecter ainsi que des difficultés à les identifier. La Section 6.2.2 détaille l'organisation de l'architecture de capteurs que nous proposons. Enfin, la Section 6.2.3 présente la solution d'identification des individus choisie et montre comment elle est mise en place sur les capteurs.

6.2.1 Question initiale

Les travaux préliminaires menés sur l'espèce des Moqueurs Gorge Blanche ont permis d'établir que les individus maintiennent des structures familiales fortes, au sein desquelles un couple d'oiseaux occupe un territoire défini et s'occupe de sa progéniture pendant un certain temps. L'objectif de ce travail est d'identifier ces *structures familiales* ainsi que leur

évolution.

Une façon simple d'inférer la structure de ces familles consiste à s'intéresser aux comportements des individus.
Par exemple, un couple d'individus a pour caractéristique commune de partager un nid et de participer à la défense du territoire. De même, les jeunes peuvent être identifiés s'ils partagent un nid avec des adultes, ou s'ils sont assistés dans la recherche de nourriture par un des parents.
Ce type de comportement reste toutefois assez difficile à identifier dans un contexte réel. Par exemple, identifier les individus qui partagent un même nid peut s'avérer particulièrement difficile, puisqu'il faut pouvoir identifier le nid ainsi que les individus qui le visitent.

Cependant, nous observons que les comportements permettant d'identifier les liens sociaux pertinents pour l'extraction des communautés sont souvent des comportements qui nécessitent une proximité spatiale des individus : partage d'un même nid, entraide dans la recherche de nourriture, etc.

Dans ce travail, nous faisons l'hypothèse que la proximité spatiale est une condition non seulement nécessaire, mais aussi suffisante pour traduire le lien familial. La fréquence des co-occurrences de deux individus sur une même zone doit être prise en compte.
La difficulté vient cependant du fait que des individus peuvent être géographiquement proches sans nécessairement faire partie de la même famille. Le cas typique est par exemple celui où deux individus, possédant des territoires adjacents, sont à la recherche de nourriture à la limite de leur territoire respectif au même moment. Dans ce travail, nous supposons que ce type d'interaction survient statistiquement assez rarement. Plus précisément, si nous considérons la fréquence des interactions basées sur la proximité géographique des individus, nous supposons que la fréquence des *interactions intra-famille* est beaucoup plus élevée que la fréquence des *interactions inter-familles*.
Les groupes à découvrir dans le réseau social sont donc les familles, qui en termes de réseau se traduisent en communautés dans la mesure où le lien social qui définit le réseau est celui de la proximité géographique. L'architecture que nous proposons s'intéresse donc uniquement à la collecte des interactions de proximité pour l'identification des communautés qui représentent ici les groupes familiaux.

6.2.2 Architecture de collecte

En plus des contraintes fonctionnelles présentées précédemment, notre choix d'architecture, à savoir un réseau de capteurs sans fil hiérarchiquement organisé sur une surface découpée en régions, nous oblige également à considérer certaines contraintes techniques : la limitation des échanges sur le réseau de capteurs sans fil, de la consommation énergétique et du nombre de dispositifs nécessaires. Nous commençons par présenter l'architecture et montrons ensuite en quoi elle répond aux différentes contraintes.

Les architectures classiques de réseau de capteurs pour la surveillance de l'habitat sont généralement divisées en trois niveaux distincts [Martincic 2005].
(i) Le niveau le plus bas se compose de capteurs autonomes, équipés de divers types de dispositifs chargés d'effectuer les tâches de collecte d'informations ou de détection d'individus. On parle de "*capteurs de détection*". Ces capteurs sont organisés en sous-réseaux appelés *clusters*.
(ii) Au sein de chaque cluster, l'un des capteurs joue le rôle de passerelle afin de recueillir les informations reçues des capteurs de détection avant un envoi vers un système centralisé. On parle de "*capteur-passerelle*". Le capteur-passerelle a souvent des capacités en calcul et en mémoire plus importantes que les autres. Il est également équipé d'une antenne capable

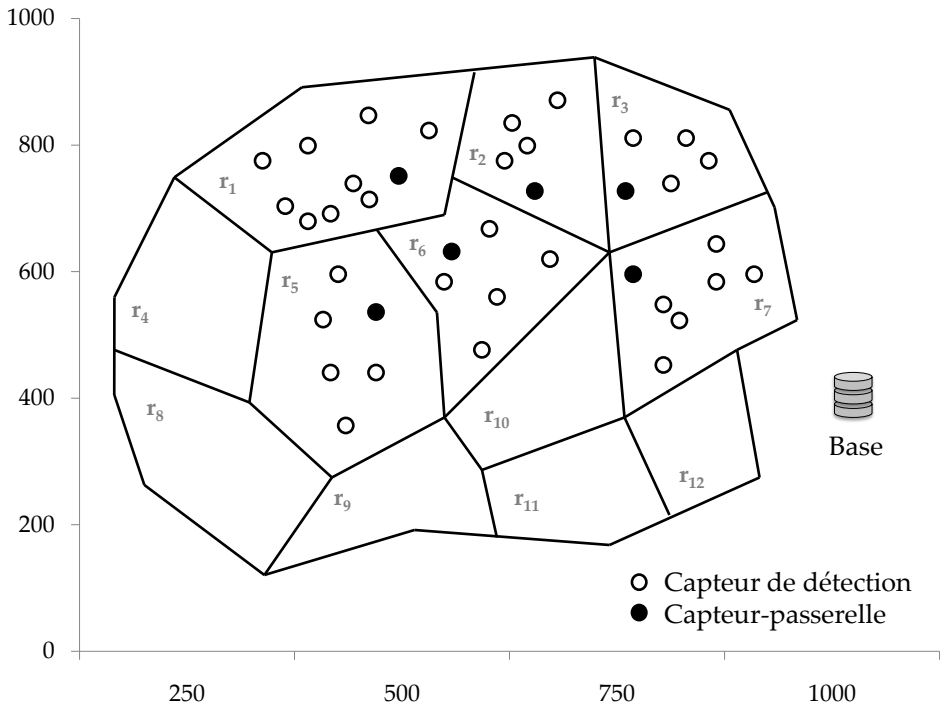


FIGURE 6.4 – Architecture de capteurs sans fils pour l'observation des Moqueurs Gorge Blanche

de transmettre des données sur de longues distances.
(iii) Enfin, une *station de base*, ou *base centrale*, agrège et traite les données reçues.

Comme l'illustre la Figure 6.4, l'architecture que nous proposons intègre ce concept de hiérarchie pour organiser les capteurs sans fil.

Premièrement, nous supposons que la zone d'étude est discrétisée en sous-zones, que nous appelons des "*régions*" et qui peuvent être délimitées par les spécialistes du domaine ornithologique.
Ce découpage peut tenir compte des caractéristiques de l'environnement pour permettre par la suite de mettre en corrélation le comportement social des individus et les caractéristiques associées à la région. Cependant, cette délimitation n'est pas toujours possible, ce qui peut parfois nous obliger à définir les régions de façon arbitraire.
Chaque région possède un identifiant unique, par exemple r_1, r_2, r_3, r_4, etc. (voir Figure 6.4).

Des capteurs de détection sont placés sur la zone d'étude et sont affiliés à la région à laquelle ils appartiennent. Par exemple, quand un capteur est placé à l'intérieur des limites de la région r_1, il est affilié à cette région. Cependant, bien que le rayon de détection du microphone d'un capteur puisse chevaucher plusieurs régions, un capteur n'est, lui, affilié qu'à une et une seule région. Les capteurs de détection sont capables de détecter la présence d'un individu en analysant les chants enregistrés.

Enfin, chaque région possède un capteur-passerelle qui collecte les informations reçues des capteurs de détection de sa région, les traite et les envoie à une base centrale, chargée de collecter les informations reçues de tous les capteurs-passerelle.

Ce système de communication en deux couches évite que tous les capteurs envoient, de façon individuelle, des informations potentiellement redondantes à la base centrale. En

effet, les informations sont traitées et triées par les capteurs-passerelle avant d'être envoyées à la base centrale. Cela a pour conséquence de limiter considérablement les échanges sur le réseau. En effet, nous savons que la zone à analyser peut être très importante. La limitation des informations transitant sur le réseau est donc un aspect important de la méthode. L'infrastructure logique de communications est présentée sur la figure 6.5. On peut y distinguer une organisation en trois niveaux :

1. Les capteurs de détection communiquent avec les capteurs-passerelle
2. Les capteurs-passerelle communiquent avec la base centrale
3. La base centrale collecte, traite et stocke les données

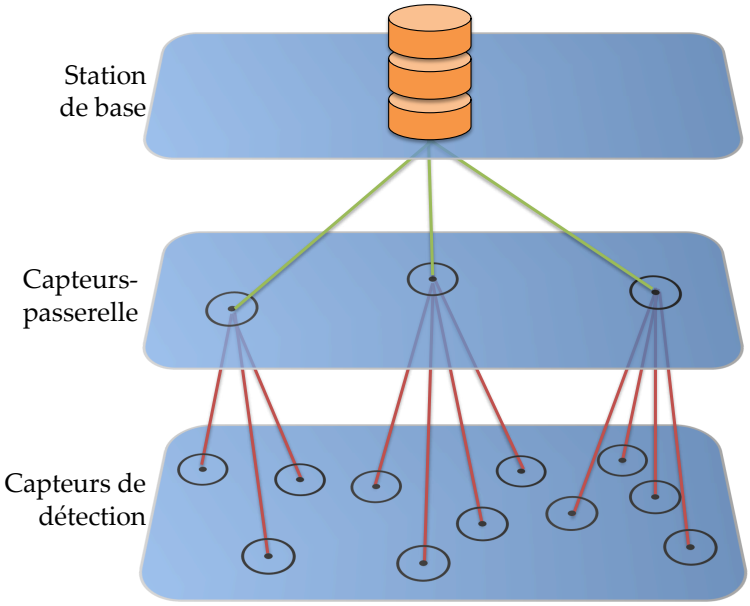


FIGURE 6.5 – Infrastructure de communications

Enfin, nous supposons que tous les capteurs de détection possèdent les mêmes capacités. Les microphones possèdent des rayons de détection équivalents, les capteurs ont la même durée de vie et connaissent la région à laquelle ils sont affiliés ainsi que le capteur-passerelle de leur région.

L'architecture que nous proposons répond aux exigences fonctionnelles et techniques du projet. Elle permet d'étudier les oiseaux sans modifier ni leur habitat, ni leur vie quotidienne en les encombrant avec des dispositifs trop gênants comme le feraient des colliers GPS. Par ailleurs, bien positionné, un même capteur peut permettre d'obtenir des données sur plusieurs individus, ce qui n'est pas permis avec un système GPS qui n'est, lui, associé qu'à un seul individu.

En ce qui concerne la limitation des échanges sur le réseau, nous avons pu observer que l'infrastructure logique de communication, organisée en deux couches, permet d'agréger, de filtrer et de traiter dans un premier temps les informations au niveau des capteurs-passerelle, avant un envoi vers la base centrale. Ce premier traitement permet de limiter les échanges sur le réseau, en n'autorisant pas une communication directe entre les capteurs de détection et la station de base. Ce système permet également de se prémunir de l'envoi d'informations redondantes à la base.

En ce qui concerne la consommation énergétique, deux facteurs concourent au maintien d'un cycle de fonctionnement optimal. Tout d'abord, les capteurs sont équipés de panneau solaire et d'une batterie rechargeable afin d'être en mesure de fonctionner sur de très longues périodes. De plus, la limitation des échanges sur le réseau réduit un peu plus la quantité d'énergie nécessaire pour acheminer les informations à la base centrale.

Enfin, le nombre de capteurs installés croît avec la surface de la zone d'étude à couvrir. La solution permet de maîtriser la quantité de dispositifs de façon à couvrir plus densément les zones très fréquentées, si cette information est connue. Cependant, une difficulté qui se pose avec cette architecture concerne le placement des capteurs, c'est-à-dire la façon de positionner les capteurs les uns par rapport aux autres. En effet, bien qu'une seule région soit affectée à un capteur, en fonction de la technique de placement choisie, on peut être confronté à deux types de phénomènes : soit des zones ne sont pas couvertes, soit le rayon de détection d'un capteur recouvre plusieurs régions. Les trois types de positionnement les plus utilisés sont :

Le positionnement uniforme. Cette configuration consiste à disposer uniformément les capteurs de façon à ce qu'ils couvrent toute la zone. Cette configuration maximise les chances de détection, et dans notre contexte peut être utilisée pour diminuer le phénomène de chevauchement des rayons de détection sur plusieurs régions.

Le positionnement dense. Ce type de placement consiste à placer un maximum de capteurs pour couvrir une large surface. Cette configuration pose naturellement des problèmes de coût et d'interférences entre les signaux.

Le positionnement aléatoire. Cette méthode consiste à disposer les capteurs de façon aléatoire sur la zone.

Dans l'étude menée à la Section 6.4, nous étudions l'effet de différentes configurations de placement sur les résultats.

6.2.3 Identification des individus

Un des avantages du système de collecte que nous proposons est de pouvoir configurer les capteurs de manière à disposer de différentes solutions d'identification des individus. En effet, en fonction des périphériques connectés aux capteurs, plusieurs méthodes de reconnaissance peuvent être envisagées, basées par exemple sur la capture d'images, de sons, de vidéos ou une combinaison de plusieurs sources. Quelle que soit la méthode de reconnaissance considérée, la construction du réseau social est basée sur le principe qu'**un individu, et la région sur laquelle il est présent peuvent être identifiés de manière unique**.

Cette première approche, se restreint aux situations dans lesquelles l'élément sonore, le chant, est le principal témoin de la présence et de l'activité d'un individu.
La méthode de reconnaissance que nous avons expérimentée est basée sur l'hypothèse que les individus d'une même espèce peuvent être identifiés en analysant leur chant. En effet, bien qu'une même espèce d'oiseaux possède un chant caractéristique, plusieurs travaux ont montré que chaque individu a des particularités propres dans son chant [Beletsky 1982, Phillmore 2002]. Le chant est d'ailleurs un signe de reconnaissance entre les individus eux-mêmes. Cette caractéristique a été mise en évidence par Laurent Dabouineau [Dabouineau 2004] qui a analysé les chants de plusieurs individus de Troglodyte et a montré que chaque individu possède des caractéristiques spécifiques dans son chant qui permettent de l'identifier. Des études similaires menées sur des Chouettes Hulottes [Galeottia 1991] ont également permis de vérifier cette propriété.
Les connaissances expertes des géographes et des ornithologues nous autorisent à penser que la même caractéristique existe chez les Moqueurs Gorge Blanche.

La méthode de reconnaissance que nous proposons s'effectue en deux étapes : (1) la première étape vise à déterminer si le chant enregistré correspond à celui de l'espèce étudiée et (2) la seconde étape a pour objectif d'identifier l'individu concerné.
Ces deux tâches sont effectuées en utilisant les techniques de *paramétrisation*, qui produisent une empreinte caractéristique à partir du signal sonore. Cette empreinte est ensuite utilisée pour identifier à la fois l'espèce et l'individu. Cette solution suppose vraie l'hypothèse d'un degré élevé de variations inter-individus et d'un faible niveau de variations pour un individu donné, dans les empreintes associées aux chants. En effet, malgré les très légères variations qui peuvent survenir sur le chant d'un même individu (selon le matériel, l'environnement, le contexte, etc.), les empreintes tirées de ses différents chants devraient être très proches, permettant ainsi de le reconnaitre.

Plus précisément, la paramétrisation est une technique qui permet de transformer un signal sonore en une représentation caractéristique, c'est-à-dire un ensemble de coefficients décrivant le signal à intervalle de temps régulier. Cette représentation permet de réduire l'information en quantité et en redondance. On parle d'*empreinte caractéristique*, ou de *traits caractéristiques*. Ces représentations sont classiquement utilisées chez les humains, en reconnaissance de la parole et pour identifier le locuteur [Mariani 2002, Levy 2006, Bredin 2006].

Il existe différentes techniques de paramétrisation telles que la méthode basée sur les MFCC (Mel-frequency cepstral coefficients) ou celle basée sur LPC (Linear predictive coding). Dans ce travail, nous utilisons la paramétrisation basée sur les MFCC, classiquement utilisés en traitement de la parole pour identifier à la fois les mots mais également le locuteur [Bredin 2006]. L'étude menée par Christophe Levy [Levy 2006] en comparant plusieurs techniques de paramétrisation pour la reconnaissance de locuteurs a également montré que la paramétrisation basée sur les MFCC présente de bons résultats sur des systèmes à capacités restreintes tels que les téléphones portables ou les capteurs. Dans cette approche, les MFCC sont utilisés car l'analyse est limitée au chant de l'oiseaux et les capacités des dispositifs sont restreintes.

Comme l'illustre la Figure 6.6, le processus d'identification peut être décrit en trois étapes selon le niveau :

(1) Le capteur de détection. Les capteurs de détection sont capables d'enregistrer les sons dans leur environnement et de déterminer s'il s'agit d'un chant de Moqueur Gorge Blanche. Plus précisément, quand un capteur enregistre un son, le signal est paramétrisé afin d'obtenir une empreinte représentative du signal. Cette empreinte est ensuite analysée par le capteur pour déterminer s'il s'agit d'un chant de l'espèce recherchée ou pas. Pour optimiser le traitement localement, nous avons opté pour une solution simple, qui consiste à comparer une empreinte donnée à une *empreinte moyenne caractéristique* stockée localement sur chaque capteur de détection. Cette empreinte moyenne est générée avant déploiement, à partir d'un grand nombre d'empreintes de différents individus connus de l'espèce. Ainsi, en fonction d'un seuil nous considérons que si l'empreinte diffère trop de l'empreinte moyenne, le son enregistré ne correspond pas à celui d'un chant de l'espèce ; sinon, nous admettons qu'il s'agit d'un chant émis par un individu de l'espèce et l'empreinte est transmise au capteur-passerelle.
Ce premier traitement effectué localement permet d'éviter d'envoyer au capteur-passerelle des empreintes générées par des sons parasites, produits par exemple par des marcheurs, le vent, la pluie ou même d'autres espèces animales.

(2) Le capteur-passerelle. À chaque instant, les passerelles reçoivent les empreintes provenant des capteurs de détection de leur région. À ce niveau, les empreintes reçues ne correspondent qu'à des individus supposés de l'espèce. Cependant, comme plusieurs

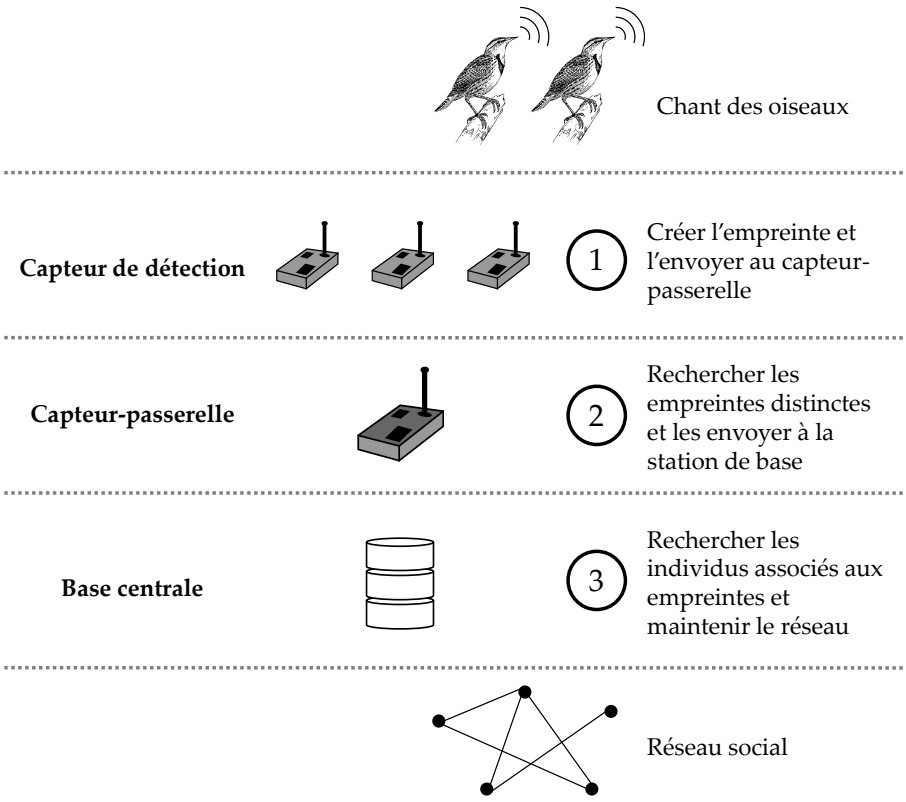


FIGURE 6.6 – Système de collecte des interactions

capteurs peuvent avoir "entendu" le même oiseau, les capteurs-passerelle effectuent une comparaison des empreintes pour ne conserver que les empreintes distinctes. Une fois cette étape achevée, les passerelles envoient les empreintes distinctes à la station de base.

L'agrégation et le traitement des données au niveau des capteurs-passerelle permettent de limiter un peu plus les informations transitant sur le réseau.

(3) La station de base. La base centrale maintient une base de données dans laquelle chaque empreinte est associée à un individu. Quand la base centrale reçoit une empreinte d'un capteur-passerelle, il y a deux possibilités. Soit cette empreinte n'est pas encore répertoriée dans la base de données, il est supposé alors qu'il s'agit donc d'un nouvel individu. Dans ce cas, l'empreinte est ajoutée à la base et un identifiant lui est affecté. Soit l'individu est déjà présent dans la base, l'identifiant associé à l'empreinte peut donc être recherché.

Ainsi, après la réception des empreintes des individus ayant chanté à un instant donné, la station de base est en possession des identifiants de ces individus et de la région sur laquelle ils ont été détectés. Ces données sont ensuite utilisées pour construire et maintenir un réseau social comme nous l'expliquons dans la Section 6.3.

Les individus ne sont identifiés dans ce contexte que par leur chant. Il nous a été difficile de vérifier cette propriété chez les Moqueurs Gorge Blanche, en raison du peu d'échantillons dont nous disposions. Dans un contexte idéal, il faudrait pouvoir disposer d'un grand nombre de chants, et pour chacun d'entre eux, pouvoir déterminer avec précision l'individu associé. Ces données sont évidemment difficiles à obtenir dans la

réalité, et particulièrement dans le contexte des oiseaux, pour lesquels les enregistrements sonores sont souvent effectués sans nécessairement qu'il y ait d'observations visuelles associées. Il devient donc très difficile, même pour des spécialistes, d'identifier les différents individus uniquement sur la base d'une écoute humaine.

Pour nos tests, nous disposions de 7 échantillons, prélevés entre septembre et octobre 2010 sur le site la Caravelle. Chaque enregistrement a été écouté par un expert qui a identifié 3 chants comme étant ceux d'un individu B_1 et les 4 autres comme étant ceux d'un second individu B_2. Deux types de tests ont été menés : (i) la reconnaissance de l'espèce et (ii) l'identification des individus. Les empreintes ont été générées en utilisant la bibliothèque JAVA *Comirva*, développée par Schedl *et al.* [Schedl 2007].

(i) En ce qui concerne la reconnaissance de l'espèce, *une empreinte moyenne caractéristique* a été générée à partir des 7 échantillons dont nous disposions. En comparant les empreintes associées aux chants d'autres espèces d'oiseaux et celle des 7 échantillons à l'empreinte moyenne, le chant du Moqueur Gorge Blanche a pu être identifié avec un taux de succès de 100%. Cela confirme la possibilité d'identifier un son comme étant un chant de l'espèce en s'intéressant aux empreintes associées.

(ii) En ce qui concerne l'identification des individus, l'un des chants identifié comme étant celui de l'individu B_2 n'a été reconnu ni comme appartenant à B_2, ni même à B_1, laissant ainsi supposer qu'il s'agissait d'un troisième individu. Il est difficile de déterminer précisément s'il s'agit d'un faux négatif ou d'un biais lié à l'écoute de l'expert. Nous pouvons cependant souligner que cette première étude semble confirmer l'hypothèse de départ, à savoir que les empreintes associées aux chants peuvent être utilisées pour distinguer les individus de Moqueur Gorge Blanche.

La Figure 6.7 illustre un exemple d'empreintes obtenues à partir de deux individus de Moqueurs Gorge Blanche. Sur la Figure 6.7(a), on peut facilement observer les fortes variations qui existent entre une empreinte de Moqueur Gorge Blanche et celle d'une autre espèce. On peut également observer sur les Figures 6.7(b) et (c) la forme générale de l'empreinte qui est caractéristique de l'espèce. Enfin, nous notons les légères variations obtenues dans les empreintes du même individu (voir Figure 6.7(b)) et les variations plus marquées quand il s'agit de deux individus différents (voir Figure 6.7(c)).

Dans la réalité, il est évident que des difficultés apparaissent. La qualité des enregistrements peut notamment être médiocre et introduire des imprécisions sur le processus d'identifications des individus.

Dans cette section, nous avons expérimenté une solution de reconnaissance des individus basée sur l'analyse des chants. En raison des difficultés rencontrées dans l'obtention d'échantillons réels fiables, nous avons posé comme axiome qu'une telle identification est possible. Nous rappelons cependant que l'approche que nous proposons pour la génération du réseau social est assez souple pour supporter différents types de méthodes de reconnaissance (son, image ou vidéo). Elle nécessite uniquement qu'un individu unique, et la région sur laquelle il est présent, puissent être identifiés.

6.3 Réseau social

Le réseau de capteurs permet la collecte de données à partir desquelles le réseau social peut être représenté par un graphe. En fonction des informations reçues, la base centrale construit et maintient ce réseau social, basé sur la fréquence des interactions de proximité entre les individus.

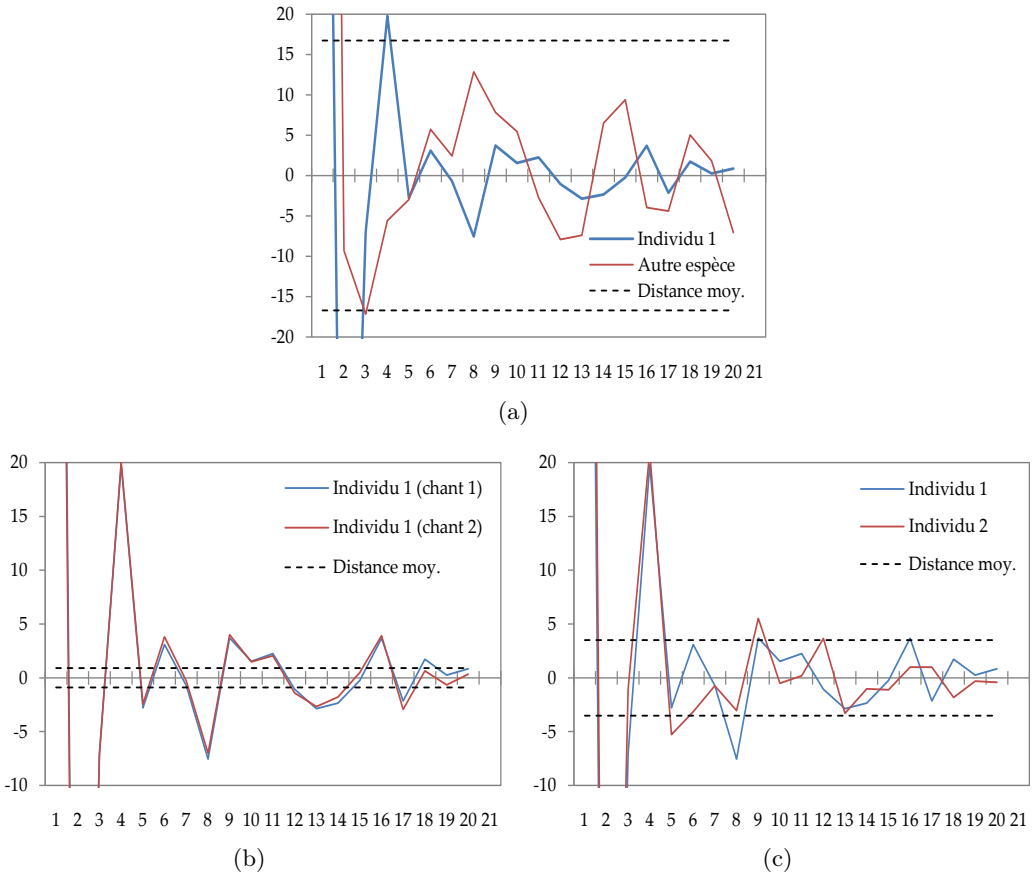


FIGURE 6.7 – Comparaison d'empreintes d'individus de Moqueurs Gorge Blanche (a) chants d'un individu et celui d'une autre espèce, (b) deux chants du même individu, (c) chants de deux individus différents

Cette section est consacrée aux deux derniers points de la méthode de Croft *et al.* La Section 6.3.1 présente l'organisation des données et la construction du réseau. La Section 6.3.2 s'interesse à la visualisation du réseau.

6.3.1 Organisation et construction

Soient $R = \{r_1, ..., r_n\}$ l'ensemble des régions qui divisent la zone d'étude, et $B = \{b_1, ..., b_m\}$ l'ensemble des oiseaux détectés. L'ensemble des oiseaux B, est alimenté progressivement par la base centrale, avec les nouveaux individus détectés.

Nous notons $T = \langle t_1, ..., t_k \rangle$, la séquence de temps sur laquelle les interactions sociales sont collectées.
Nous notons également $F_t^r : B \to \{0, 1\}$, la fonction de détection qui renvoie, pour un élément $b \in B$, 1 si b a été détecté sur la région r à l'instant t, et 0 dans le cas contraire.

Enfin, nous définissons $D_{i,j}$ comme l'*ensemble de détection instantané*, qui contient la liste de tous les oiseaux qui ont été détectés sur la région $r_i \in R$ à l'instant $t_j \in T$.

$$\forall r_i \in R,\ \forall t_j \in T \quad D_{i,j} = \{b \in B\ ;\ F_{t_j}^{r_i}(b) = 1\} \tag{6.1}$$

De cette façon, la base centrale entretient un *tableau de détection* P à deux dimensions de taille $|R| \times |T|$, avec $|R| = n$ et $|T| = k$, dans lequel chaque élément $P[i][j]$ correspond à l'ensemble des individus détectés sur la région $r_i \in R$ à l'instant $t_j \in T$, c'est-à-dire $P[i][j] = D_{i,j}$.

$$P = \begin{bmatrix} D_{1,1} & D_{1,2} & ... & D_{1,k} \\ D_{2,1} & D_{2,2} & ... & D_{2,k} \\ ... & ... & ... & ... \\ D_{n,1} & D_{n,2} & ... & D_{n,k} \end{bmatrix} \tag{6.2}$$

Le tableau P est le support du réseau social créé entre les individus détectés. Ses liens sont pondérés par la fréquence des interactions. P sauvegarde également l'historique de l'évolution du réseau social.
Les liens du réseau social correspondent à des interactions sémantiquement proches d'interactions de type "*contacts de proximité*". En effet, l'architecture que nous proposons ne permet pas d'évaluer la position exacte des individus sur la zone. Ainsi, la signification d'un "contact de proximité", correspond à la situation où des individus sont détectés sur une même région. Nous considérons en effet qu'en présence d'une telle situation, les individus sont suffisamment proches pour être en relation.
Plus précisément, nous créons des liaisons entre tous les oiseaux détectés sur la même région. Si ces liaisons existent déjà, la valeur de leur fréquence associée est incrémentée.

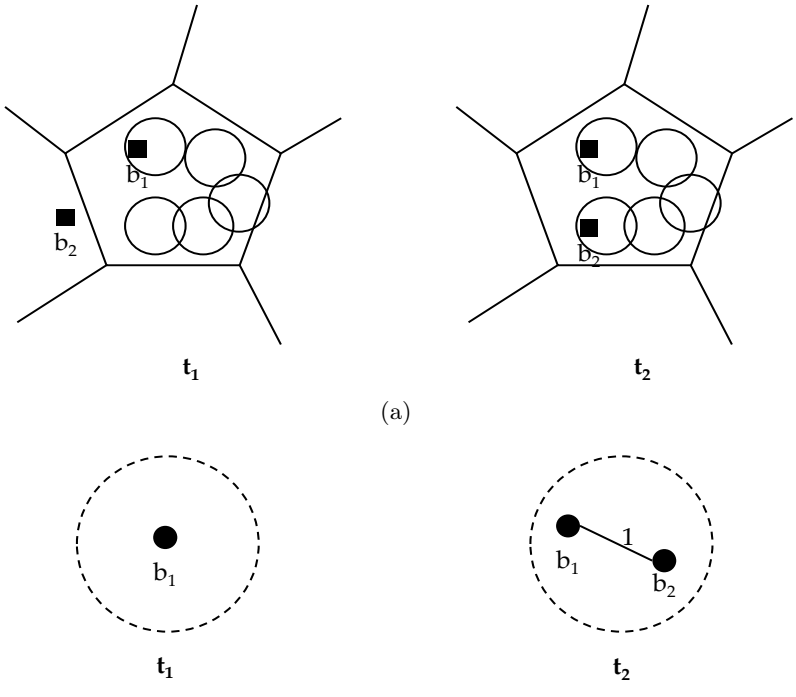


FIGURE 6.8 – Construction du réseau social
(a) Situation sur le terrain, (b) Mise à jour du réseau

Par exemple, comme l'illustre la Figure 6.8, si à l'instant t_1 l'individu b_1 est détecté sur la région r_1 (voir Figure 6.8(a)), il est ajouté comme noeud au réseau (voir Figure 6.8(b)). De même, si à l'instant t_2, les individus b_1 et b_2 sont détectés sur la région r_1, c'est-à-dire $D_{1,2} = \{b_1, b_2\}$, l'individu b_2 est ajouté comme nouveau noeud au réseau et le lien $e_1 = (b_1, b_2)$ est créé.
La procédure de construction du graphe à partir du tableau de détection P est décrite dans l'Algorithme 10.

algorithme 10 Construction du réseau social à partir du tableau de détection P

Précondition : P : **Tableau de détection**
1. V : **Ensemble de noeuds** $\leftarrow \emptyset$ (avec $V \subset B$)
2. E : **Ensemble de liens** $\leftarrow \emptyset$
3. G : **Réseau social** $\leftarrow (V, E)$
4. **pour** t de 1 à k **faire**
5. %*Création de tous les liens survenus au temps t*
6. **pour** r de 1 à n **faire**
7. $D_{r,t} \leftarrow P[r][t]$
8. **pour** tous les couples distincts $e \leftarrow (b_1, b_2)$, tel que $b_1, b_2 \in D_{r,t}$ **faire**
9. %*Vérification des noeuds*
10. **si** $b_1 \notin V$ **alors**
11. Ajouter b_1 à V
12. **fin si**
13. **si** $b_2 \notin V$ **alors**
14. Ajouter b_2 à V
15. **fin si**
16. %*Vérification du lien*
17. **si** $e \notin E$ **alors**
18. Ajouter e à E
19. **sinon**
20. Incrémenter fréquence de e dans E
21. **fin si**
22. **fin pour**
23. **fin pour**
24. **fin pour**
25. **retour** G

Nous obtenons ainsi un réseau liant les individus entre eux, qui représente la fréquence des détections des individus dans la même région. Une telle approche exige de restreindre le découpage des régions à une taille adéquate. En effet, si les régions sont trop étendues, de nombreux individus risquent d'être détectés dans la même région. Inversement, si le découpage est trop fin, les individus réellement proches risquent d'être détectés comme étant dans des régions différentes.

6.3.2 Visualisation

L'outil de simulation *Lypus* que nous avons développé dans ce travail permet de suivre en temps réel la construction et l'évolution de ce réseau. Dans l'interface de visualisation de *Lypus* les oiseaux sont représentés par un cercle jaune et les liens entre eux par un trait rouge (voir Figure 6.9). L'épaisseur du trait correspond à la fréquence de l'interaction ; plus un trait est épais entre deux individus et plus la fréquence de cette interaction est élevée.

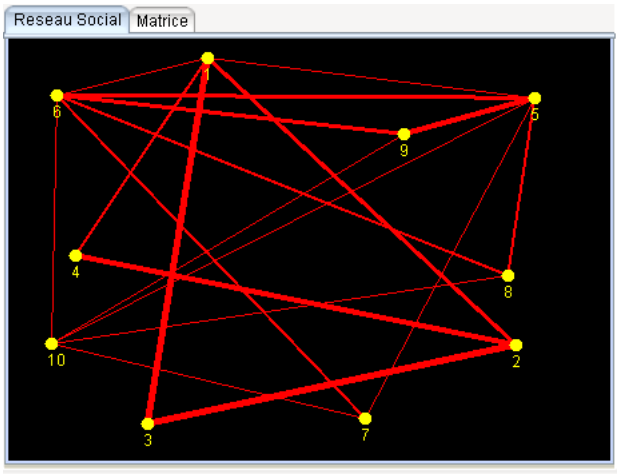


FIGURE 6.9 – Vue du réseau social des oiseaux

L'approche proposée permet également de générer un réseau biparti liant les oiseaux aux régions sur lesquelles ils sont détectés. Ainsi, chaque fois qu'un individu est détecté dans une région, un lien est créé, ou incrémenté, entre cet individu et la région sur laquelle il a été détecté. Un tel réseau peut par exemple être utilisé pour identifier les zones les plus fréquentées par les individus.

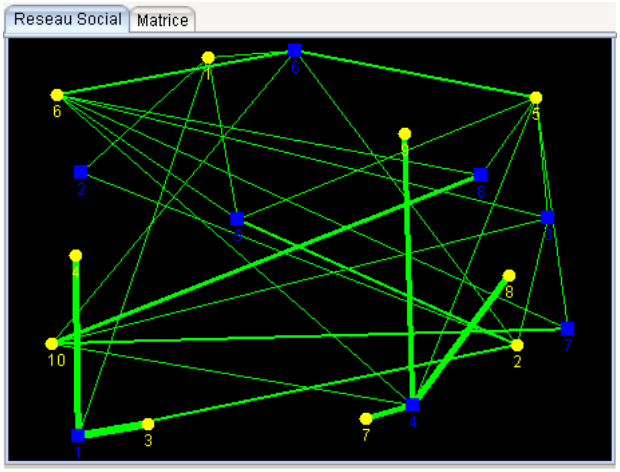


FIGURE 6.10 – Vue du réseau liant les individus aux régions

L'outil *Lypus* permet de suivre l'évolution de ce réseau en temps réel (voir Figure 6.10). Les cercles jaunes représentent les individus et les carrés bleus les régions. Comme précédemment, l'épaisseur du trait correspond à la fréquence de la détection.
Évidemment, il est également possible de suivre l'évolution des deux réseaux sur la même visualisation comme le montre la Figure 6.11).

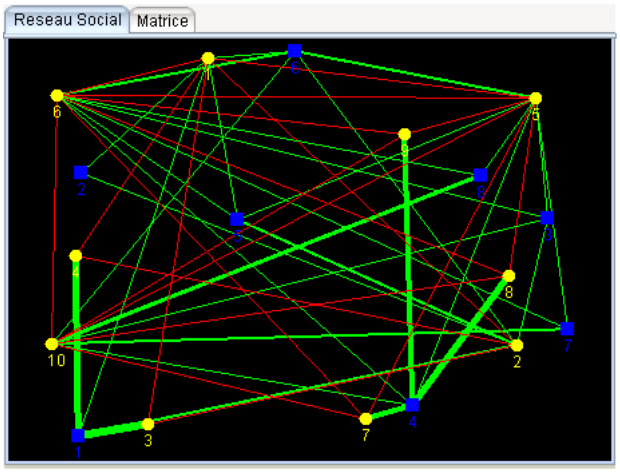


FIGURE 6.11 – Combinaison des deux vues

6.4 Expérimentations

Avant de déployer l'architecture en condition réelle sur le site de *la Caravelle*, une première étape importante consiste à étudier les facteurs qui influencent les résultats. Dans cet objectif, les simulations sont considérées comme des étapes essentielles pour la compréhension des phénomènes. Elles peuvent en effet permettre de tester différentes configurations et détecter les paramètres les plus importants pour la collecte des données. Plus généralement, elles peuvent être utilisées pour révéler des informations pertinentes pour l'optimisation du réseau de capteurs qui sera déployé sur le terrain.

Cette Section est consacrée à *Lypus*, l'outil de simulation que nous avons développé pour étudier les performances de notre solution, ainsi qu'aux résultats obtenus.
La Section 6.4.1 détaille l'outil, la Section 6.4.2 présente l'environnement de test utilisé et la Section 6.4.3 décrit les résultats obtenus.

6.4.1 Simulateur *Lypus*

Dans le but d'évaluer l'efficacité de l'architecture proposée pour la collecte d'informations sociales sur les oiseaux, nous avons développé l'outil *Lypus*, un environnement de simulation en 2D réalisé en JAVA, qui modélise les oiseaux dans leur habitat. *Lypus* a pour objectif de représenter un environnement naturel simulé, capable de reproduire virtuellement l'habitat des oiseaux dans lequel le réseau de capteurs peut être introduit. Le splash screen de *Lypus* est présenté sur la Figure 6.12.

Dans l'outil, la zone d'étude est tout d'abord divisée en régions uniformes. Bien que la prise en compte des caractéristiques des régions puisse permettre d'obtenir des informations utiles sur les comportements, nous cherchons ici à mesurer l'efficacité de l'architecture dans la collecte de données sociales. Ainsi, les caractéristiques particulières que peuvent présenter les différentes régions ne sont pas considérées dans l'outil.

Une fois la zone d'étude divisée en régions, les capteurs sont placés sur la zone et sont affiliés à la région à laquelle ils appartiennent. Tous les capteurs possèdent les mêmes capacités. Le rayon de détection des microphones et la durée de vie des batteries sont les mêmes. Les capteurs de détection communiquent avec leur capteur-passerelle, et les capteurs-passerelle communiquent avec la base centrale. Dans l'outil, nous ne nous intéressons pas à l'optimisa-

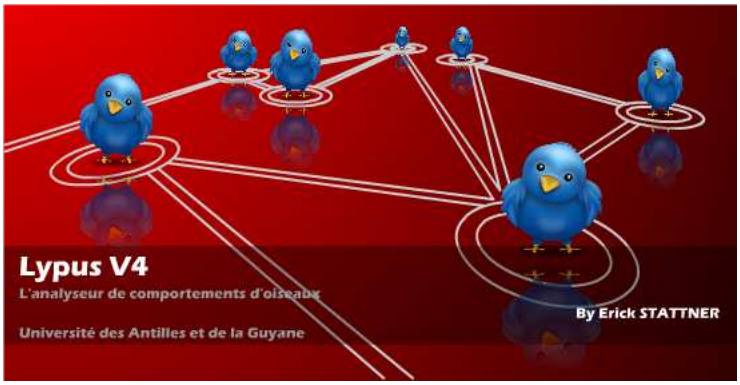


FIGURE 6.12 – Splash screen de *Lypus*

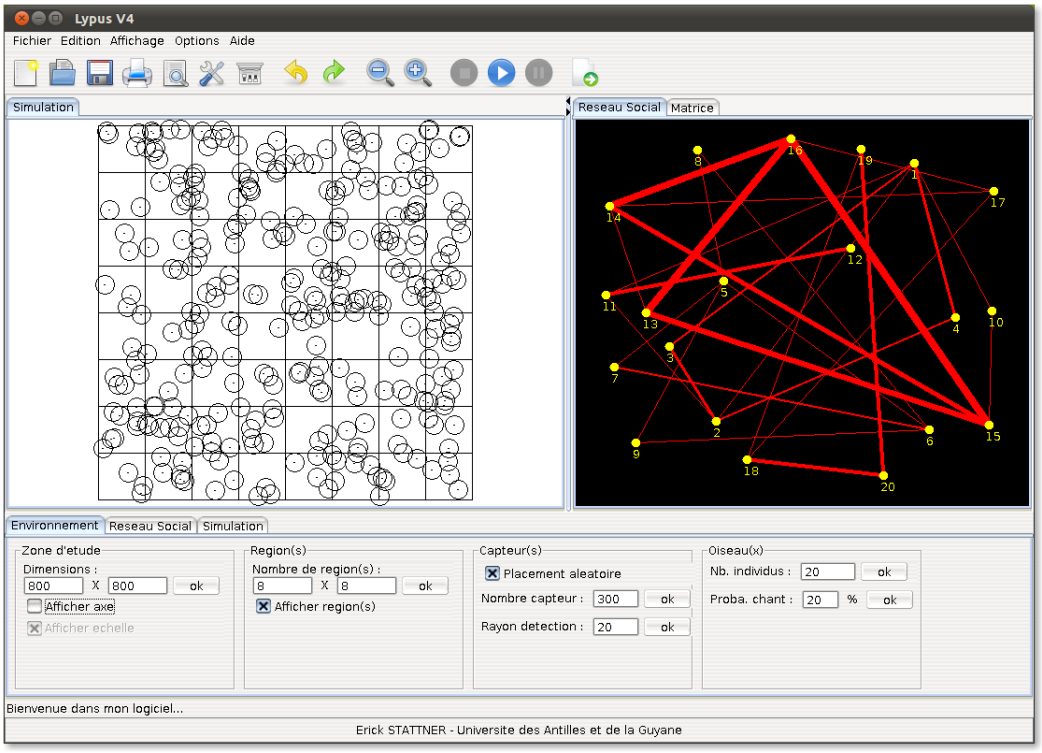
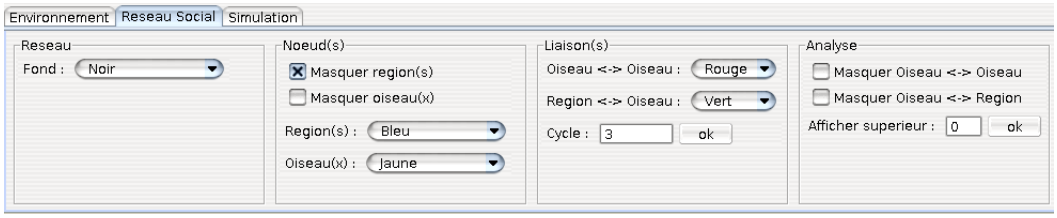


FIGURE 6.13 – Interface de *Lypus*

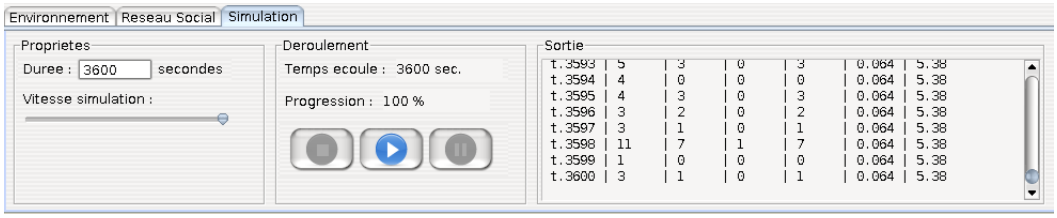
tion des échanges, ni aux éventuelles pertes ou corruptions d'information. Nous supposons donc que les communications sont toujours possibles.

Lypus est entièrement paramétrable. Il est possible de définir la taille de la zone d'étude, le nombre de régions et leur taille, le nombre de capteurs, la taille des rayons de détection, le type de placement (selon configuration souhaitée : aléatoire, dense ou uniforme), la taille de la population d'oiseaux, la probabilité de chant des oiseaux, la probabilité de trouver de la nourriture, etc. La Figure 6.13 montre une capture de l'interface de visualisation de *Lypus*.

L'interface de l'outil se présente sous la forme d'une fenêtre qui contient trois panneaux principaux.
(i) Le panneau de gauche représente l'environnement virtuel dans lequel se déroule la simulation. On peut y distinguer la zone d'étude et la séparation en régions uniformes représentée par des rectangles noirs. Les capteurs et les rayons de détection de leur microphone sont représentés sous la forme de cercles noirs. Sur l'exemple de la Figure 6.13 les capteurs sont placés de façon aléatoire. On peut d'ailleurs observer deux phénomènes intéressants avec ce type de placement : (1) la surface de la zone d'étude n'est pas couverte entièrement, (2) certains capteurs possèdent des rayons de détection qui chevauchent plusieurs régions. Enfin pendant la simulation, les individus qui chantent sont représentés par des carrés rouges sur la zone.
(ii) Le panneau de droite est le panneau de visualisation du réseau social. Il permet de suivre en temps réel la construction du réseau en fonction des informations collectées par la base. L'utilisateur peut interagir avec le réseau en déplaçant les noeuds, ou en appliquant les transformations proposées par le panneau de configuration sur le bas de la fenêtre.
(iii) Le panneau du bas, organisé sous la forme de trois onglets, permet d'interagir avec l'outil. (1) L'onglet "*Environnement*" permet de calibrer la simulation en fixant les différents paramètres présentés précédemment (taille des régions, nombre de capteurs, type de placement, etc.). (2) L'onglet "*Réseau social*" permet d'interagir avec le réseau social et fournit quelques outils d'analyse. (3) L'onglet "*Simulation*" permet de contrôler et de suivre l'évolution de la simulation. Il fournit également à chaque itération des détails sur les données collectées comme le nombre réel d'individus à avoir chanté, le nombre d'individus détectés, le nombre d'interactions identifiées, le nombre de communautés, etc.
Ces différents onglets sont présentés sur la figure 6.14.



(a)



(b)

FIGURE 6.14 – Différentes vues de l'interface *Lypus*
Vue pour (a) le paramétrage du réseau social (b) la simulation et les résultats

Nous nous sommes basés sur les comportements connus de l'espèce, de façon à ce que les oiseaux virtuels modélisés au sein de *Lypus* reproduisent le plus fidèlement les comportements du Moqueur Gorge Blanche.
Les oiseaux s'organisent en famille au sein desquelles trois types d'individus peuvent être

distingués : (i) les petits, (ii) les femelles et (iii) les mâles. Chaque catégorie à des comportements qui lui sont propres en fonction de sa situation.

(i) Le petit :

- Un petit n'a que deux types de comportements.
- Soit il est de *niveau 1*, c'est-à-dire qu'il reste dans le nid et reçoit de la nourriture de ses parents.
- Sinon il est de *niveau 2* et il est capable de suivre un de ses parents dans ses déplacements.

(ii) La femelle :

- Dans l'outil, nous considérons que toutes les femelles sont en couple, car nous n'avons aucune information sur le comportement de femelle solitaire. Ainsi dans ce monde virtuel, toutes les femelles partagent nécessairement un nid avec un mâle.
- Soit le couple n'a pas de petit, la femelle recherche alors de la nourriture sur le territoire dominé par le mâle.
- Soit la femelle est en incubation, elle reste dans le nid et couve ses oeufs.
- Soit les petits sont de niveau 1, elle recherche de la nourriture à proximité du nid pour elle et ses petits.
- Sinon les petits sont de niveau 2 et elle recherche de la nourriture pour elle et ses petits et peut être accompagnée d'un ou plusieurs petits.

(iii) Le mâle :

- Tous les mâles (en couple ou solitaire) occupent un territoire défini. Les mâles défendent leur territoire si un autre mâle, généralement solitaire, s'y aventure.
- Si le mâle est solitaire, il parcourt son territoire et s'aventure parfois dans celui des autres à la recherche d'une femelle.
- En revanche, si le mâle est en couple, nous distinguons alors plusieurs comportements. S'il n'a pas de petits, il recherche simplement de la nourriture sur son territoire. Si la femelle est en couvaison, le mâle porte régulièrement de la nourriture à la femelle. Si les petits sont de niveau 1, il recherche de la nourriture à proximité du nid pour lui et ses petits. Si les petits sont de niveau 2, il recherche de la nourriture pour lui et ses petits et peut être accompagné d'un ou plusieurs petits.

Les communautés auxquelles nous nous intéressons dans ce travail sont les familles. Cette espèce a pour caractéristique d'évoluer dans des espaces géographiques limités, ce qui rend notre approche pertinente pour détecter les communautés.

6.4.2 Environnement de test

Pour nos expériences, nous avons utilisé *Lypus* pour mettre en place un environnement de test composé de 20 individus formant 7 communautés. La simulation a été calibrée avec des paramètres que nous jugeons proches de la réalité.
Les individus évoluent sur une zone d'étude de $800m \times 800m$, divisée en 64 régions de $100m \times 100m$. Chaque capteur possède un rayon de détection de $20m$. Nous supposons que quand plusieurs oiseaux chantent à proximité d'un même capteur, cela génère du bruit dans "*l'environnement du capteur*", ce qui a pour conséquence de mettre en échec le processus de reconnaissance. Ces oiseaux ne sont donc pas pris en compte.

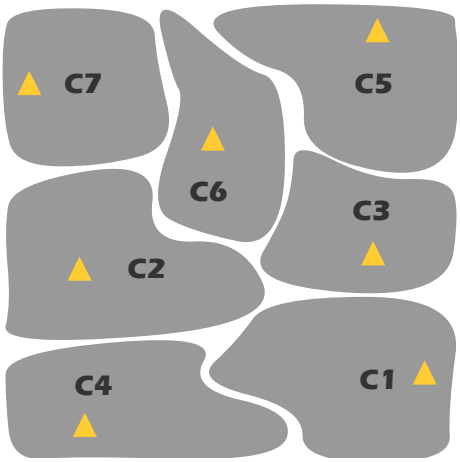


FIGURE 6.15 – Répartition des communautés et de leur territoire sur la zone d'étude

La répartition des communautés sur la zone d'étude est présentée sur la Figure 6.15. Les zones grisées représentent le territoire dominé par le mâle de chaque communauté et les triangles jaunes représentent l'emplacement des nids des différentes communautés.
Les communautés $C1$ et $C2$ sont des familles d'oiseaux composées chacune d'un couple qui a des petits de niveau 1, deux pour la première et trois pour la seconde. Les communautés $C3$ et $C6$ sont formées par des mâles solitaires. Les communautés $C5$ et $C7$ sont formés de couples qui ont des petits de niveaux 2, deux pour la première et un pour la seconde. Enfin, la communauté $C4$ est formée par un mâle et une femelle en couvaison.
Comme dans la réalité, l'architecture de collecte n'a aucune connaissance sur les communautés et leur territoire. Ainsi, lors de la division de la zone d'étude en régions, une même région peut couvrir plusieurs territoires.

Les individus se déplacent en fonction de leur catégorie selon les comportements présentés précédemment. Dans nos tests, la probabilité de chant des individus est fixée à 0.1 et la probabilité de trouver de la nourriture est fixée à 0.2.
De plus, nous considérons que les comportements n'évoluent pas au cours du temps. Typiquement, un petit de niveau 1 ne passe pas au niveau 2 pendant une simulation. De même, une femelle n'entame pas une phase de couvaison durant la simulation. Enfin, on ne considère pas non plus les cas de décès, ou de nouvel individu qui pourrait arriver sur la zone.
Les résultats ont été moyennés sur 20 itérations et obtenus avec la configuration matérielle suivante : Intel Core 2 Duo P8600 2.4Ghz, 4Go RAM, Linux Ubuntu 10.10, JDK 1.6.

Ainsi, l'objectif est de vérifier que l'architecture est effectivement en mesure de retrouver les différentes communautés et leur composition. Nous avons collecté les données sur une période de 60 minutes en mesurant toutes les 5 minutes l'évolution du réseau.

Les performances de l'approche ont été analysées selon les quatre configurations illustrées sur la Figure 6.16.
La *configuration 1* utilise un placement aléatoire de 300 **capteurs** (voir Figure 6.16(a)). Cette configuration crée deux phénomènes susceptibles d'impacter les résultats : la zone d'étude n'est pas entièrement couverte et les rayons de détection des capteurs peuvent chevaucher plusieurs régions.
La *configuration 2* correspond à un placement aléatoire de 600 **capteurs** (voir Figure 6.16(b)). Contrairement à la configuration précédente, elle a l'avantage d'augmen-

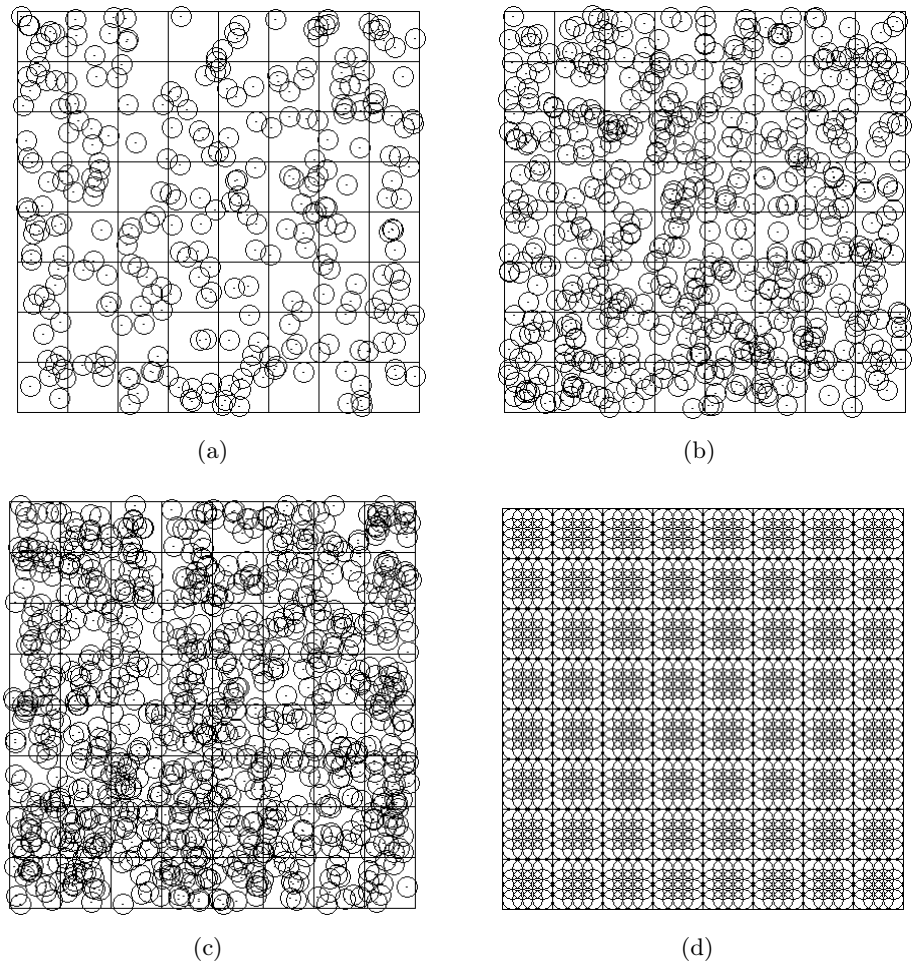


FIGURE 6.16 – Placement de capteurs dans *Lypus*
(a) configuration 1, (b) configuration 2, (c) configuration 3 et (d) Configuration 4

ter la surface couverte, mais elle augmente également le phénomène de chevauchement des rayons de détection des microphones sur plusieurs régions.

La *configuration 3* correspond à un placement dense et aléatoire de 900 **capteurs** (voir Figure 6.16(c)). Elle a pour objectif de couvrir complètement la zone d'étude. Comme pour la configuration 2, elle augmente la surface couverte, mais également le phénomène de chevauchement des rayons de détection.

Enfin **la *configuration 4* représente un placement dense et uniforme des capteurs** (voir Figure 6.16(d)). Elle correspond à une situation idéale dans laquelle toute la zone d'étude est couverte et dans laquelle les capteurs sont placés de façon à ce qu'ils ne couvrent que la région à laquelle ils sont affiliés.

Il est important de préciser que la division de la zone d'étude en régions ne tient pas compte du territoire réel des individus puisque cette information ne peut souvent pas être connue. Ainsi, toutes les configurations sont susceptibles de créer des situations dans lesquelles une même région est à l'intersection de plusieurs territoires, ce qui créera inévitablement des interactions entre des individus de communautés différentes lorsque des individus chantent

au même moment à la limite de leur territoire.

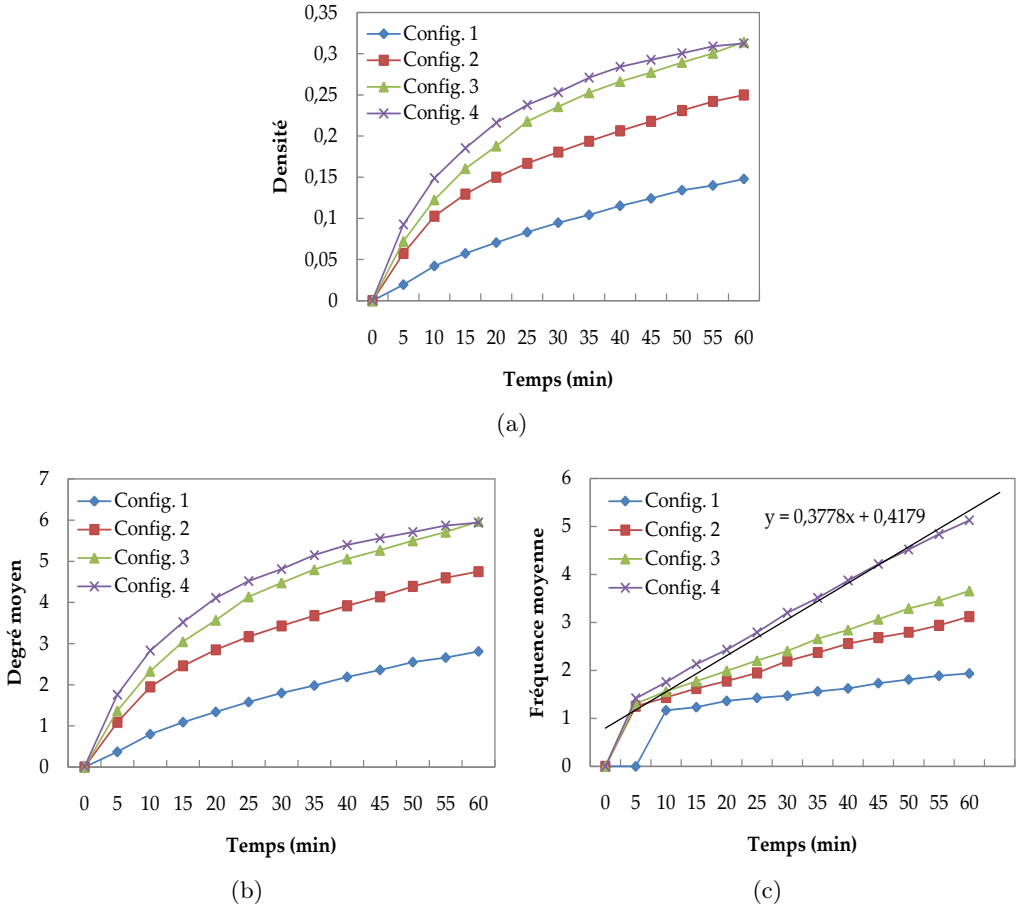
6.4.3 Résultats expérimentaux

Comme nous avons pu le voir au Chapitre 5, il existe plusieurs méthodes capables d'extraire des groupes de noeuds d'un réseau social. Dans ce travail, nous utilisons une approche simple, proposée par Croft *et al.* [Croft 2008a] sur des réseaux pondérés d'animaux, qui effectue la recherche en deux temps. (1) Les liens "*faibles*" sont dans un premier temps supprimés du réseau. Les liens faibles sont les liens les moins pertinents, c'est-à-dire ceux qui ont une fréquence basse. (2) Une fois la suppression des liens faibles effectuée, chaque composante connexe est considérée comme étant une famille d'individus.
Dans nos tests, nous considérons tous les liens ayant une fréquence comprise dans le premier quartile comme des liens faibles.

La difficulté du processus d'extraction vient ici du fait que ces liaisons minoritaires dans le réseau peuvent être générées par trois situations distinctes :
(i) La durée de collecte n'est pas suffisamment élevée et ne permet pas d'obtenir un grand nombre d'interactions entre les individus d'une même communauté.
(ii) Les rayons de détection des capteurs chevauchent plusieurs régions, ce qui génère un lien lorsque deux individus de communautés différentes se trouvent dans la limite de leur territoire et chantent en même temps.
(iii) Une même région couvre deux territoires. Comme dans le cas (ii) si deux individus sont à la limite de leur territoire et chantent en même temps, une interaction est détectée.
Ainsi, les liens générés par les situations (ii) et (iii) surviennent rarement puisqu'ils nécessitent une coïncidence spatio-temporelle forte, c'est-à-dire qu'il faut que les individus soient à la limite de leur territoire au même moment et chantent en même temps. Nous pouvons donc supposer que si la collecte est menée sur une période de temps suffisamment longue, nous réduisons la probabilité que le cas (i) surviennent, permettant ainsi de distinguer la situation (i) des situations (ii) et (iii).

Dans une première approche, nous nous sommes intéressés à l'évolution des propriétés du réseau au cours du temps. La Figure 6.17 présente une analyse comparative des configurations (configurations 1, 2, 3 et 4) dans l'évolution, (a) de la densité du réseau (b) du degré moyen des noeuds et (c) de la fréquence moyenne des interactions au cours du temps.

Comme attendu, quelle que soit la configuration, nous observons globalement que la densité du réseau croît avec le temps, puisque de nouvelles liaisons sont détectées à chaque itération (voir Figure 6.17(a)).
Cependant, nous pouvons également observer que pour toutes les configurations, cette croissance est très forte pendant les premières minutes de la collecte, puis connait une phase de quasi-stabilité ou seuls quelques nouveaux liens sont détectés. Par exemple pour la configuration 4, la densité passe de 0.09 à 0.15 entre la $5e$ et la $10e$ minute, alors qu'elle passe de 0.300 à 0.308 entre la $50e$ et la $55e$ minute. Cela s'explique par le fait que quand la durée de la collecte est suffisamment longue, toutes les interactions pertinentes sont collectées durant les premières minutes de la collecte. Le réseau est ensuite alimenté de quelques interactions marginales, jusqu'à atteindre un état de stabilité.
Quand on compare les résultats obtenus entre les différentes configurations, on observe que la densité du réseau tend à croître avec le nombre de capteurs. Par exemple, après une heure de collecte, la densité du réseau obtenue avec la configuration 1 est de 0.31, alors que celle obtenue avec la configuration 4 est de 0.14. Cela est évidemment dû au taux de couverture de la zone d'étude. En effet, plus la zone est couverte et plus le nombre d'individus détectés augmente, générant ainsi potentiellement un nombre plus important de liaisons détectées.



(a)

(b)

(c)

FIGURE 6.17 – Évolution des propriétés du réseau au cours du temps (a) densité, (b) degré moyen et (c) fréquence moyenne des interactions

Le degré moyen des noeuds suit les mêmes tendances que celles observées pour la densité (voir Figure 6.17(b)).
En effet, comme de nouveaux liens sont détectés à chaque itération, le degré moyen croît avec le temps. Par exemple pour la configuration 1 le degré moyen est de 0.019 après 5 minutes de collecte contre 0.14 après 60 minutes.
Comme pour la densité, nous observons une phase de croissance rapide durant les premières minutes de la collecte, puis une diminution de cette croissance au cours du temps. Pour la configuration 4 par exemple, le degré moyen est de 1.76 à la $5e$ minute et de 2.83 à la $10e$, alors qu'il est de 5.71 à la $50e$ minute contre 5.87 à la $55e$. Cela est dû au nombre de nouveaux liens détectés qui connait également une phase de décroissance.
Enfin, le degré moyen est plus élevé pour les configurations ayant le plus de capteurs. Cela s'explique par l'augmentation du nombre de liens détectés quand la zone est plus couverte.

En ce qui concerne la fréquence moyenne des interactions (voir Figure 6.17(c)), il est intéressant d'observer que quelle que soit la configuration, la fréquence croît quasi-linéairement avec le temps. Par exemple, pour la configuration 4, la fréquence moyenne peut être approchée par la fonction $y = 0.3778 \times t + 0.4179$.

Dans une deuxième étape, nous avons extrait, à chaque instant, les communautés des réseaux générés, selon l'approche proposée par Croft *et al.* [Croft 2008a]. La Figure 6.18 permet de comparer au cours du temps et pour chaque configuration, l'erreur absolue entre le nombre de communautés détectées et le nombre de communautés attendues.

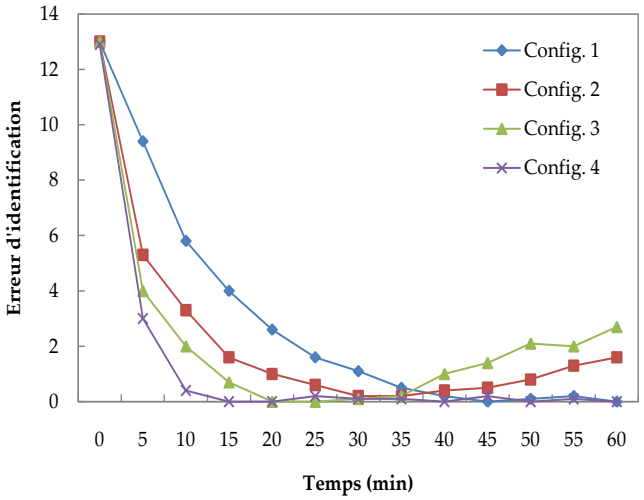


FIGURE 6.18 – Erreur sur la détection des communautés

Globalement, nous pouvons observer que quelle que soit la configuration, l'erreur est en général décroissante. Par exemple, pour la configuration 1, l'erreur est de 9.4 à la $5e$ minute contre 5.8 à la $10e$ minute et 0 à la $45e$ minute. Ce résultat suggère (1) que l'accumulation des données tend à améliorer les performances de la solution et (2) qu'il existe un temps d'accumulation nécessaire pour extraire les communautés.
Une observation intéressante concerne le temps à partir duquel les différentes configurations fournissent des réseaux capables d'extraire efficacement les familles d'individus. Nous pouvons en effet observer que plus la zone est couverte par des capteurs et plus le temps d'accumulation nécessaire est faible. Par exemple, dès la $15e$ minute, le réseau obtenu par la configuration 4 permet d'extraire les communautés d'oiseaux sans erreur. En revanche, il faut respectivement attendre la $20e$, la $30e$ et la $45e$ minute pour extraire les communautés sans erreur des réseaux obtenus par les configurations 3, 2 et 1. Plus généralement, ces résultats confirment d'une part l'efficacité de l'approche dans le processus d'extraction des communautés et montrent d'autre part que le temps nécessaire à une collecte efficace des données diminue lorsque la zone est densément couverte.
Enfin, pour les configurations 2 et 3, on observe que l'erreur connaît une phase de croissance à partir d'un certain temps : la $40e$ minute pour la configuration 3 et la $45e$ pour la configuration 2. Cela peut s'expliquer par les phénomènes de chevauchement des rayons de détection, très présents sur ces deux types de configuration. En effet, plus le temps de collecte est élevé et plus on détecte des interactions entre des individus de communautés différentes, donc des cas de faux positifs. Après un certain temps de collecte, l'accumulation de ces faux positifs est telle, que le premier quartile n'est plus suffisant pour caractériser les liens non-pertinents. Ces résultats suggèrent ainsi que les performances de la solution sont améliorées quand les rayons de détection sont confinés aux régions associées.

6.5 Conclusion

La collecte des données est un sujet important pour l'étude des réseaux. Si des jeux de données ont pu être générés à partir de sites d'échange et de partage, ou à partir de périphériques mobiles capables de collecter en temps réel des données spatio-temporelles sur des individus, il existe des situations dans lesquelles ce type d'approche ne peut pas être utilisé.

Dans ce chapitre, nous nous sommes intéressés à la question de la collecte dans le contexte particulier d'études menées sur le Moqueur Gorge Blanche, une espèce d'oiseaux endémique à la Martinique, pour laquelle les périphériques mobiles traditionnellement utilisés, tels que les colliers GPS ou les puces RFID, ne peuvent pas être envisagés. Notre objectif dans ce travail était de proposer une architecture de collecte de données sociales et de mesurer son efficacité en vérifiant qu'elle permettait d'identifier les structures familiales. Les contributions de ce chapitre peuvent être résumées comme suit.

(i) Nous avons proposé **une architecture de collecte basée sur un réseau de capteurs sans fil**. Nous avons ensuite montré comment l'architecture peut être utilisée pour générer un réseau social pondéré, basé sur la fréquence des présences dans une même région géographique. Ce réseau est ensuite utilisé pour identifier les structures familiales. Bien que nous ayons expérimenté une méthode d'identification des individus basée sur l'analyse des chants, l'architecture proposée est assez souple pour supporter différents types de techniques de reconnaissance, la seule condition étant que les individus et les régions sur lesquelles ils sont présents puissent être identifiés.

(ii) Avant un déploiement réel de la solution sur le terrain, nous avons développé **l'outil de simulation *Lypus*** pour étudier d'une part la faisabilité de l'approche et analyser d'autre part les paramètres qui pouvaient influencer les performances de la solution. *Lypus* est un environnement de simulation en 2D qui permet de recréer virtuellement une zone d'étude sur laquelle des oiseaux ayant des comportements similaires à ceux de l'espèce étudiée sont recréés et sur laquelle un réseau de capteurs virtuel est introduit. L'outil permet ainsi de mesurer la qualité des données obtenues selon différents paramètres.

(iii) Une première étude a mis en application l'outil pour étudier **l'impact de différentes configurations de capteurs** sur les performances de la collecte. Nous nous sommes dans un premier temps intéressés à l'évolution des propriétés des réseaux obtenus. Puis nous avons vérifié à chaque instant si le réseau obtenu permettait d'identifier les familles. Nous avons ainsi pu observer que le nombre de capteurs avait une influence directe sur le temps d'accumulation nécessaire pour une identification efficace des familles d'individus.

Le travail présenté dans ce chapitre constitue une étape préliminaire et nécessaire avant tout déploiement réel de la solution sur le terrain. Il est un garant de la faisabilité de l'approche et permet d'envisager un déploiement sur le terrain à court terme.

Chapitre 7

Conclusion et perspectives

Les travaux de recherche présentés dans ce mémoire apportent un certain nombre de contributions à l'étude des réseaux sociaux.

Les Chapitres 3 et 4, consacrés à l'étude des **phénomènes de diffusion sur des réseaux dynamiques** contribuent à une meilleure compréhension de ces processus sur les réseaux du monde réel. En effet, en partant du constat que les travaux traditionnels ne tenaient pas compte de la dynamique des réseaux, nous avons proposé une approche générique, permettant la modélisation d'un phénomène de diffusion sur des réseaux en évolution. En identifiant ensuite plusieurs facteurs qui peuvent être à l'origine de modifications structurelles du réseau, nous avons mis en application cette approche pour comprendre l'impact de ces facteurs sur le processus de diffusion. Si les résultats obtenus ont démontré le profond impact de la dynamique du réseau sur ces phénomènes, nous avons également pu mettre en évidence l'influence de l'espace, aussi bien social que géographique, sur ces processus.
Les travaux menés ont également des implications directes pour l'étude des réseaux, puisqu'ils nous amènent à nous interroger sur les méthodes de modélisation de la dynamique et des processus prenant place sur les réseaux, ainsi que leurs interdépendances.

Les travaux décrits dans le Chapitre 5 sur **l'analyse conceptuelle des réseaux sociaux** permettent d'extraire un nouveau type de connaissance des réseaux. En effet, après avoir constaté que certaines questions pertinentes sur la structure des réseaux ne trouvaient pas de réponses à travers les méthodes classiques de fouille de réseaux sociaux, nous avons proposé une nouvelle approche qui exploite à la fois la structure des réseaux et les attributs des noeuds. Cette approche présente deux intérêts majeurs pour l'étude des réseaux sociaux. Elle permet tout d'abord d'extraire une connaissance statistiquement pertinente sur les groupes de noeuds les plus connectés du réseau, puis elle fournit une représentation sémantique du réseau en synthétisant la connaissance acquise. Les expériences menées ont permis de confirmer la pertinence de l'approche au regard des motifs extraits et les bonnes performances de l'algorithme d'extraction proposé.
D'une façon plus générale, ces travaux s'inscrivent dans une branche nouvelle de la recherche sur les réseaux, qui consiste à considérer toutes les informations disponibles sur le réseau lors de la phase d'extraction de connaissances.

Enfin, les travaux du Chapitre 6, menés sur **la collecte automatique de données sociales** permettent aujourd'hui d'envisager un champ d'applications à venir très large et le recueil de nouveaux jeux de données, autres que ceux traditionnellement extraits des sites d'échange et de partage. En effet, de nombreux types d'interactions sociales échappent encore aux travaux menés sur les réseaux en raison des difficultés liées à la disponibilité des données. Si des travaux ont déjà été menés sur l'utilisation de dispositifs mobiles pour la collecte de contacts de proximité, nous avons pu observer que ces dispositifs ne pouvaient pas être déployés dans toutes les situations. Le travail précurseur décrit dans ce chapitre propose une architecture composée de capteurs fixes qui répond aux exigences fonctionnelles et techniques de la collecte de données sociales au sein d'une population animale. Les résultats obtenus en simulation ont ainsi montré l'intérêt de cette architecture pour collecter les interactions et extraire les liens qui structurent la société animale étudiée.

Du point de vue de la recherche sur les réseaux, cette architecture peut être utilisée pour collecter des données dans des situations spécifiques où ni des observations humaines, ni des sites d'échange ne permettent d'obtenir des données pertinentes.

Perspectives

Les travaux entamés dans ce mémoire ouvrent de nouvelles perspectives de recherche intéressantes, que nous prévoyons d'aborder dans nos travaux futurs.

En ce qui concerne la diffusion sur des réseaux sociaux dynamiques, les travaux menés dans ce mémoire se sont uniquement intéressés à des contacts "*directs*", c'est-à-dire que nous supposions que la transmission s'effectuait d'un individu à l'autre. Cependant, dans de nombreuses situations, la transmission peut également survenir de manière "*indirecte*", à travers des vecteurs de diffusion intermédiaires tels qu'une poignée de porte, un tableau noir, une chaise, un mouchoir, etc. La dimension spatiale intrinsèque du modèle *ER* est par exemple tout à fait adaptée pour modéliser ce type de situation, caractérisée par la présence de zones géographiques.
Une évolution de cette approche serait de supposer que même dans leur état de mobilité, les individus peuvent passer plus ou moins de temps sur les zones qu'ils visitent : cinéma, centre commercial, musée, hôpital, etc. Comme l'illustre la Figure 7.1, sur laquelle l'intensité de la couleur (points chauds) correspond au temps que passe un individu sur la zone, une première évolution du modèle *ER* a déjà été implémentée pour modéliser cette situation. Un des objectifs est d'étudier comment le temps d'immobilité passé sur une zone influence le processus de diffusion.

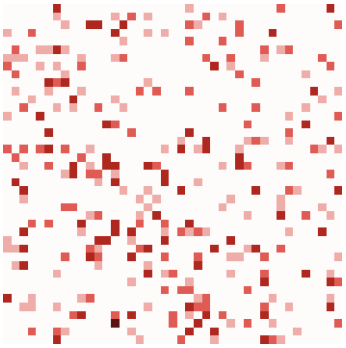
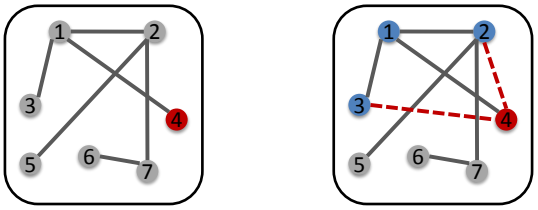


FIGURE 7.1 – Extension du modèle *ER* avec points chauds distribués dans l'espace (intensité de la couleur correspond au temps que passe un individu sur la zone)

Une autre piste intéressante concerne l'étude de l'interdépendance entre le comportement des individus et l'évolution du phénomène. En effet, dans le contexte de la diffusion de maladies par exemple, il est fréquent que les individus répondent à l'apparition et au développement d'une épidémie en adoptant des comportements de défense, qui peuvent se traduire au niveau du réseau social sous-jacent par des suppressions de liens avec des individus infectés ou la création de nouveaux liens avec des individus sains. La dynamique sur le réseau est alors le résultat d'une part de comportements individuels ou sociaux, mais pourrait également être modifiée ou adaptée selon l'état du voisinage des noeuds ou l'évolution globale du phénomène.
Le modèle *D2SNet* a été conçu pour intégrer ces dépendances. L'intérêt de cette nouvelle approche serait de modéliser différents types de comportements de défense et de vérifier s'ils

permettent véritablement de limiter la propagation, ou au contraire, si ces comportements n'ont finalement aucun impact sur la diffusion globale.

Enfin, s'intéresser à des réactions individuelles, qui visent à immuniser localement un individu ou un groupe d'individus, nous amène directement à nous interroger sur les stratégies d'intervention actuellement mises en place. En effet, nous avons pu mettre en évidence l'impact de la dynamique du réseau sur le processus de diffusion. Or, les stratégies d'intervention actuelles ne prennent pas encore en compte cette réalité. Ainsi, une piste de recherche prometteuse serait de proposer des stratégies d'interventions adaptées, qui tiendraient compte de la dynamique des réseaux.
Dans cet objectif, les méthodes de prédiction de liens pourraient par exemple avoir des applications intéressantes. Ces méthodes pourraient en effet être utilisées pour prédire la formation de liens entre des individus *susceptibles* et des individus *infectés* et proposer des stratégies d'intervention tenant compte de l'évolution du réseau, comme l'illustre la Figure 7.2.



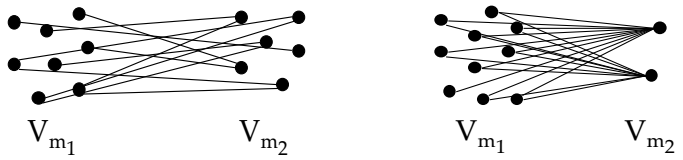
(a) (b)

FIGURE 7.2 – Prédiction de liens appliquée au contrôle de l'épidémie
(a) Réseau instant t, (b) Prédiction et intervention appliqués à l'instant $(t + \Delta)$

Par exemple, partant d'un réseau contenant un individu *infecté* (individu 4 sur la Figure 7.2(a)), une stratégie d'intervention pourrait consister à vacciner les individus en contact avec l'individu *infecté* et les individus ayant une forte probabilité de l'être (voir individus $1, 2$ et 3 sur la Figure 7.2(b)) de façon à neutraliser tous les liens susceptibles de véhiculer la maladie autour de l'individu *infecté*.

En ce qui concerne l'extraction de liens conceptuels, un de nos premiers objectifs concerne le passage à l'échelle de l'algorithme *MFCL-Min*. En effet, bien que les résultats obtenus aient pu mettre en évidence les bonnes performances de notre solution par rapport à une approche naïve, nous avons également pu constater que le nombre d'attributs est un paramètre qui augmente significativement le temps de calcul. Une de nos pistes d'étude pour la réduction du temps de calcul consiste à limiter la recherche des liens conceptuels aux motifs impliquant uniquement des itemsets fréquents. Bien que cette hypothèse soit théoriquement fausse, puisque comme l'illustre la Figure 7.3 les liens conceptuels peuvent être trouvés dans deux types de configurations, soit les deux itemsets impliqués sont relativement fréquents, soit au moins un des deux est très fréquent, nos premiers résultats ont montré que la réalité des comportements humains fait que le cas deuxième cas ne survient que très rarement.

Typiquement, sur le jeu de données utilisé dans les tests, nous avons pu observer qu'aucun itemset ayant un support inférieur à 0.15 n'intervenait dans un lien conceptuel, ce qui laisse supposer qu'il existe un seuil de fréquence de l'itemset à partir duquel il est impliqué dans un lien conceptuel. Cette observation pourrait permettre d'adapter l'algorithme de façon à limiter l'espace de recherche aux propriétés, elles-mêmes, fréquentes. Cependant, elle soulève également une question majeure : comment déterminer le seuil idéal qui permet-



(a) (b)

FIGURE 7.3 – Configurations d'apparition des liens conceptuels

trait d'extraire tous les liens conceptuels sans aucune perte ? Peut-être qu'à cet effet, des corrélations entre des propriétés du réseau (distribution des degrés, coefficient de clustering, etc.) et le seuil à utiliser peuvent être mises en évidence.

Ce travail a également des implications directes pour le domaine de la science des réseaux, et plus particulièrement pour l'axe de la modélisation. En effet, plusieurs modèles ont été proposés pour générer des réseaux sociaux. Cependant, nous observons que la plupart des modèles tentent de reproduire des propriétés structurelles particulières : petit-monde, scale-free, etc. Ainsi, les liens conceptuels extraits de données réelles pourraient permettre d'aborder le problème de la génération de jeux de données sous un angle nouveau, par la conception de nouveaux modèles de génération qui tenteraient de reproduire des structures conforment aux vues conceptuelles extraites de données réelles.

Enfin, d'une façon plus générale, nous pensons que la prise en compte de la nature et de la sémantique des liens est un aspect fondamental des travaux à venir sur l'étude des réseaux sociaux. En effet, dans de nombreuses situations, la sémantique du lien peut permettre de comprendre et d'expliquer certains phénomènes ou comportements. Par exemple, des liens de concurrence au sein d'une entreprise ou d'un groupe, peuvent souvent être à l'origine de conflits. Ainsi, en complément des mesures d'intérêt déjà proposées, la pertinence des liens conceptuels pourrait être mesurée à travers de nouvelles mesures qui tiennent compte de la sémantique des liens.

En ce qui concerne la collecte de données sociales, les études ornithologiques récentes menées sur l'espèce ont permis d'améliorer la connaissance que l'on avait du comportement des individus et notamment de leur cycle de vie de la naissance à l'âge adulte. Ces nouvelles connaissances seront intégrées au simulateur, de façon à ce qu'il soit en mesure de reproduire le plus fidèlement possible le comportement de ces animaux.
Nous utiliserons les résultats obtenus en simulation pour déployer une architecture réelle sur le site de la Caravelle et collecter les interactions sociales survenant entre les individus de Moqueurs Gorge Blanche. Les interactions ainsi collectées devraient permettre de comprendre l'organisation sociale de l'espèce et éventuellement de mettre à jour des comportements encore insoupçonnés.

Les méthodes présentées dans ce mémoire pourraient également trouver des applications pertinentes. Typiquement, étudier comment une maladie se propage chez cette espèce en voie de disparition permettrait aux scientifiques de mieux se prémunir contre ce type de danger en identifiant par exemple les individus dangereux de ce point de vue. De même, si des informations sur les différents individus de l'espèce peuvent être connues (age, sexe, poids, etc.), rechercher des liens conceptuels pourrait permettre d'améliorer les connaissances sur cette espèce en identifiant les caractéristiques qui lient les individus de l'espèce.

Enfin à long terme, une piste intéressante serait de mettre au point une méthode générique pour la conception d'architecture répondant aux besoins fonctionnels, techniques et opérationnels pour la collecte de données sociales chez tout type de population.

Bibliographie

[Agrawal 1994] R. Agrawal et R. Srikant. *Fast Algorithms for Mining Association Rules in Large Databases*. In Proceedings of the 20th International Conference on Very Large Data Bases, pages 487–499, 1994. (Cité en pages 110 et 115.)

[Albano 2012] Alice Albano, Jean-Loup Guillaume et Benedicte Le Grand. *File Diffusion in a Dynamic Peer-to-peer Network*. Mining Social Network Dynamic 2012 Workshop (MSND), 2012. (Cité en pages 44, 47 et 48.)

[Albert 2000] R. Albert, H. Jeong et A.L. Barabási. *Error and attack tolerance of complex networks*. Nature, vol. 406, no. 6794, pages 378–382, 2000. (Cité en page 28.)

[Albert 2002] R. Albert et A. L. Barabasi. *Statistical mechanics of complex networks*. Reviews of Modern Physics, vol. 74, 2002. (Cité en pages 22, 23, 54, 56 et 62.)

[Barabasi 1999] A. Barabasi et R. Albert. *Emergence of Scaling in Random Networks*. Science, vol. 286(5439), pages 509 – 512, 1999. (Cité en pages 3, 21, 46 et 55.)

[Barabasi 2002] A. L. Barabasi. Linked : The new science of networks. Perseus Books, 2002. (Cité en pages 2 et 54.)

[Barabasi 2003] A. L. Barabasi et E. Bonabeau. *Scale-free networks*. Scientific American, vol. 288, no. 5, pages 60–69, 2003. (Cité en page 20.)

[Barabasi 2009] Albert-Laszlo Barabasi et James Fowler. *The century of networks*. Seed salon mahazine, 2009. (Cité en page 1.)

[Barnes 1954] J. A. Barnes. *Class and Committees in a Norwegian Island Parish*. Human Relations, vol. 7, pages 39–58, 1954. (Cité en page 23.)

[Barrett 2008] C. L. Barrett, K. R. Bisset, Stephen G. Eubank, X. Feng et M. V. Marathe. *EpiSimdemics : an efficient algorithm for simulating the spread of infectious disease over large realistic social networks*. In Proceedings of the 2008 ACM/IEEE conference on Supercomputing, 2008. (Cité en pages 5, 47 et 124.)

[Bastian 2009] M. Bastian, S. Heymann et M. Jacomy. *Gephi : An Open Source Software for Exploring and Manipulating Networks*. 2009. (Cité en pages 103 et 134.)

[Beletsky 1982] L. David Beletsky. *The Role of Song in Individual Recognition in the Indigo Bunting*. Animal Behaviour, vol. 31(2), pages 355–362, 1982. (Cité en page 147.)

[Belik 2011] V. Belik, T. Geisel et D. Brockmann. *Recurrent host mobility in spatial epidemics : beyond reaction-diffusion*. European Physical Journal B-Condensed Matter, vol. 84, no. 4, page 579, 2011. (Cité en pages 75, 76, 80 et 82.)

[Bettstetter 2004] C. Bettstetter, H. Hartenstein et X. Pérez-Costa. *Stochastic properties of the random waypoint mobility model*. Wireless Networks, vol. 10, no. 5, pages 555–567, 2004. (Cité en page 78.)

[Blondel 2008] V.D. Blondel, J. L. Guillaume, R. Lambiotte et E. Lefebvre. *Fast unfolding of communities in large networks*. Journal of Statistical Mechanics : Theory and Experiment, vol. 2008, page P10008, 2008. (Cité en page 107.)

[Boccaletti 2006] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez et D.U. Hwang. *Complex networks : Structure and dynamics*. Physics reports, vol. 424, no. 4, pages 175–308, 2006. (Cité en pages 16, 22 et 62.)

[Boguna 2003] M. Boguna, R. Pastor-Satorras, A. Diaz-Guilera et A. Arenas. *Emergence of clustering, correlations, and communities in a social network model*. Arxiv preprint cond-mat 0309263, 2003. (Cité en pages 46, 54, 55, 62 et 63.)

[Bollobas 2001] B. Bollobas. Random graphs, volume 73. Cambridge Univ Pr, 2001. (Cité en page 55.)

[Borgatti 2002a] S.P. Borgatti. *NetDraw Software for Network Visualization. Analytic Technologies : Lexington, KY*. 2002. (Cité en pages 103 et 134.)

[Borgatti 2002b] S.P. Borgatti, M.G. Everett et L.C. Freeman. *Ucinet for Windows : Software for social network analysis*. Harvard Analytic Technologies, vol. 2006, 2002. (Cité en pages 103 et 134.)

[Borner 2007] K. Borner, S. Sanyal et A. Vespignani. *Network Science*. Blaise Cronin (Ed) Annual Review of Information Science and Technology, vol. 41, pages 537–607, 2007. (Cité en pages 2, 14, 16, 22, 23 et 62.)

[Bott 1957] E. Bott. Family and social network. The Free Press, 1957. (Cité en pages 3, 14, 19 et 24.)

[Bredin 2006] H. Bredin et G. Chollet. *Synchronisation voix/levres pour la verification d'identite basee sur les visages parlants*. 2006. (Cité en page 148.)

[Brin 1998] S. Brin et L. Page. *The anatomy of a large-scale hypertextual Web search engine*. Computer Networks and ISDN Systems, vol. 30, 1998. (Cité en page 25.)

[Brisson 2008] L. Brisson et M. Collard. *How to Semantically Enhance a Data Mining Process ?* In ICEIS, pages 103–116, 2008. (Cité en page 122.)

[Broadbent 1957] Simon Broadbent et John Hammersley. *Percolation processes I. Crystals and mazes*. Proceedings of the Cambridge Philosophical Society, vol. 53, pages 629–641, 1957. (Cité en page 27.)

[Camp 2002] T. Camp, J. Boleng et V. Davies. *A survey of mobility models for ad hoc network research*. Wireless communications and mobile computing, vol. 2, no. 5, pages 483–502, 2002. (Cité en pages 75, 76 et 79.)

[Chavoya 2008] A. Chavoya et Y. Duthen. *A cell pattern generation model based on an extended artificial regulatory network*. Biosystems, vol. 94, pages 95–101, 2008. (Cité en page 14.)

[Chen 2007a] D. Chen, J. Yang, R. Malkin et H. D. Wactlar. *Detecting social interactions of the elderly in a nursing home environment*. ACM Trans. Multimedia Comput. Commun. Appl., vol. 3, 2007. (Cité en pages 39 et 142.)

[Chen 2007b] Y.-D. Chen, C. Tseng, C.-C. King, T.-S. J. Wu et H. Chen. *Incorporating geographical contacts into social network analysis for contact tracing in epidemiology : a study on Taiwan SARS data*. In NSF conference on Intelligence and security informatics : BioSurveillance, 2007. (Cité en page 85.)

[Cheng 2010] H. Cheng, X. Yan et J. Han. *Mining Graph Patterns*. Managing and Mining Graph Data, pages 365–392, 2010. (Cité en pages 26 et 110.)

[Christakis 2007] N.A. Christakis et J.H. Fowler. *The spread of obesity in a large social network over 32 years*. New England Journal of Medicine, vol. 357, no. 4, pages 370–379, 2007. (Cité en pages 1 et 36.)

[Christakis 2008] N.A. Christakis et J.H. Fowler. *The collective dynamics of smoking in a large social network*. New England journal of medicine, vol. 358, no. 21, pages 2249–2258, 2008. (Cité en page 1.)

[Christakis 2010] N. A. Christakis et J. H. Fowler. *Social network sensors for early detection of contagious outbreaks*. PloS one, vol. 5(9), 2010. (Cité en pages 35, 43, 48 et 58.)

[Christensen 2010] C. Christensen, I. Albert, B. Grenfell et R. Albert. *Disease dynamics in a dynamic social network*. Physica A : Statistical Mechanics and its Applications, vol. 389(13), pages 2663–2674, Février 2010. (Cité en page 47.)

[Christley 2005] R. M. Christley, G. L. Pinchbeck, R. G. Bowers, D. Clancy, N. P. French, R. Bennett et J. Turner. *Infection in Social Networks : Using Network Analysis to Identify High-Risk Individuals*. American Journal of Epidemiology, vol. 162(10), pages 1024–1031, 2005. (Cité en pages 3, 25, 32, 34, 43, 44 et 48.)

[Cohen 2000] R. Cohen, K. Erez, D. Ben-Avraham et S. Havlin. *Resilience of the Internet to random breakdowns*. Physical Review Letters, vol. 85, no. 21, pages 4626–4628, 2000. (Cité en pages 27 et 33.)

[Cohen 2001] R. Cohen, K. Erez, D. Ben-Avraham et S. Havlin. *Breakdown of the Internet under intentional attack*. Physical Review Letters, vol. 86, no. 16, pages 3682–3685, 2001. (Cité en pages 28 et 33.)

[Collard 2007] M. Collard et J. Vansnick. *How to measure interestingness in data mining : a multiple criteria decision analysis approach*. In RCIS, pages 395–400, 2007. (Cité en page 122.)

[Collard 2012] Martine Collard, Philippe Collard et Erick Stattner. *Mobility and information flow : percolation in a multi-agent model*. 3rd International Conference on Ambient Systems, Networks and Technologies, 2012. (Cité en page 5.)

[Combe 2012] D. Combe, C. Largeron, E. Egyed-Zsigmond et M Gery. *Combining relations and text in scientific network clustering*. International Conference on Advances in Social Networks Analysis and Mining, 2012. (Cité en page 26.)

[Corley 2008] Courtney D. Corley, Armin R. Mikler, Diane J. Cook et Karan Singh. *Dynamic intimate contact social networks and epidemic interventions*. International Journal of Functional Informatics and Personalised Medicine, vol. 1(2), pages 171 – 188, 2008. (Cité en page 70.)

[Cosley 2010] D. Cosley, D. Huttenlocher, J. Kleinberg, X. Lan et S. Suri. *Sequential influence models in social networks*. In Proc. 4th International Conference on Weblogs and Social Media, 2010. (Cité en page 3.)

[Croft 2008a] D P Croft, R James et J Krause. Exploring animals social networks. Princeton University Press, 2008. (Cité en pages 4, 14, 35, 139, 143, 161 et 163.)

[Croft 2008b] Darren P. Croft, Richard James et Jens Krause. Exploring animals social networks, chapitre 3. Visual Exploration. Princeton, 2008. (Cité en page 72.)

[Crooks 2009] Andrew Crooks, Andrew Hudson-Smith et Joel Dearden. *Agent Street : An Environment for Exploring Agent-Based Models in Second Life*. Journal of Artificial Societies and Social Simulation, vol. 12, no. 4, page 10, 2009. (Cité en page 77.)

[Dabouineau 2004] Laurent Dabouineau. *Pourquoi les oiseaux chantent-ils ?* Le rale d'eau, vol. 119, pages 10–14, 2004. (Cité en page 147.)

[Daley 1965] DJ Daley et D.G. Kendall. *Stochastic rumours*. IMA Journal of Applied Mathematics, vol. 1, no. 1, pages 42–55, 1965. (Cité en page 48.)

[Daly 2007] Elizabeth M. Daly et Mads Haahr. *Social network analysis for routing in disconnected delay-tolerant MANETs*. In Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing, MobiHoc '07, pages 32–40, New York, NY, USA, 2007. ACM. (Cité en page 23.)

[Daude 2005] Eric Daude et Emmanuel Eliot. *Exploration de l'effet des types de mobilites sur la diffusion des epidemies*. In Proceedings of Theo Quant'05., 2005. (Cité en pages 80 et 81.)

[Davidsen 2002] J Davidsen, H Ebel et S Bornholdt. *Emergence of a Small World from Local Interactions : Modeling Acquaintance Networks*. Phys Rev Lett, vol. 88, no. 12, page 128701, 2002. (Cité en pages 62 et 70.)

[Deo 2004] N. Deo. Graph theory with applications to engineering and computer science. PHI Learning Pvt. Ltd., 2004. (Cité en page 14.)

[Dezsho 2002] Z. Dezsho et A.L. Barabasi. *Halting viruses in scale-free networks*. Physical Review E, vol. 65, no. 5, 2002. (Cité en page 34.)

[Dorogovtsev 2002] S.N. Dorogovtsev et J.F.F. Mendes. *Evolution of networks*. Adv. Phys, 2002. (Cité en pages 46 et 54.)

[Easley 2010] David Easley et Jon Kleinberg. Networks, crowds, and markets : Reasoning about a highly connected world, chapitre 21. Epidemics. Cambridge University Press, 2010. (Cité en page 32.)

[Ebel 2003] H. Ebel, J. Davidsen et S. Bornholdt. *Dynamics of social networks*. Complexity, vol. 8, pages 24–27, 2003. (Cité en page 55.)

[Ekman 2008] F. Ekman, A. Keranen, J. Karvo et J. Ott. *Working day movement model*. In Proceeding of the 1st ACM SIGMOBILE workshop on Mobility models, pages 33–40. ACM, 2008. (Cité en pages 75, 80 et 82.)

[Eliot 2006] Emmanuel Eliot et Eric Daude. *Spatial diffusion of epidemics : Approaches of complexities in geography*. Espace populations societes, vol. 2(3), pages 403–416, 2006. (Cité en pages 29 et 80.)

[Epstein 2008] J.M. Epstein. *Why model ?* Journal of Artificial Societies and Social Simulation, vol. 11, no. 4, page 12, 2008. (Cité en page 81.)

[Erdos 1960] P. Erdos et A. Renyi. On the evolution of random graphs. Akad. Kiado, 1960. (Cité en pages 20 et 55.)

[Erdos 2006] P. Erdos, A. Meir, V. T. Sos et P. Turan. *On some applications of graph theory, I*. Discrete Mathematics, vol. 306, no. 10-11, 2006. (Cité en page 14.)

[Eubank 2005] S. Eubank. *Network based models of infectious disease spread*. Japanese journal of infectious diseases, vol. 58(6), 2005. (Cité en pages 29, 47, 54 et 56.)

[Eugster 2004] P.T. Eugster, R. Guerraoui, A.M. Kermarrec et L. Massoulié. *From epidemics to distributed computing*. IEEE computer, vol. 37, no. 5, pages 60–67, 2004. (Cité en page 48.)

[Euler 1741] L. Euler. *Solutio problematis ad geometriam situs pertinentis*. Commentarii academiae scientiarum Petropolitanae, vol. 8, pages 128–140, 1741. (Cité en page 12.)

[Fortunato 2009] Santo Fortunato. *Community detection in graphs*. Physics Reports, vol. 486, pages 75–174, 2009. (Cité en pages 3, 26, 72, 111 et 134.)

[Fowler 2008] J.H. Fowler et N.A. Christakis. *The dynamic spread of happiness in a large social network*. BMJ : British medical journal, vol. 337, page a2338, 2008. (Cité en page 1.)

[Fruchterman 1991] T. M. J. Fruchterman et E. M. Reingold. *Graph drawing by force-directed placement*. Softw. Pract. Exper., vol. 21, pages 1129–1164, November 1991. (Cité en page 33.)

[Galeottia 1991] P. Galeottia et G. Pavanb. *Individual recognition of male Tawny owls (Strix aluco) using spectrograms of their territorial calls*. Ethology Ecology and Evolution, vol. 3(2), pages 113–126, 1991. (Cité en page 147.)

[Ganter 2005] Bernhard Ganter, Gerd Stumme et Rudolf Wille. *Formal Concept Analysis, Foundations and Applications*. Lecture Notes in Computer Science, vol. 3626, 2005. (Cité en page 116.)

[Garrison 1960] W.L. Garrison. *Connectivity of the interstate highway system*. Papers in Regional Science, vol. 6, no. 1, pages 121–137, 1960. (Cité en page 14.)

[Gaume 2010] B. Gaume, E. Navarro et H. Prade. *A parallel between extended formal concept analysis and bipartite graphs analysis*. Computational Intelligence for Knowledge-Based Systems Design, pages 270–280, 2010. (Cité en pages 108 et 112.)

[Getoor 2005] L Getoor et C P Diehl. *Link mining : a survey*. SIGKDD Explor., vol. 7, pages 3–12, 2005. (Cité en pages 3, 25, 72, 107 et 109.)

[Gonzalez 2008] M. C. Gonzalez, C. A. Hidalgo et A. Barabasi. *Understanding individual human mobility patterns*. Nature, vol. 453, pages 779–782, 2008. (Cité en pages 75, 76, 80 et 81.)

[Granovetter 1973] M.S. Granovetter. *The strength of weak ties*. American journal of sociology, pages 1360–1380, 1973. (Cité en page 55.)

[Granovetter 1978] M. Granovetter. *Threshold models of collective behavior*. American journal of sociology, pages 1420–1443, 1978. (Cité en page 48.)

[Gross 2006] Thilo Gross, Carlos J. Dommar D'Lima et Bernd Blasius. *Epidemic Dynamics on an Adaptive Network*. Physical Review Letters, vol. 96, no. 20, page 208701, May 2006. (Cité en page 46.)

[Gruhl 2004] D. Gruhl, R. Guha, D. Liben-Nowell et A. Tomkins. *Information diffusion through blogspace*. In Proceedings of the 13th international conference on World Wide Web, pages 491–501. ACM, 2004. (Cité en page 48.)

[Guille 2012] Adrien Guille et Hakim Hacid. *A predictive model for the temporal dynamics of information diffusion in online social networks*. In Proceedings of the 21st international conference companion on World Wide Web, WWW '12 Companion, pages 1145–1152, New York, NY, USA, 2012. ACM. (Cité en page 47.)

[Guimera 2002] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt et A. Arenas. *Self-similar community structure in organisations*. Arxiv preprint cond-mat/0211498, 2002. (Cité en page 62.)

[Gutman 1972] I. Gutman et N. Trinajstic. *Graph theory and molecular orbitals. Total [phi]-electron energy of alternant hydrocarbons*. Chemical Physics Letters, vol. 17, no. 4, pages 535–538, 1972. (Cité en page 14.)

[Haas 1997] Z.J. Haas. *A new routing protocol for the reconfigurable wireless networks*. In IEEE 6th International Conference on Universal Personal Communications, volume 2, pages 562–566, 1997. (Cité en page 79.)

[Hammersley 1980] JM Hammersley et DJA Welsh. *Percolation theory and its ramifications*. Contemporary Physics, vol. 21, no. 6, pages 593–605, 1980. (Cité en page 27.)

[Henderson 2004] T. Henderson, D. Kotz et I. Abyzov. *The changing usage of a mature campus-wide wireless network*. In 10th international conference on Mobile computing and networking, pages 187–201. ACM, 2004. (Cité en page 78.)

[Inokuchi 2000] A. Inokuchi, T. Washio et H. Motoda. *An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data*. In 4th European Conference on Principles of Data Mining and Knowledge Discovery, pages 13–23, 2000. (Cité en page 110.)

[Jardosh 2003] A. Jardosh, E.M. Belding-Royer, K.C. Almeroth et S. Suri. *Towards realistic mobility models for mobile ad hoc networks.* In Proceedings of the 9th annual international conference on Mobile computing and networking, pages 217–229. ACM, 2003. (Cité en page 78.)

[Kamada 1989] T. Kamada et S. Kawai. *An algorithm for drawing general undirected graphs.* Inf. Process. Lett., vol. 31, pages 7–15, April 1989. (Cité en page 33.)

[Kephart 1993] J.O. Kephart et S.R. White. *Measuring and modeling computer virus prevalence.* In Research in Security and Privacy, 1993. Proceedings., 1993 IEEE Computer Society Symposium on, pages 2–15. IEEE, 1993. (Cité en page 48.)

[Kermack 1927] W. O. Kermack et A. G. McKendrick. *A Contribution to the Mathematical Theory of Epidemics.* Proceedings of the Royal Society of London. Series A, vol. 115, pages 700–721, 1927. (Cité en pages 29 et 47.)

[Kermarrec 2003] A.M. Kermarrec, L. Massoulié et A.J. Ganesh. *Probabilistic reliable dissemination in large-scale systems.* Parallel and Distributed Systems, IEEE Transactions on, vol. 14, no. 3, pages 248–258, 2003. (Cité en page 48.)

[Kesten 1982] H. Kesten. Percolation theory for mathematicians, volume 2. Birkhauser, 1982. (Cité en page 27.)

[Klovdahl 1985] A. S. Klovdahl. *Social networks and the spread of infectious diseases : the AIDS example.* Soc Sci Med, vol. 21(11), pages 1203–16, 1985. (Cité en page 32.)

[Kossinets 2006] G. Kossinets et D.J. Watts. *Empirical analysis of an evolving social network.* Science, vol. 311, no. 5757, pages 88–90, 2006. (Cité en pages 55 et 62.)

[Kumpula 2007] J.M. Kumpula, J.P. Onnela, J. Saramäki, K. Kaski et J. Kertész. *Emergence of communities in weighted networks.* Physical review letters, vol. 99, no. 22, page 228701, 2007. (Cité en pages 54, 55, 62 et 70.)

[Kuramochi 2001] M. Kuramochi et G. Karypis. *Frequent Subgraph Discovery.* In Proceedings of the 2001 IEEE International Conference on Data Mining, pages 313–320, 2001. (Cité en pages 4 et 110.)

[Kuramochi 2005] M. Kuramochi et G. Karypis. *Finding Frequent Patterns in a Large Sparse Graph.* Data Min. Knowl. Discov., vol. 11, pages 243–271, 2005. (Cité en pages 3 et 110.)

[Kuznetsov 2009] S.O. Kuznetsov et D.I. Ignatov. *Concept stability for constructing taxonomies of web-site users.* Arxiv preprint arXiv :0905.1424, 2009. (Cité en pages 108 et 111.)

[Laibowitz 2006] Mathew Laibowitz, Jonathan Gips, Ryan Aylward, Alex Pentland et Joseph A. Paradiso. *A Sensor Network for Social Dynamics.* Information Processing In Sensor Networks, vol. 5, pages 483–491, 2006. (Cité en pages 39, 54 et 141.)

[Le Grand 2009] B. Le Grand, M. Soto et M.A. Aufaure. *Conceptual and Spatial Footprints for Complex Systems Analysis : Application to the Semantic Web.* In Database and Expert Systems Applications, 2009. (Cité en pages 111, 112 et 134.)

[Leskovec 2007] Jure Leskovec, Lada A. Adamic et Bernardo A. Huberman. *The dynamics of viral marketing.* ACM Trans. Web, vol. 1, 2007. (Cité en page 1.)

[Levy 2006] Christophe Levy, Georges Linares, Pascal Nocera et Jean-Francois Bonastre. *Gmm-based acoustic modeling for embedded speech recognition.* International Conference on Speech Communication and Technology, 2006. (Cité en page 148.)

[Li 2004] Sheng Li, Meng Meng et Hongru Ma. *Epidemic Spreading in Dynamic Small World Networks.* page 9, 2004. (Cité en pages 44 et 46.)

[Liben-Nowell 2007] David Liben-Nowell et Jon Kleinberg. *The link-prediction problem for social networks*. J. Am. Soc. Inf. Sci. Technol., vol. 58, pages 1019–1031, 2007. (Cité en pages 3 et 26.)

[Liquiere 2006] M. Liquiere. *Some links between formal concept analysis and graph mining*. Mining graph data, pages 227–252, 2006. (Cité en page 111.)

[Lloyd 2001] A.L. Lloyd et R.M. May. *How viruses spread among computers and people*. Science, vol. 292, no. 5520, pages 1316–1317, 2001. (Cité en page 33.)

[Lu 2003] Qing Lu et Lise Getoor. *Link-based Classification*. In ICML, pages 496–503, 2003. (Cité en page 26.)

[Lusseau 2007] D. Lusseau. *Evidence for social role in a dolphin social network*. Evolutionary ecology, vol. 21(3), pages 357–66, 2007. (Cité en page 141.)

[Mariani 2002] J. Mariani. Analyse, synthese et codage de la parole, traitement du langage parle. Hermes, 2002. (Cité en page 148.)

[Marsili 2004] M. Marsili, F. Vega-Redondo et F. Slanina. *The rise and fall of a networked society : a formal model*. Proceedings of the National Academy of Sciences of the United States of America, vol. 101, no. 6, page 1439, 2004. (Cité en pages 62 et 70.)

[Martincic 2005] F. Martincic et L. Schwiebert. Handbook of sensor networks : Algorithms and architectures, chapitre Introduction to wireless sensor networking. 2005. (Cité en page 144.)

[May 1965] K.O. May. *The origin of the four-color conjecture*. JSTOR, vol. 56, pages 346–348, 1965. (Cité en page 13.)

[McPherson 2001] M. McPherson, L. Smith-Lovin et J.M. Cook. *Birds of a feather : Homophily in social networks*. Annual review of sociology, 2001. (Cité en page 63.)

[Meyers 2005] L. A. Meyers, B. Pourbohloul, M. E. Newman, D. M. Skowronski et R. C. Brunham. *Network theory and SARS : predicting outbreak diversity*. Journal of theoretical biology, vol. 232, no. 1, pages 71–81, January 2005. (Cité en pages 29 et 47.)

[Milgram 1967] S. Milgram. *The Small World Problem*. Psychology Today, vol. 1, pages 61–67, 1967. (Cité en pages 3, 19, 20, 24 et 55.)

[Musolesi 2009] M. Musolesi et C. Mascolo. *Mobility models for systems evaluation. A survey*. Middleware for Network Eccentric and Mobile Applications, pages 43–62, 2009. (Cité en page 79.)

[Nekovee 2007] M. Nekovee, Y. Moreno, G. Bianconi et M. Marsili. *Theory of rumour spreading in complex social networks*. Physica A : Statistical Mechanics and its Applications, vol. 374, no. 1, pages 457–470, 2007. (Cité en page 48.)

[Newman 2003] M.E.J. Newman. *The structure and function of complex networks*. SIAM review, pages 167–256, 2003. (Cité en pages 23, 28 et 62.)

[Newman 2010] M. E. J. Newman. Networks : An introduction, volume 1, chapitre 17. Epidemics on Networks, pages 627–676(50). Oxford University Press, 2010. (Cité en pages 2, 14, 32 et 58.)

[Nijssen 2005] Siegfried Nijssen et Joost N. Kok. *The Gaston Tool for Frequent Subgraph Mining*. Electr. Notes Theor. Comput. Sci., vol. 127, no. 1, pages 77–87, 2005. (Cité en page 110.)

[Noymer 2001] A. Noymer. *The transmission and persistence of urban legends : Sociological application of age-structured epidemic models*. Journal of Mathematical Sociology, vol. 25, no. 3, pages 299–323, 2001. (Cité en page 48.)

[Olguin 2008] D.O. Olguin et A. Pentland. *Social sensors for automatic data collection*. Americas Conference on Information Systems, 2008. (Cité en pages 4, 39, 54 et 141.)

[Ostapenko 1991] VV Ostapenko et AP Yakovleva. *Mathematical questions of modeling and control in water distribution problems*. Control Cybern, vol. 20, pages 99–111, 1991. (Cité en page 13.)

[Pajot 2001] Stephane Pajot. *Percolation et economie*. PhD thesis, Universite de Nantes, 2001. (Cité en page 28.)

[Pastor-Satorras 2001] R. Pastor-Satorras et A. Vespignani. *Immunization of complex networks*. Physical Review E, vol. 65, 2001. (Cité en page 34.)

[Pauwels 2007] Eric J. Pauwels, Albert A. Salah et Romain Tavenard. *Sensor Networks for Ambient Intelligence*. Multimedia Signal Processing, 2007. MMSP 2007. IEEE 9th Workshop on, pages 13–16, 2007. (Cité en page 40.)

[Pham 2004] D. Pham. From agent-based computational economics towards cognitive economics. in Bourgine P., Nadal J.P eds : Cognitive Economics : An Interdisciplinary Approach. Springer verlag, 2004. (Cité en page 83.)

[Phillmore 2002] L. S. Phillmore, C. B. Sturdy, M. R. Turyk et R. G. Weisman. *Discrimination of individual vocalizations by black-capped chickadees*. Animal learning and behavior, vol. 30(1), pages 43–52, 2002. (Cité en page 147.)

[Press 2005] National Academies Press, editeur. National research council committee on network science for future army applications. Network Science, 2005. (Cité en page 2.)

[Quinn 1979] N. Quinn et M. Breuer. *A forced directed component placement procedure for printed circuit boards*. Circuits and Systems, vol. 26(6), pages 377–388, 1979. (Cité en page 33.)

[Read 2008] Jonathan M Read, Ken T D Eames et W John Edmunds. *Dynamic social networks and the implications for the spread of infectious disease.* Journal of the Royal Society Interface the Royal Society, vol. 5, Issue : 26, pages 1001–1007, 2008. (Cité en pages 16, 32, 44 et 46.)

[Rhodes 1996] C. J. Rhodes et R. M. Anderson. *Dynamics in a lattice epidemic model*. Physics Letters A, vol. 210, pages 183–188, 1996. (Cité en page 47.)

[Riadh 2009] T.M. Riadh, B. Le Grand, M.A. Aufaure et M. Soto. *Conceptual and statistical footprints for social networks' characterization*. In Proceedings of the 3rd Workshop on Social Network Mining and Analysis, page 8. ACM, 2009. (Cité en pages 108, 111 et 134.)

[Rumble 2001] M.A. Rumble, L. Benkobi, F. Lindzey et R.S. Gamo. *Evaluating elk habitat interactions with GPS collars*. Tracking Animals with GPS, 2001. (Cité en page 141.)

[Ryan 2004] P. G. Ryan, S. L. Petersen, G. Peters et D. Gremillet. *GPS tracking a marine predator : the effects of precision, resolution and sampling rate on foraging tracks of African Penguins*. Marine Biology, vol. 145(2), 2004. (Cité en page 141.)

[Sade 1994] D. S. Sade et M. M. Dow. *Primate social networks*. In Advances in social network analysis, ed. S. Wasserman and J. Galaskiewicz. California : Sage Publications, pages 152–66, 1994. (Cité en page 141.)

[Salathe 2010a] Marcel Salathe et James H. Jones. *Dynamics and Control of Diseases in Networks with Community Structure*. PLoS Comput Biol, vol. 6, 2010. (Cité en pages 32, 34, 43 et 70.)

[Salathe 2010b] Marcel Salathe, Maria Kazandjieva, Jung W. Lee, Philip Levis, Marcus W. Feldman et James H. Jones. *A High-Resolution Human Contact Network for Infectious Disease Transmission*. Aug 2010. (Cité en pages 4, 48 et 85.)

[Sanchez 2001] M. Sanchez et P. Manzoni. *ANEJOS : a java based simulator for ad hoc networks*. Future Generation Computer Systems, vol. 17, no. 5, pages 573–583, 2001. (Cité en page 78.)

[Schedl 2007] M. Schedl, P. Knees, K. Seyerlehner et T. Pohle. *The comirva toolkit for visualizing music-related data*. In Eurographics IEEE VGTC Symposium on Visualization, 2007. (Cité en page 150.)

[Shorrocks 2006] B. Shorrocks et D. P. Croft. *Giraffe necks and networks*. Mpala News, vol. 3, page 3, 2006. (Cité en page 141.)

[Simon 2006] Gary Simon. *Hot spot spatial models for data reported as counts over geographic areas*, 2006. (Cité en page 88.)

[Snasel 2008] V. Snasel, Z. Horak et A. Abraham. *Understanding social networks using formal concept analysis*. In Web Intelligence and Intelligent Agent Technology, volume 3, pages 390–393, 2008. (Cité en pages 3, 108 et 111.)

[Snasel 2009] V. Snasel, Z. Horak, J. Kocibova et A. Abraham. *Analyzing social networks using FCA : complexity aspects*. In Web Intelligence and Intelligent Agent Technologies, 2009. WI-IAT'09. IEEE/WIC/ACM International Joint Conferences on, volume 3, pages 38–41. IEEE, 2009. (Cité en page 112.)

[Song 2010] C. Song, Z. Qu, N. Blumm et A.L. Barabasi. *Limits of predictability in human mobility*. Science, vol. 327, pages 1018–1021, 2010. (Cité en pages 76 et 80.)

[Spitzer 2001] F. Spitzer. Principles of random walk, volume 34. Springer Verlag, 2001. (Cité en pages 78 et 81.)

[Stattner 2010a] Erick Stattner, Martine Collard, Philippe Hunel et Nicolas Vidot. *A Data Collection Framework for Tracking Collective Behaviour Patterns*. In RCIS, pages 43–50, 2010. (Cité en page 7.)

[Stattner 2010b] Erick Stattner, Martine Collard, Philippe Hunel et Nicolas Vidot. *Detecting movement patterns with wireless sensor networks : application to bird behavior*. In MoMM, pages 251–258, 2010. (Cité en page 7.)

[Stattner 2011a] Erick Stattner, Martine Collard, Philippe Hunel et Nicolas Vidot. *Wireless sensor networks for social network data collection*. In IEEE Conference on Local Computer Networks (LCN), pages 867–874, 2011. (Cité en page 7.)

[Stattner 2011b] Erick Stattner, Martine Collard et Nicolas Vidot. *Diffusion in Dynamic Social Networks : Application in Epidemiology*. 22nd International Conference on Database and Expert Systems Applications, 2011. (Cité en page 5.)

[Stattner 2011c] Erick Stattner, Martine Collard et Nicolas Vidot. *Towards Merging Models of Information Spreading and Dynamic Phenomena in Social Networks*. World Summit on the Knowledge Society, 2011. (Cité en page 5.)

[Stattner 2012a] E. Stattner et M. Collard. *FLMin : An Approach for Mining Frequent Links in Social Networks*. International Conference on Networked Digital Technologies, 2012. (Cité en page 6.)

[Stattner 2012b] E. Stattner et M. Collard. *GT-FLMin : Un Outil Graphique pour l'Extraction de Liens Frequents dans les Reseaux Sociaux*. Conference Internationale Francophone sur l'Extraction et la Gestion de Connaissance, 2012. (Cité en page 132.)

[Stattner 2012c] E. Stattner et M. Collard. *Social-Based Conceptual Links : Conceptual Analysis Applied to Social Networks*. International Conference on Advances in Social Networks Analysis and Mining, 2012. (Cité en pages 6 et 132.)

[Stattner 2012d] E. Stattner, M. Collard et N. Vidot. *Network-based Modeling in Epidemiology : An Emphasis on Dynamics*. International Journal of Information System Modeling and Design (IJISMD), vol. 3(3), 2012. (Cité en page 5.)

[Stattner 2012e] E. Stattner, M. Collard et N. Vidot. *Sociability VS Network Dynamics : Impact of Two Aspects of Human Behavior on Diffusion Phenomena*. International Conference on Advances in Social Networks Analysis and Mining, 2012. (Cité en pages 5 et 70.)

[Stattner 2012f] Erick Stattner et Martine Collard. *Extracion de Liens Frequents dans les Reseaux Sociaux*. Conference Internationale Francophone sur l'Extraction et la Gestion de Connaissance, 2012. (Cité en page 6.)

[Stattner 2012g] Erick Stattner et Martine Collard. *Frequent Links : An Approach that Combines Attributes and Structure for Extracting Frequent Patterns in Social Networks*. 16th East-European Conference on Advances in Databases and Information Systems, 2012. (Cité en page 6.)

[Stattner 2012h] Erick Stattner et Martine Collard. *How to Extract Frequent Links with Frequent Itemsets in Social Networks ?* Sixth International Conference on Research Challenges in Information Science (RCIS), 2012. (Cité en pages 133 et 134.)

[Stattner 2012i] Erick Stattner et Martine Collard. *MAX-FLMin : An Approach for Mining Maximal Frequent Links and Generating Semantical Structures from Social Networks*. 23rd International Conference on Database and Expert Systems Applications, 2012. (Cité en page 6.)

[Stattner 2012j] Erick Stattner, Martine Collard et Nicolas Vidot. *D2SNet : Dynamics of diffusion and dynamic human behaviour in social networks*. Computers in Human Behavior, 2012. (Cité en pages 5 et 70.)

[Stehle 2011] J. Stehle, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.F. Pinton, M. Quaggiotto, W. Van Den Broeck, C. Regis, B. Lina*et al. High-resolution measurements of face-to-face contact patterns in a primary school*. PloS one, vol. 6, no. 8, 2011. (Cité en page 38.)

[Steinhaeuser 2010] Karsten Steinhaeuser et Nitesh V. Chawla. *Identifying and evaluating community structure in complex networks*. Pattern Recogn. Lett., vol. 31, pages 413–421, 2010. (Cité en page 110.)

[Tian 2008] Y. Tian, R.A. Hankins et J.M. Patel. *Efficient aggregation for graph summarization*. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data, pages 567–580. ACM, 2008. (Cité en page 111.)

[Toivonen 2009] R. Toivonen, L. Kovanen, M. Kivela, J.P. Onnela, J. Saramaki et K. Kaski. *A comparative study of social network models : network evolution models and nodal attribute models*. Social Networks, vol. 31, 2009. (Cité en pages 3, 46, 54, 55 et 62.)

[Vaillant 2011] J. Vaillant, G. Puggioni, L. Waller et J.-H. Daugrois. *A spatio-temporal analysis of the spread of sugar cane yellow leaf virus*. Journal of Time Series Analysis, vol. 32, pages 396–406., 2011. (Cité en page 88.)

[Volz 2007] Erik Volz et Lauren Ancel Meyers. *Susceptible-infected-recovered epidemics in dynamic contact networks*. Proceedings of the Royal Society B Biological Sciences, vol. 274, no. 1628, pages 2925–2933, 2007. (Cité en page 46.)

[Wallace 1991] Rodrick Wallace. *Social disintegration and the spread of AIDS : Thresholds for propagation along 'sociogeographic' networks*. Social Science & Medicine, vol. 33, no. 10, pages 1155 – 1162, 1991. (Cité en pages 29 et 47.)

[Wang 2003] J. Wang, H. Zeng, Z. Chen, H. Lu, L. Tao et W.Y. Ma. *ReCoM : reinforcement clustering of multi-type interrelated data objects*. In Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, pages 274–281. ACM, 2003. (Cité en page 111.)

[Watts 1998] D. J. Watts et S. H. Strogatz. *Collective dynamics of 'small-world' networks*. Nature, vol. 393, pages 440–42, 1998. (Cité en pages 20, 33, 46, 54 et 62.)

[Watts 2004] Duncan J. Watts. *The 'New' Science of Networks*. Annual Review of Sociology, vol. 30, pages 243–270, 2004. (Cité en page 2.)

[West 2000] Douglas B. West. Introduction to graph theory (2nd edition). Prentice Hall, 2000. (Cité en pages 2 et 12.)

[(WHO) 2007] World Health Organization (WHO). *International spread of disease threatens public health security*. 2007. (Cité en page 104.)

[Wierman 2004] J.C. Wierman et D.J. Marchette. *Modeling computer virus prevalence with a susceptible-infected-susceptible model with reintroduction*. Computational statistics & data analysis, vol. 45, no. 1, pages 3–23, 2004. (Cité en pages 29 et 48.)

[Wilensky 1999] U. Wilensky. *Center for Connected Learning and Computer-Based Modeling*, 1999. http ://ccl.northwestern.edu/netlogo/. (Cité en page 83.)

[Wilensky 2009] U. Wilensky et W. Rand. *An introduction to agent-based modeling : Modeling natural, social and engineered complex systems with NetLogo*. 2009. (Cité en page 83.)

[Wille 1984] R. Wille. *Line diagrams of hierarchical concept systems*. International classification, vol. 11, no. 2, pages 77–86, 1984. (Cité en page 111.)

[Yan 2002] X. Yan et J. Han. *gSpan : Graph-Based Substructure Pattern Mining*. In Proceedings of the 2002 IEEE International Conference on Data Mining, 2002. (Cité en page 110.)

[Yoneki 2008] Eiko Yoneki, Pan Hui et Jon Crowcroft. *Bio-Inspired Computing and Communication*. chapitre Wireless Epidemic Spread in Dynamic Human Networks, pages 116–132. Springer-Verlag, Berlin, Heidelberg, 2008. (Cité en page 44.)

[Yoon 2011] Seok-Ho Yoon, Suk-Soon Song et Sang-Wook Kim. *Efficient link-based clustering in a large scaled blog network*. In Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication, ICUIMC '11, pages 71 :1–71 :5, New York, NY, USA, 2011. ACM. (Cité en pages 110 et 111.)

[Zanette 2001] D.H. Zanette. *Critical behavior of propagation on small-world networks*. Physical Review E, vol. 64, no. 5, page 050901, 2001. (Cité en page 48.)

[Zanette 2002] D.H. Zanette. *Dynamics of rumor propagation on small-world networks*. Physical review E, vol. 65, no. 4, 2002. (Cité en pages 1, 29 et 48.)

[Zhou 2009] Y. Zhou, H. Cheng et J.X. Yu. *Graph clustering based on structural/attribute similarities*. VLDB Endowment, vol. 2, no. 1, pages 718–729, 2009. (Cité en pages 107, 108 et 111.)

Contributions à l'étude des réseaux sociaux :
propagation, fouille, collecte de données

Résumé. Le concept de réseau offre un modèle de représentation pour une grande variété d'objets et de systèmes, aussi bien naturels que sociaux, dans lesquels un ensemble d'entités homogènes ou hétérogènes interagissent entre elles. Il est aujourd'hui employé couramment pour désigner divers types de structures relationnelles. Pourtant, si chacun a une idée plus ou moins précise de ce qu'est un réseau, nous ignorons encore souvent les implications qu'ont ces structures dans de nombreux phénomènes du monde qui nous entoure. C'est par exemple le cas de processus tels que la diffusion d'une rumeur, la transmission d'une maladie, ou même l'émergence de sujets d'intérêt commun à un groupe d'individus, dans lesquels les relations que maintiennent les individus entre eux et leur nature s'avèrent souvent être les principaux facteurs déterminants l'évolution du phénomène. C'est ainsi que l'étude des réseaux est devenue l'un des domaines émergents du 21e siècle appelé la "Science des réseaux". Dans ce mémoire, nous abordons trois problèmes de la science des réseaux : le problème de la diffusion dans les réseaux sociaux, où nous nous sommes intéressés plus particulièrement à l'impact de la dynamique du réseau sur le processus de diffusion, le problème de l'analyse des réseaux sociaux, dans lequel nous avons proposé une solution pour tirer parti de l'ensemble des informations disponibles en combinant les informations sur la structure du réseau et les attributs des noeuds et le problème central de la collecte de données sociales, où nous nous sommes intéressés au cas particulier de la collecte de données en milieux sauvages.

Mots-clés : Réseaux sociaux, percolation, diffusion, modélisation, dynamique des réseaux, analyse des réseaux sociaux, fouilles des réseaux sociaux, collecte de données

---

Contributions to the study of social networks :
propagation, mining, data collection

Abstract. The concept of network provides a model for representing a wide variety of objects and systems, both natural and social, in which a set of homogeneous or heterogeneous entities interact. It is now widely used to describe various kinds of relational structures. However, if everyone has an idea of the concept of network, we often ignore the implications that these structures have in real world phenomena. This is for example the case of processes such as the spread of a rumor, the disease transmission, or even the emergence of subjects of common interest for a group of individuals, in which the relations maintained between individuals, and their nature, often prove to be the main factors determining the evolution of the phenomenon. This is the reason why the study of networks has become one of the emerging areas in the 21st century called the "Science of networks." In this thesis, we address three issues of the domain of the science of networks : the problem of diffusion in social networks, where we have addressed more particularly the impact of the network dynamics on the diffusion process, the problem of the analysis of social networks, in which we have proposed a solution to take full advantage of all information available on the network by combining information on both structure and node attributes and the central problem of the social data collection, for which we have focused on the particular case of the data collection in a wild environment.

Keywords : Social networks, percolation, diffusion, modelisation, network dynamics, social network analysis, social network mining, data collection