De la manipulation dans les systèmes multi-agents : une étude sur les jeux hédoniques et les systèmes de réputation

Thibaut Vallée

Université de Caen Basse-Normandie

U.F.R. : Sciences

Ecole doctorale : *SIMEM*

Thèse de doctorat

présentée

par

Thibaut VALLEE

et soutenue

le 9 décembre 2015

en vue de l'obtention du

Doctorat de l'Université de Caen Normandie

Spécialité : *Informatique et Applications*

(Arrêté du 7 août 2006)

De la manipulation dans les systèmes multi-agents : une étude sur les jeux hédoniques et les systèmes de réputation

MEMBRES du JURY

Nicolas MAUDET	Professeur des Universités	U. Pierre et Marie Curie	(rapporteur)
Laurent VERCOUTER	Professeur des Universités	INSA de Rouen	(rapporteur)
Zahia GUESSOUM	Maître de conférences	U. Pierre et Marie Curie	
Grégory BONNET	Maître de conférences	U. Caen Normandie	
François BOURDON	Professeur des Universités	U. Caen Normandie	(directeur)

Remerciements

Cette thèse a été pour moi une grande expérience tant sur le plan scientifique qu'humain. Elle résulte avant tout d'une collaboration et c'est pourquoi je tiens à remercier tous ceux et celles qui, de près comme de loin, m'ont aidé et soutenu pendant ces trois années. À tous celles et ceux dont le nom n'apparait pas sur cette page malgré toute l'aide que vous m'avez apportée, je m'en excuse et je tiens à vous dire merci.

La première personne que je tiens à remercier est François Bourdon, mon directeur de thèse. Dès ma première année de DUT, tu as su m'amener à voir plus loin et à me poser les bonnes questions. Merci pour tes conseils avisés et d'avoir partagé avec moi ton point de vue original. Un grand merci à Grégory Bonnet, pour son encadrement et ses conseils. Merci Grégory, pour la confiance que tu as eu en moi, ton soutien et ton amitié m'ont été d'une aide précieuse durant ces trois années. Sans toi, cette thèse n'aurait pas été la même.

Je remercie Zahia Guessoum d'avoir accepté d'être examinatrice de ma thèse. Je remercie également mes rapporteurs Laurent Vercouter et Nicolas Maudet pour l'intérêt qu'ils ont porté à mes travaux et dont le point de vue extérieur m'a permis d'envisager mon travail sous un autre angle. Merci particulièrement à Nicolas pour son accueil lors de FairDiv2015.

Mes remerciements vont maintenant à l'ensemble du personnel du GREYC. Merci particulièrement à Bruno Zannutini, un directeur d'équipe exceptionnel, pour tous les conseils et l'aide qu'il m'a apporté. Merci également à Agnès, Arielle, Davy, Edwige, Pierre, Renaud, Véronique et Virginie pour la qualité de leur travail et sans qui le GREYC ne fonctionnerait pas si bien. Je remercie aussi tout ceux avec qui j'ai partagé le bureau 368 : Henri, Krystian et plus récement Florian, mais aussi Arthur, Esther, Guillaume, Jennifer et Nicolas pour venir avec votre bonne humeur perturber le travail du mardi. Merci à Laurent pour le temps passé à satisfaire les préférences des uns et des autres dans l'emploi du temps de l'IUT.

Ces trois années ont été une expérience riche tant sur le plan scientifique qu'humain. Merci donc à Boris, Céline, Charlotte, Cyril, Gaël, Grégory, Jean-Philippe, Mathieu, Romain pour ces soirées catu-patoch, jorky, DDR, netrunner ... Merci notamment à Jean-Philippe pour m'avoir amené à tester la stochocratie et à Charlotte pour toute l'aide que tu m'as apportée dans cette fin de thèse.

Je remercie également Marian, Margaux, Jonas pour nos nombreux débats (pas toujours très constuctifs) sur le cours du café en Azerbaïdjan et sur la reproduction des crevettes naines albinos du Honduras septentrional. Merci à Raphy, notamment pour tes comparaisons toujours très imagées. Merci à Benjamin et Léo pour m'avoir supporté comme colocataire durant deux années. Merci également aux colocataires du dimanche : T.Salles, Maël, Anne-Marie pour avoir participé à la vie de l'appartement.

Bien qu'il s'agisse d'une espèce en voie d'extinction, je remercie tous les Bruéroliens qui m'ont soutenu et notamment Franck pour partager sans langue de bois ton point de vue éclairé sur le monde. Merci à tous les membres d'Éphémère, ces rendez-vous hebdomadaires sont toujours des moments de détente agréable et conviviaux. Une pensée à la fine équipe des Grilles, Lucile, Pierre, Po, Manon, avec vous j'ai partagé plus qu'une passion. Merci à vous quatre de partager cette joie de vivre qui vous caractérise.

Amandine, merci à toi pour ces 25 années d'inestimable amitié et ce malgré la distance.

Enfin, je dédicace cette thèse à mes parents, ma sœur et mon frère qui m'ont encouragé et soutenu pendant toutes ces années.

Table des matières

Introduction

Contexte

Durant les dernières années, le développement informatique a évolué du logiciel conçu pour résoudre individuellement des tâches spécifiques à des logiciels autonomes interconnectés, appelés *agents artificiels* (ou plus simplement *agents*). Les agents partagant un même environnement interagissent au sein d'un *système multi-agent*. À partir de connaissances qui leur sont propres et de connaissances communes, ces agents doivent prendre des décisions afin d'atteindre leurs objectifs respectifs. Ces objectifs peuvent être communs ou, au contraire, en opposition. C'est pourquoi les interactions entre les agents sont régies par des règles formant un protocole. Cependant, le respect de ce protocole amène parfois des agents à être insatisfaits car certaines décisions prises collectivement peuvent se faire à l'encontre d'objectifs individuels.

Forcés de respecter le protocole, certains de ces agents insatisfaits peuvent alors désirer manipuler le système ou les autres agents pour en tirer profit. Ces agents manipulateurs propagent de fausses informations, usurpent l'identité d'un autre, interceptent les communications d'agents tiers ou mettent en œvre des stratégies plus complexes afin d'altérer les connaissances des autres agents et de les inciter à prendre des décisions qui leur seraient plus favorables. Cela n'aurait aucune conséquence si les manipulations ne se faisaient généralement pas au détriment des agents honnêtes.

Pour garantir à ces agents honnêtes qu'ils peuvent participer à un système multi-agent sans être victimes de manipulations, il est important de définir des stratégies de défense. Celles-ci doivent assurer que le système est robuste aux manipulations sans pour autant remettre en cause ses propriétés désirées. Malheureusement, vouloir satisfaire à la fois des propriétés comme l'absence de dictature (un unique agent du système décide pour tous les autres) ou l'ouverture (la possibilité pour un agent de rejoindre ou quitter sans coût le système) tout en garantissant une robustesse à toutes les manipulations est mis à mal par des théorèmes d'impossibilité. Il est donc important que les stratégies de défense mises en œuvre soient les plus performantes possibles au regard des limites théoriques qu'elles ont.

En effet, l'existance de manipulations se fonde sur certaines propriétés du système multi-agent considéré. Par exemple, l'absence d'une autorité centrale qui pourrait vérifier des informations partagées ou encore la volonté de construire des systèmes ouverts et interconnectés permettent l'introduction de fausses identités. Dans ce manuscrit, nous proposons d'étudier les manipulations au regard des propriétés dont elles ont besoin pour être mises en œuvre de manière efficace ainsi que les stratégies de défense qui viennent renforcer ou affaiblir ces propriétés en fonction de leur influence sur les manipulations.

Contributions

Dans ce manuscrit, nous faisons porter notre étude sur deux familles de systèmes multi-agents : les jeux hédoniques où des agents doivent décider collectivement comment se répartir en sous-ensembles d'agents afin d'atteindre leurs objectifs respectifs, et les systèmes de réputation où des agents partagent des informations et décident individuellement à quels autres agents ils peuvent faire confiance. Le choix de ces deux familles de systèmes reposent sur la complémentarité de leurs propriétés qui nous permettent d'aborder chaque point de notre problématique. En effet, les jeux hédoniques sont des systèmes statiques dans le temps qui impliquent une décision collective au regard d'utilités ordinales. Les systèmes de réputation sont des systèmes dynamiques dans le temps qui impliquent des décisions individuelles au regard d'utilités cardinales.

De plus, chaque famille nous permet d'aborder un point de notre problématique : caractériser les propriétés nécessaires à la mise en œuvre de manipulations et identifier des stratégies de défense au regard de ces propriétés. En effet, à notre connaissance, la question des manipulations dans les jeux hédoniques n'a pas encore été étudiée et nous présentons donc dans ce manuscrit une première étude de leur robustesse avec pour spécificité de se placer du point de vue des agents manipulateurs. Dans le contexte des systèmes de réputation, de nombreux travaux antérieurs ont mis en lumière les propriétés fondamentales de ces systèmes qui étaient mises en défaut par les agents manipulateurs et nous proposons deux stratégies de défense qui n'affaiblissent pas ces propriétés.

Nos contributions ont fait l'objet de plusieurs publications dans des conférences internationales [Vallée *et al.*, 2014c, Vallée *et al.*, 2014b, Vallée et Bonnet, 2015], nationales [Vallée *et al.*, 2013, Vallée *et al.*, 2014a] et dans une revue nationale [Vallée *et al.*, 2015]. Dans le cadre des jeux hédoniques, nous proposons :

1. une définition générique des manipulations dans un tel contexte. Cette définition d'une manipulation repose sur deux points : le partage de *fausses informations* et l'introduction dans le jeu de multiples fausses identités, appelées *agents Sybil*. Pour ces manipulations, nous considérons la notion de rationalité représentant le fait que sa mise en œuvre amène le système dans un autre jeu plus favorable pour l'agent manipulateur ;
2. une caractérisation formelle des conditions nécessaires sur la structure du jeu autorisant une manipulation rationnelle, et cela sur trois concepts de solution : la *stabilité au sens de Nash*, la *stabilité individuelle* et la *stabilité au sens du cœur*. Nous montrons empiriquement que la stabilité au sens de Nash est un concept de solution robuste aux manipulations contrairement à la stabilité individuelle et à la stabilité au sens du cœur. Cette robustesse repose sur la complexité algorithmique de décider si une manipulation est rationnelle et sur la fréquence de satisfaction des conditions nécessaires à sa rationalité.

Dans le cadre des systèmes de réputation, nous proposons :

1. d'étudier l'utilisation des valeurs de réputation dans le processus de décision des agents. Pour cela, nous modélisons le problème de décision des agents dans un système de réputation en considérant des outils issus d'un autre domaine de recherche : celui des *bandits manchots* et de leurs politiques d'apprentissage par renforcement. Nous analysons empiriquement les forces et faiblesses de plusieurs politiques de décision couramment employées afin d'éclairer un concepteur sur le choix le plus judicieux à faire ;
2. de détecter et filtrer les *faux témoignages*. Pour cela, nous proposons une nouvelle *mesure de crédibilité* des témoignages fondée sur la divergence de Kullback-Leibler, mesurant un gain d'information à croire un témoignage. Nous proposons ensuite trois méthodes permettant d'écarter les témoignages jugés comme non crédibles dans le processus de décision.

Ces différents mécanismes de défense sont soumis à des évaluations empiriques permettant de montrer leur efficacité contre des stratégies oscillatoires employées par des agents manipulateurs.

Organisation du document

Ce manuscrit est organisé en trois grandes parties : un état de l'art, des contributions sur les jeux hédoniques, des contributions sur les systèmes de réputation.

La première partie est consacrée à l'état de l'art. Il s'agit ici de présenter les principales notions que nous utilisons dans la suite de ce manuscrit. Dans le chapitre 1, nous présentons dans leur globalité les systèmes multi-agents, en particulier les systèmes multi-agents ouverts décentralisés où des agents égoïstes doivent prendre des décisions rationnelles. Nous présentons ensuite le problème des manipulations et de nombreuses formes qu'elles peuvent prendre. Nous présentons enfin les principales techniques de défense utilisées dans la littérature. Le chapitre 2 est une présentation détaillée des deux familles de systèmes multi-agents que nous considérons dans ce manuscrit : les jeux hédoniques et les systèmes de réputation.

La seconde partie de ce document est entièrement consacrée à nos contributions sur les jeux hédoniques. Dans le chapitre 3, nous présentons un modèle de jeu hédonique et un modèle de manipulation, en particulier la notion de rationalité d'une manipulation. Le chapitre 4 est une étude de la robustesse des jeux hédoniques utilisant la stabilité au sens de Nash comme concept de solution. Dans le chapitre 5, nous remettons en cause l'une des hypothèses utilisées dans le chapitre précédent et étendons notre étude à d'autres concepts de solution.

La troisième partie de ce manuscrit est dédiée aux problèmes des manipulations dans les systèmes de réputation. Dans le chapitre 6, nous présentons un modèle générique d'interactions entre agents utilisant un système de réputation abstrait pour estimer le comportement futur des autres agents ainsi qu'un ensemble de manipulations sur un tel système. Dans le chapitre 7, nous montrons par analogie comment un agent peut utiliser les politiques de bandits manchots afin de décider à qui faire confiance. Une étude empirique illustre les forces et faiblesses de différentes politiques. Dans le chapitre 8, nous proposons une nouvelle mesure de crédibilité qui se fonde sur la divergence de Kullback-Leibler pour évaluer la qualité des témoignages échangés. Nous présentons ensuite trois fonctions de filtrage écartant les faux témoignages et évaluons leur efficacité à l'aide d'une étude empirique.

Enfin, nous clôturons ce manuscrit par le chapitre 9 dans lequel nous présentons un synthèse de notre travail ainsi que plusieurs pistes pour des travaux futurs permettant d'étendre nos résultats. À titre d'exemple, si nous caractérisons dans ce manuscrit des stratégies de manipulation qui peuvent êre mise en œuvre par des agents malhonnêtes ou malveillants, il pourrait être intéressant de se demander si un agent honnête ne pourrait lui-même pas se servir de ces caractérisation pour mettre en œuvre des stratégies de manipulation à l'encontre des agents manipulateurs.

Première partie

État de l'art

Chapitre 1

Des manipulations dans les systèmes multi-agents

Sommaire

Résumé.

Ce chapitre a pour objectif d'introduire le contexte général de la thèse. Dans un premier temps, nous présentons les systèmes multi-agents et les différents contextes d'utilisation de ces systèmes. Nous présentons ensuite le problème de manipulation dans le cadre des systèmes multi-agents et les principales approches proposées pour empêcher leurs manipulations. Nous concluons ce chapitre en nous positionnant par rapport à ces différentes approches.

1.1 Systèmes multi-agents

1.1.1 Des agents dans un environnement

Si la notion d'agent est largement répandue en intelligence artificielle, il en existe de nombreuses définitions [Russell et Norvig, 1995, Jennings et Wooldridge, 1996, Franklin et Graesser, 1997, Serenko et Detlor, 2004, Panait et Luke, 2005]. Par exemple, selon [Jennings et Wooldridge, 1996], un agent est une entité autonome capable de contrôler son processus de décision et d'agir à partir de sa perception de son environnement, et ce afin d'atteindre son ou ses objectifs. [Panait et Luke, 2005] considèrent quant à eux un agent comme un programme informatique exhibant un haut degré d'autonomie (c'est-à-dire capable de prendre seul des décisions), tout en étant capable d'effectuer des actions dans son environnement à partir des informations qu'il perçoit de ce dernier.

Ces définitions, comme beaucoup d'autres, comportent trois caractéristiques principales. Un agent est doté :

- d'un processus de décision autonome ;
- d'une perception de son environnement ;
- d'une capacité d'action.

La principale caractéristique d'un agent est qu'il s'agit d'une entité autonome, c'est-à-dire qu'un agent est capable de prendre seul des décisions afin de réaliser des buts donnés a priori. Lorsque [Panait et Luke, 2005] considèrent un *haut degré d'autonomie*, cela exprime le fait qu'un agent peut, sous certaines conditions, être amené à suivre une décision venant d'une autre entité, qu'elle soit une décision d'un opérateur ou utilisateur humain, d'un autre agent ou d'un collectif d'agents. [Bradshaw *et al.*, 2003] représentent le degré d'autonomie d'un agent comme l'influence que peut exercer une autre entité sur le processus de décision de l'agent. Remarquons que la notion d'autonomie est plus large. Par exemple, l'autonomie d'un agent est considérée comme *ajustable* si ce dernier peut faire varier son degré d'autonomie. De plus, si les définitions précédentes considèrent qu'un agent dispose d'un ou plusieurs objectifs fixés a priori, [Luck et d'Inverno, 2001] considèrent l'autonomie d'un agent comme sa capacité à définir ses propres objectifs. D'autres approches, telles que celle de [Conte *et al.*, 1999], considèrent l'autonomie d'un agent par le fait qu'il puisse générer ses propres règles sociales et décider, éventuellement, de ne pas les respecter.

La seconde caractéristique d'un agent est qu'il est capable de percevoir des informations de son environnement. La perception d'un agent définit l'ensemble des informations dont il dispose en provenance l'environnement dans lequel il évolue. Celle-ci est caractérisée par sa complétude (ou incomplétude) et sa certitude (ou incertitude). La complétude et l'incomplétude permettent de représenter le fait que l'agent connaisse ou non en temps réel l'ensemble des éléments qui constituent son environnement. La certitude et l'incertitude permettent de représenter un degré de fiabilité dans les éléments perçus par l'agent. Par exemple, un robot peut connaître la carte du monde dans lequel il évolue, mais être incertain quant à sa position exacte, suite à des défauts de capteurs ou encore à un décalage temporel entre l'émission et la réception de l'information. L'environnement, quant à lui, est le contexte dans lequel l'agent évolue. Il regroupe un ensemble d'éléments sur lesquels l'agent peut agir. Cet environnement peut être physique (par exemple, il peut s'agir de la pièce dans laquelle un robot se déplace [Kuipers et Byun, 1991]) ou virtuel (telle qu'une place de marché pour des agents de trading haute fréquence [Dasgupta *et al.*, 1999]).

Enfin, un agent est capable d'agir, c'est-à-dire qu'il peut effectuer des actions qui vont modifier son état et son environnement. Les actions d'un agent ont pour but de réaliser ses objectifs, c'est-à-dire ce pour quoi il a été conçu [Jennings et Wooldridge, 1996]. Un agent peut avoir

plusieurs objectifs, parfois contradictoires, et son processus de décision doit décider de l'action ou suite d'actions à exécuter dans ces circonstances.

Au vu de ces caractéristiques, nous considérons la définition d'un agent suivante.

Définition 1.1.1 - *Agent* : Un *agent* désigne une entité munie d'un processus de décision, capable de percevoir et d'agir et d'interagir dans et avec son environnement (physique ou virtuel) afin de réaliser son ou ses objectifs.

Plusieurs agents peuvent partager le même environnement. Nous parlons alors de systèmes multi-agents. Les agents disposent alors de ressources propres, d'informations privées et de capacités de communication [Ferber, 1999]. C'est le fait d'avoir des ressources propres (capacités de calcul ou de mémorisation, actionneurs différents, etc.) et d'informations privées (perceptions différentes, connaissances spécifiques, etc.) qui rend les agents autonomes, à la fois vis-à-vis de leur environnement et des autres agents. Cette autonomie est d'autant plus renforcée par leur capacité à communiquer et à partager avec les autres agents une partie de leurs connaissances privées. Par ailleurs, plusieurs agents peuvent partager leurs ressources afin de réaliser collectivement une même tâche ou déléguer à un agent tiers sa réalisation. On parle alors d'*interaction entre les agents*.

Si [Ferber, 1999] décrit l'environnement d'un système multi-agents comme un ensemble d'entités actives (les agents) et d'entités passives (appelées artefacts) sur lesquels les agents peuvent agir, la notion d'environnement est plus large. Par exemple, [Parunak, 1997] considère l'environnement comme un tuple $\langle Etats, Processus \rangle$, où $Etats$ désigne l'ensemble des éléments de l'environnement (incluant les agents) et $Processus$ désigne un ensemble de processus influant sur l'état de l'environnement. Cette modélisation permet de considérer des facteurs externes aux agents qui peuvent modifier l'état de l'environnement. [Weyns *et al.*, 2007] présentent de multiples définitions et propriétés de l'environnement dans le cadre des systèmes multi-agents. Il décrit trois niveaux de structure de l'environnement :

- structure physique : topologie spatiale des agents et des artefacts ;
- structure communicationnelle : infrastructure permettant le transfert de messages explicites et/ou modification de l'environnement pour des communications indirectes ;
- structure sociale : organisation des agents et des artefacts en fonction de leurs rôles.

Au vu de ces définitions, nous considérons la définition d'un système multi-agents suivante :

Définition 1.1.2 - *Système multi-agents* : Un *système multi-agents* (SMA) désigne un système composé d'un ensemble N d'agents partageant le même environnement et où les agents

- ont des objectifs individuels ou collectifs ;
- disposent de ressources individuelles ;
- disposent d'informations privées ;
- communiquent entre eux ;
- ont une perception (complète ou incomplète, sûre ou non sûre) de leur environnement ;
- agissent sur les éléments de leur environnement ;
- interagissent entre eux pour atteindre leurs objectifs.

1.1.2 Différents types de systèmes

Les systèmes multi-agents peuvent servir pour diverses applications. Nous distinguons trois grandes catégories : la simulation multi-agents, la résolution coopérative de tâches et la régulation de systèmes ouverts.

Simulation multi-agents

La simulation multi-agents consiste à reproduire dans un environnement contrôlé [Railsback *et al.*, 2006] des phénomènes biologiques [Resnick, 1994], sociologiques [Drogoul et Ferber, 1994], économiques [Janssen et Jager, 2003] ou autres. L'objectif de la simulation est de permettre d'observer l'évolution de cet environnement en y faisant varier certains paramètres, et ainsi de prédire l'influence de ces paramètres dans un environnement non contrôlé.

Les systèmes multi-agents permettent en effet de modéliser des systèmes complexes dans lesquels des entités autonomes interagissent. Dans cette approche, l'environnement est défini par un ensemble de propriétés que le concepteur suppose être vraies. Le comportement des agents est décrit par un ensemble de règles simples, influencées par l'environnement et les actions des autres agents. Les enjeux principaux de la simulation multi-agents sont : la prédiction de l'influence des paramètres sur les processus de décision [Barsalou, 2009], la vérification et la validation d'hypothèses émises sur un modèle [Kleijnen, 1995], l'étude de comportements et de phénomènes émergents [Gilbert, 1995].

Résolution coopérative de tâches

Dans la résolution coopérative de tâches, les agents sont conçus pour atteindre un objectif commun et ils coopèrent en répartissant ressources, informations et sous-objectifs entre eux. Ces systèmes sont dits *systèmes coopératifs*. Un cas classique de système coopératif est celui de la RobotCup Rescue [Kitano *et al.*, 1999] où de multiples agents doivent se coordonner et s'affecter des sous-objectifs distincts afin de sauver un maximum de civil dans une situation de catastrophe naturelle.

La résolution coopérative de tâches s'intéresse à la modélisation et la distribution des sous-objectifs entre les différents agents afin qu'ils puissent résoudre de manière distribuée et/ou décentralisée un problème. Ces systèmes utilisent des protocoles d'affectation de tâches, comme ContractNet [Smith, 1980, Sandholm, 1993], des techniques de satisfaction de contraintes distribuées [Matsui *et al.*, 2008] ou des techniques de maintien de plans d'actions cohérents et efficaces [Pinson et Moraitis, 1997, Shen *et al.*, 2006, Bernstein *et al.*, 2000].

Une autre problématique de la résolution coopérative de tâches est celle du partage d'informations. Ce problème est abordé par exemple par [Seuken et Zilberstein, 2008, Renoux *et al.*, 2014] dans le cadre des DEC-POMPD où les agents doivent décider de l'utilité à transmettre certaines informations, utilité fondée sur le gain que l'information apporte aux agents qui pourraient la recevoir.

Régulation de systèmes ouverts

Si les systèmes coopératifs permettent de résoudre un problème multi-agent en faisant l'hypothèse que les agents sont conçus ensemble et partagent des modèles, des méthodes de résolution et des objectifs communs de haut niveau, il est également possible de concevoir des systèmes multi-agents où ceux-ci ne se pas conçus de concert et peuvent partager les mêmes objectifs sans le savoir ou encore avoir des objectifs opposés et en compétition. De tels systèmes sont appelés *systèmes ouverts* car en raison de cette hétérogénéité dans la conception, il est fait l'hypothèse que les agents peuvent à tout instant rejoindre ou quitter le système [Huynh *et al.*, 2006].

Ainsi, dans un système multi-agents ouvert, aucun agent ne peut disposer d'une perception parfaite et complète de l'environnement et il ne peut pas y avoir d'autorité centrale capable de contrôler la totalité des agents [Huynh *et al.*, 2006]. De même, le manque de prédictibilité du

comportement des agents amène ne pouvoir modéliser l'environnement que comme une infrastructure définissant un protocole de communication et d'interaction commun [Mazouzi *et al.*, 2002]. Ainsi, les seules hypothèses qu'un agent peut faire sur un autre agent afin de tenir compte de son comportement est que ce dernier est *pleinement autonome* et *rationnel*.

Un agent pleinement autonome est un agent dont les décisions ne peuvent pas être dictées par des entités extérieures, sans pour autant ne tenir pas compte de ces entités dans sa décision. Un agent rationnel est un agent qui prend la meilleure décision en fonction de ses informations et de ses objectifs personnels [Wooldridge, 2009].

Un agent pleinement autonome et rationnel peut être amené à coopérer temporairement avec d'autres agents même si certains de leurs objectifs sont opposés [Jensen et Meckling, 1994]. C'est le cas par exemple dans les *systèmes de recommandation* [Balabanović et Shoham, 1997] où les agents disposent d'informations privées sur des artefacts (par l'intermédiaire d'utilisations antérieures par exemple) et partagent ces informations avec les autres afin d'évaluer au mieux l'ensemble des artefacts du système. Ce partage d'informations privées est appelé un *témoignage*. Le système de recommandation permet alors d'ordonner les artefacts en fonction de leur utilité supposée pour chaque agent grâce à des mécanismes d'agrégation de témoignages [Breese *et al.*, 1998], de similitude entre artefacts [Basu *et al.*, 1998] ou de proximité sociale entre les agents [Walter *et al.*, 2008]. Certains systèmes de recommandation sont appelés *systèmes de réputation* et permettent aux agents d'évaluer non plus les artefacts, mais les agents eux-mêmes en fonction de leurs interactions passées. En partageant des témoignages sur leurs *confiances* (estimation personnelle des autres agents) respectives et en agrégeant ces dernières les agents définissent une valeur de *réputation* (estimation collective) sur les autres agents.

Ces deux notions, pleine autonomie et rationalité, ont toutes deux une influence importante sur l'une des caractéristiques fondamentales d'un agent : son processus de décision [Faratin *et al.*, 1998, Olfati-Saber *et al.*, 2007].

1.1.3 Décisions pour des agents rationnels

Dans un système multi-agents ouvert, le processus de décision des agents peut être influencé par le choix des autres agents. Cela conduit à considérer le problème de décision des agents de deux points de vue : un point de vue individuel et un point de vue collectif. Du point de vue individuel, l'agent doit prendre une décision à partir de ses connaissances du système et de ses connaissances sur les autres agents. Du point de vue collectif, les agents doivent prendre une décision coordonnée en faisant un compromis entre les objectifs de chaque agent.

Décision individuelle

Une des manières d'aborder la question de la décision individuelle consiste à s'intéresser au cadre de la théorie des jeux. La théorie des jeux considère le problème de décision comme un jeu stratégique entre agents rationnels [von Neumann et Morgenstern, 1944].

Classiquement, on représente un jeu stratégique comme un triplet $\langle N,\Sigma,\mathcal{U}\rangle$ où N désigne l'ensemble des agents, Σ l'ensemble des actions possibles pour chaque agent et U l'ensemble des fonctions d'utilité des agents. Le processus de décision de l'agent consiste alors à calculer une *stratégie*, c'est-à-dire un plan d'action qu'il va devoir mettre en œuvre. La fonction d'*utilité* de chaque agent est une fonction qui associe à chaque option possible une valeur réelle indiquant son degré de satisfaction si cette option se produit. Ce degré de satisfaction est appelé la *récompense* de l'agent.

Un agent rationnel calcule alors sa stratégie afin de maximiser sa récompense à long terme,

appelé le *gain* [Binmore *et al.*, 1998]. La rationalité des agents doit conduire à des solutions dites d'équilibres, c'est-à-dire des situations où la solution est acceptable par tous les agents. Il existe de nombreuses notions d'équilibre [Harsanyi et Selten, 1988]. Par exemple, l'équilibre de Nash ([Nash, 1950]) correspond à une situation où individuellement, en connaissant les décisions des autres, aucun agent n'a d'intérêt à changer de décision. Si ces concepts de solution permettent de garantir certaines propriétés telles que l'optimalité de la solution vis-à-vis du bien-être social (gains cumulés de l'ensemble des agents), nombre de jeux ne disposent pas de solutions les respectant.

L'un des exemples classiques de la théorie des jeux est le dilemme du prisonnier [Poundstone *et al.*, 1993]. La table 1.1 est un exemple de matrice représentant les coûts (fonction d'utilité) associés aux décisions des agents.

		Prisonnier 1	
		Se tait	Dénonce
Prisonnier 2	Se tait	1 1	0 7
	Dénonce	7 0	5 5

Tableau 1.1 – Une matrice des coûts pour le dilemme du prisonnier

Il existe de nombreux modèles de jeux [Myerson, 2013]. Par exemple, les jeux à somme nulle permettent de considérer des situations où les objectifs des agents sont strictement opposés [Duersch *et al.*, 2012] ou encore les jeux dans lesquels les fonctions d'utilité sont remplacées par des *profils de préférence* désignant une relation d'ordre sur l'ensemble des solutions possibles [Osborne et Rubinstein, 1994]. La théorie des jeux permet aussi de modéliser des systèmes d'enchères où chaque agent doit décider du prix à payer pour obtenir une ressource [Guttman et Maes, 1998, Sandholm, 2002].

Décision collective

Même si les agents pleinement autonomes et rationels prennent des décisions individuellement, il leur est parfois nécessaire de prendre des décisions collectives. Par exemple, décider d'une répartition de ressources entre tous les agents du système. Ce problème de prise de décision collective est étudié dans le cadre de la théorie des choix sociaux [Sen, 1986, Coleman, 1986].

Classiquement, un problème de choix social se présente par un ensemble d'agents qui doivent prendre une décision parmi un ensemble d'options possibles. Pour ce faire, les agents définissent un *profil de préférence*, une relation d'ordre sur l'ensemble des options possibles. Ce profil peut éventuellement provenir de fonctions d'utilité. Ce cadre permet de capturer des problèmes divers comme la formation de coalition (les agents doivent décider collectivement avec qui coopérer) ou l'affectation de ressources (les agents doivent décider du compromis entre répartition équitable et efficace des biens). Dans tous les cas, les agents doivent alors se mettre d'accord pour définir un ordre commun sur l'ensemble des options et de nombreux mécanismes définissant cet ordre existent [Chevaleyre *et al.*, 2007] :

– Les systèmes de votes dans lesquels des règles connues de tous les agents agrègent les profils de préférence pour extraire un profil global qui respecte au mieux un certain nombre d'axiomes désirés [Arrow, 1963, Bartholdi III et Orlin, 1991, de Condorcet, 1785, Young, 1995]. Il existe de nombreuses règles de vote [Brams et Fishburn, 2002,Tideman, 2006,Oo et Aung, 2014] comme le vote majoritaire [Coughlin, 1982] ou le système de Borda [de Borda, 1781] ;
– Les protocoles de négociation dans lesquels les agents vont échanger itérativement des propositions satisfaisant leur profil de préférence et vont modifier ce dernier pour converger vers un consensus [Chevaleyre *et al.*, 2006, Johansson *et al.*, 2008]. Ces protocoles doivent satisfaire certaines propriétés comme des notions d'équité ou d'efficacité [Endriss *et al.*, 2006, Bertsimas *et al.*, 2011] ;
– La théorie des jeux coopératifs, aussi appelés jeux de coalitions, dans laquelle les agents forment des groupes avec les agents avec qui ils vont interagir. Ces jeux se divisent en deux grandes catégories : les jeux à utilité transférable généralement utilisés lorsque les agents cherchent à maximiser des gains et les jeux hédoniques lorsque les agents cherchent à satisfaire des préférences sur les groupes auxquelles ils peuvent appartenir [Nash, 1950, Drèze et Greenberg, 1980, Shehory et Kraus, 1998].

Cependant, l'option collective choisie par l'ensemble des agents n'est pas toujours celle désirée individuellement par les agents. Un agent égoïste peut alors se demander s'il lui est possible de définir une stratégie afin que la décision collective lui soit favorable. Un tel comportement stratégique est appelé une *manipulation*.

1.2 Des agents manipulateurs

1.2.1 Qu'est-ce qu'une manipulation ?

La robustesse d'un système face aux manipulations est l'une des grandes problématiques des systèmes multi-agents ouverts où les agents sont rationnels et pleinement autonomes. Ce problème est étudié dans de nombreux domaines tels que celui de la théorie des choix sociaux [Gärdenfors, 1976], les processus de décision individuelle tels que les systèmes d'enchère [Robinson, 1985], les systèmes de réputation pour des sites commerciaux [Schafer *et al.*, 1999], la sécurité des réseaux [Alpcan et Başar, 2010] et bien d'autres. La définition et la mise en œuvre d'une manipulation sont souvent spécifiques à chacun de ces domaines. Par exemple, dans le cadre des systèmes de vote, [Gibbard, 1973] définit une manipulation ainsi :

Définition 1.2.1 - *Manipulation d'un système de vote* : Un individu *manipule un système de vote* si, en fournissant un faux profil de préférence, il s'assure d'un résultat qu'il préfère à celui normalement obtenu s'il avait fourni son véritable profil de préférence.

Dans le cadre des réseaux (où l'on utilise plus généralement le terme d'attaque), une manipulation consiste à s'introduire dans un système pour en perturber son fonctionnement [Ellison *et al.*, 1997]. Dans ce manuscrit, nous considérons la définition d'une manipulation suivante :

Définition 1.2.2 - *Manipulation* : Une manipulation est une stratégie, permettant à un agent a_i d'influencer et de contrôler le processus de décision individuel (ou collectif) d'un ensemble d'agents à l'aide de fausses informations, afin que ces derniers prennent une décision favorable à l'agent a_i.

Un agent manipulateur peut alors avoir deux raisons de manipuler. La première est de vouloir utiliser le système pour ce pour quoi il est conçu tout en le manipulant afin d'augmenter son gain indépendamment du gain des autres agents. Nous parlons alors d'*agents malhonnêtes*. Par exemple, bourrer les urnes a pour but de bel et bien obtenir un résultat de vote favorable. La seconde raison est de perturber le fonctionnement du système, c'est-à-dire en l'empêchant de réaliser les fonctions pour lesquels il a été conçu. Nous parlons alors d'*agents malveillants*. Par exemple, supprimer arbitrairement des messages qui transitent dans un réseau pour faire croire à une défaillance de l'agent émetteur. Notons cependant qu'il n'existe pas de frontière stricte entre agents malhonnêtes et agents malveillants. En effet, la malhonnêteté d'un agent induit généralement des baisses de gains pour les autres agents. De même, comme le gain d'un agent malveillant peut être modélisé par l'opposé du gain des autres agents, être malhonnête lui suffit parfois à être malveillant.

Définition 1.2.3 - *Agent malhonnête* : Un *agent malhonnête* est un agent qui manipule un système afin de maximiser son gain, indépendamment du gain obtenu par les autres agents.

Définition 1.2.4 - *Agent malveillant* : Un *agent malveillant* est un agent qui manipule un système afin de minimiser le gain d'un sous-ensemble d'agents tiers.

La distinction entre ces deux types d'agents est à rapprocher de celle introduite par [Conitzer *et al.*, 2003] dans les systèmes de vote. Ils considèrent en effet deux types de manipulations : les manipulations constructives où un agent manipule le système pour faire élire l'option qu'il préfère (malhonnêteté) et les manipulations destructives où un agent manipule le système pour faire perdre une option donnée (malveillance).

Si un seul agent manipulateur peut individuellement influencer les décisions des autres agents, plusieurs agents manipulateurs se regroupant autour du même objectif peuvent avoir une influence plus importante. C'est un phénomène de *collusion* tel que décrit par [Robinson, 1985] dans les systèmes d'enchères. De manière générale, nous définissons une collusion comme suit :

Définition 1.2.5 - *Collusion* : Une *collusion* est une coalition d'agents malhonnêtes ou malveillants s'accordant pour définir une manipulation commune.

Il existe de nombreux types de manipulations [Douceur, 2002, Chang, 2002, Bachrach et Elkind, 2008, Bilge *et al.*, 2009, Hoffman *et al.*, 2009, Waggoner *et al.*, 2012]. La figure 1.1 présente une taxonomie (non exhaustive au niveau des feuilles) des différentes manipulations dans les systèmes multi-agents. De manière générale, les processus de décision des agents sont fondés en partie sur leurs connaissances. Ainsi, manipuler un agent consiste à biaiser ses connaissances. Pour ce faire, l'agent manipulateur peut soit partager *explicitement* avec sa cible des informations qu'il sait être fausses, soit *implicitement* amener ces derniers à déduire de fausses connaissances.

1.2.2 Manipulations explicites

Les agents n'ayant pas une perception complète et parfaite de l'environnement, ils peuvent s'échanger une partie de leurs informations afin de renforcer mutuellement leurs connaissances [Stone et Veloso, 2000]. Ces échanges d'informations leur permettent ainsi de prendre de meilleures décisions, qu'elles soient individuelles comme dans les systèmes de réputation [Pazzani et Billsus, 2007] ou collectives comme dans les systèmes de vote [Gärdenfors, 1976]. Nous pouvons distinguer deux catégories d'informations : les *informations privées* d'un agent, c'est-à-dire sa

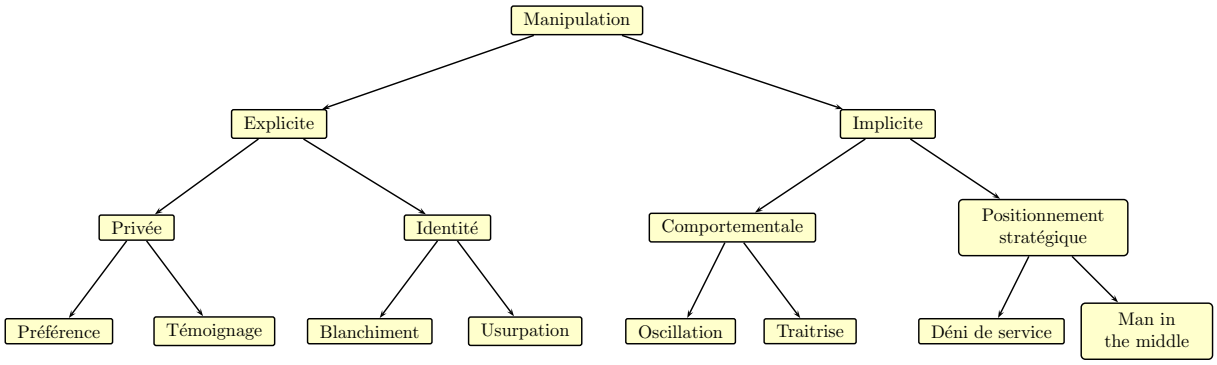

FIGURE 1.1 – Taxonomie des manipulations dans les SMA

représentation interne de l'environnement et les *informations publiques*, c'est-à-dire l'ensemble des connaissances observables par tous les agents du système.

Sur les informations privées

Les processus de communication entre agents permettent à un agent de partager avec un agent tiers une partie de ses informations privées. Cependant, ses informations étant internes à chaque agent, un agent manipulateur peut volontairement fournir à un agent tiers de fausses informations. Or, l'apport du partage d'informations privées repose toujours sur la véracité de ces dernières. Par exemple, dans le cadre de la prise de décision collective, les agents partagent avec les autres agents tout (ou une partie de) leur *profil de préférence* afin de les agréger par une fonction de choix social. En ayant connaissance du processus d'agrégation des profils de préférence, un agent manipulateur peut se demander si le fait de fournir un *faux profil de préférence*, c'est-à-dire mentir sur ses préférences, peut modifier le résultat de la fonction de choix social et ainsi obtenir un résultat qu'il préfère.

Exemple 1.2.1 - Considérons trois agents a_1, a_2 et a_3 devant faire un choix parmi quatre options A, B, C et D. Considérons les profils de préférence suivants :

- $a_1 : A \succ_1 B \succ_1 C \succ_1 D$;
- $a_2 : C \succ_2 B \succ_2 A \succ_2 D$;
- $a_3 : B \succ_3 A \succ_3 D \succ_2 C$.

Considérons le système de vote [de Borda, 1781] où, pour 4 options, chaque agent donne 3 points à l'option qu'il préfère, 2 à la seconde, 1 à la troisième et 0 à la dernière. Le vainqueur du vote est l'option réunissant le plus grand nombre de points. Dans cette situation, si tous les agents fournissent leur véritable profil de préférence, le vainqueur est le candidat B comme le montre le tableau 1.2.

Cependant, l'agent a_1 peut mentir sur ces préférences et fournir le profil de préférence $\succ'_1$: $B \succ'_1 D \succ'_1 C \succ'_1 A$. Le tableau 1.3 nous montre alors que, dans cette situation, le vainqueur devient l'option A qui est préférée par l'agent a_1 à l'option B.

Si l'échange des profils de préférence permet aux agents de prendre une décision collective, d'autres systèmes tels que les systèmes de recommandation utilisent l'échange d'informations dans les processus de décision individuelle. Dans ces systèmes, les agents s'échangent une partie de leurs connaissances par le biais de *témoignages* sur des objets et les agrègent afin d'obtenir

		Options			
		A	B	C	D
Votes	a_1	3	2	1	0
	a_2	1	2	3	0
	a_3	2	3	0	1
Score de Borda		6	7	4	1

Tableau 1.2 – Scores de Borda si les agents sont honnêtes

		Options			
		A	B	C	D
Votes	a_1	3	0	1	2
	a_2	1	2	3	0
	a_3	2	3	0	1
Score de Borda		6	5	4	3

Tableau 1.3 – Score de Borda si l'agent a_1 ment sur son profil de préférence

une connaissance plus précise sur ces derniers. L'utilisation de *faux témoignages* permet alors à un agent manipulateur de biaiser les connaissances d'un autre agent et ainsi influencer son processus de décision. La problématique des faux témoignages est d'autant plus importante dans le contexte des systèmes de réputation [Hoffman *et al.*, 2009] où les informations partagées portent sur des agents tiers. Dans ces deux systèmes, une fonction d'agrégation des connaissances permet d'obtenir un rang entre les objets (artefacts ou agents). Les faux témoignages ont pour objectif de modifier cet ordre. Nous distinguons deux catégories de faux témoignages :

- les *promotions* où les faux témoignages permettent d'améliorer le rang d'un objet ;
- les *diffamations* où les faux témoignages permettent de diminuer le rang d'un objet.

Sur l'identité, une information publique

Dans les systèmes multi-agents, les agents se distinguent les uns des autres par l'utilisation d'une information publique personnelle : leur *identité*. L'identité d'un agent est une représentation abstraite qu'il fournit afin de se faire reconnaître par les autres agents dans le système. Par exemple, dans le cadre des réseaux, il s'agit d'une adresse IP. Dans le cadre des réseaux sociaux, il peut s'agir d'une adresse email. De manière générale, il est fait l'hypothèse dans les systèmes multi-agents qu'un agent ne dispose que d'une et une seule identité. Ainsi, manipuler l'identité correspond à soit se faire passer pour une autre, soit en créer une fausse de toute pièce.

Le premier type de manipulation est l'*usurpation d'identité* [Koops et Leenes, 2006, Bilge *et al.*, 2009, Angin *et al.*, 2010]. [Koops et Leenes, 2006] décrivent cette manipulation comme le fait de se faire passer pour un autre agent sans son consentement, et obtenir ainsi des données privées auxquelles il n'aurait pas eu accès.

Un second type de manipulation consiste à se donner une nouvelle identité créée de toute pièce, en abandonnant la précédente. Nous parlons alors de *blanchiment* [Feldman *et al.*, 2006]. En effet, certains systèmes multi-agents tels que les systèmes de réputation permettent aux agents d'apprendre le comportement des autres à partir de leurs interactions passées. Se blanchir permet alors à un agent de perturber ce mécanisme d'apprentissage en quittant le système et en le réintégrant en tant que nouvel agent. Ainsi, les autres agents le considèrent comme un nouvel

agent venant de rejoindre le système et ne tiennent plus compte de ce qu'ils ont appris sur lui.

Cependant, un agent se blanchissant peut conserver tout de même sa précédente identité. Cela correspond à s'introduire dans le système sous de *multiples fausses identités*. Cette manipulation est tour à tour appelée *attaque Sybil* [Douceur, 2002, Cheng et Friedman, 2005] ou *false-name manipulation* [Bachrach et Elkind, 2008, Aziz et Paterson, 2009, Waggoner *et al.*, 2012]. Si l'ensemble de ces fausses identités, appelées agents Sybil, est associé en pratique au même agent, elles apparaissent pour les autres agents du système comme autant d'agents distincts avec lesquels ils peuvent interagir. Cette manipulation permet à un agent de construire une collusion virtuelle et donc d'utiliser tous les avantages de la collusion sans pour autant avoir besoin de complices.

Exemple 1.2.2 - Dans l'exemple 1.2.1, l'agent a_1 peut effectuer une attaque Sybil en intégrant dans l'ensemble des votants un nouvel agent a_4 avec le profil de préférence $A \succ_4 B \succ_4 C \succ_4 D$. Comme le montre le tableau 1.4, l'utilisation de l'agent Sybil permet ainsi à l'agent a_1 de faire gagner l'option A qu'il préfère. affecter

		Options			
		A	B	C	D
Votes	a_1	3	2	1	0
	a_2	1	2	3	0
	a_3	2	3	0	1
	a_4	3	2	1	0
Score de Borda		9	8	5	1

Tableau 1.4 – Scores de Borda lorsque a_1 introduit un agent Sybil a_4

1.2.3 Manipulations implicites

À l'inverse des manipulations explicites, une *manipulation implicite* consiste à interagir avec le système afin que les autres agents déduisent de leurs observations de fausses connaissances. Nous considérons dans ce manuscrit les *manipulations comportementales* lorsque l'agent manipulateur fournit de fausses informations par l'intermédiaire d'un comportement observable particulier et les manipulations par *positionnement stratégique* lorsque l'agent manipulateur va agir afin de réduire la capacité d'observation des autres agents.

Manipulations comportementales

Dans certains systèmes multi-agents, tel que les jeux répétés [Foster et Young, 2003] et les systèmes de réputation [Resnick *et al.*, 2000], les agents estiment le comportement futur des autres agents en utilisant des techniques d'apprentissages sur leurs observations lors d'interactions passées. L'hypothèse est faite que les agents suivent le même comportement au cours du temps. À partir de cette estimation sur le comportement d'un agent tiers, un agent peut calculer l'utilité d'interagir à nouveau avec lui. L'utilité résultante d'une interaction avec un agent a_i permet de considérer son comportement comme :

- fiable si interagir avec a_i apporte une récompense ;
- non fiable si interagir avec a_i apporte une perte.

Les *manipulations comportementales* consistent à effectuer une succession d'actions qui vont biaiser le processus d'apprentissage. Par exemple, dans les systèmes de réputation, le comportement appelé *traîtrise* consiste à adopter un comportement fiable pendant une période de temps afin d'être identifié en tant que tel puis subitement adopter un comportement non fiable, mais associé à un gain important [Marti et Garcia-Molina, 2006].

Exemple 1.2.3 - Considérons un site de vente en ligne. Un vendeur malveillant peut vendre plusieurs objets de bonne qualité (comportement fiable), puis une fois qu'il a gagné la confiance d'un client, lui vendre un objet de mauvaise qualité à un fort coût (comportement non fiable).

Si une traîtrise consiste à changer soudainement de comportement, d'autres manipulations consistent à alterner entre comportement fiable et non fiable. Par exemple, dans les systèmes de réputation, l'*attaque oscillante* combine promotion, diffamation et traîtrise [Srivatsa *et al.*, 2005, Hoffman *et al.*, 2009] : les agents en collusion se divisent en deux groupes M_1 et M_2. Les agents de M_1 présentent un comportement fiable et promeuvent les agents de M_2. Les agents de M_2 ont un comportement non fiable et diffament les agents n'appartenant pas à la collusion. Lorsque la réputation des agents de M_2 est inférieure à un seuil, les deux groupes inversent leurs rôles. Ainsi, les agents de M_1 profitent de leur bonne réputation pour avoir un comportement non fiable et ceux de M_2 font progressivement remonter leurs valeurs de réputation en présentant un comportement fiable.

Positionnements stratégiques

Dans les systèmes multi-agents organisés qui forment une structure topologique, un agent manipulateur peut profiter d'une position stratégique dans cette structure pour perturber le système. Ces manipulations s'attaquent au réseau d'accointance et de communication des agents.

L'une de ces manipulations, appelée *man-in-the-middle*, consiste pour un agent à se positionner dans le réseau afin d'intercepter les communications entres deux agents pour éventuellement les déformer [Meyer et Wetzel, 2004]. Cette manipulation se généralise en une *attaque éclipse* où une collusion d'agents malveillants va isoler du réseau un sous-ensemble des agents afin que ces derniers ne puissent pas interagir avec les autres agents du système [Specht et Lee, 2004, Singh *et al.*, 2006].

Exemple 1.2.4 - Considérons un ensemble $N = \{a_1, \ldots, a_9\}$ d'agents organisés en un réseau comme le montre la figure 1.2. Dans ce réseau, l'agent a_9 ne peut communiquer directement qu'avec les agents a_6, a_7 et a_8. Si ces derniers forment une collusion, ils peuvent intercepter tous les messages transmis en provenance ou à destination de l'agent a_9, le rendant dans l'incapacité d'interagir avec les agents $\{a_1, \ldots, a_5\}$. Par ailleurs, si l'agent a_9 ne dispose pas d'une connaissance a priori de l'organisation du réseau, il ne considérera que le sous-réseau composé des agents $\{a_6, a_7, a_8, a_9\}$.

En isolant une partie des agents du système, l'attaque éclipse s'inscrit dans une plus large famille de comportements malveillants étudiée pour la sécurité des réseaux : le *déni de service*. Une stratégie de déni de service est l'ensemble des actions qui réduisent la capacité d'un système à réaliser les fonctions pour lesquels il a été conçu [Wood et Stankovic, 2002]. Par exemple, dans le cas des réseaux, il s'agit d'empêcher un agent de fournir ses services en lui envoyant une grande quantité de requêtes [Chang, 2002]. Ceci a pour effet de consommer les ressources de l'agent ciblé et ainsi de l'empêcher de les utiliser pour fournir ses services aux autres agents. Ceci a pour autre conséquence de surcharger le réseau de messages, ce qui augmente les délais de

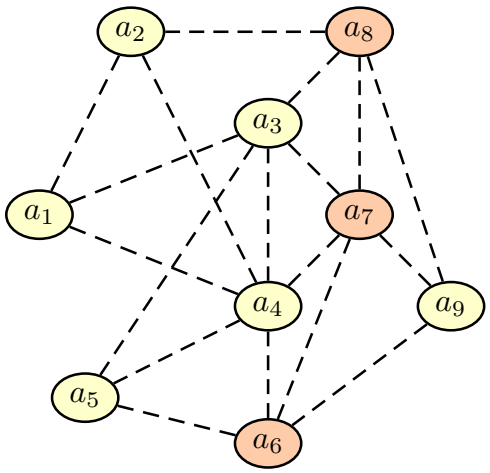

FIGURE 1.2 – Attaque Eclipse

communication entre les agents [Lau *et al.*, 2000]. Dans le cas des systèmes de partage de fichiers, il est possible de faire du déni de service en surchargeant le système de fichiers corrompus afin d'empêcher les agents d'accéder à un fichier original [Dumitriu *et al.*, 2005].

1.3 Stratégies de défense

Si les manipulations dans les systèmes multi-agents peuvent prendre de multiples formes, les stratégies pour lutter contre sont tout aussi variées. Elles peuvent être classées en trois catégories :

1. l'axiomatisation du système ;
2. la complexité algorithmique ;
3. l'authentification.

1.3.1 Axiomatisation

L'approche par axiomatisation consiste à définir des propriétés sur le système qui, si elles sont respectées, rendent ce système robuste aux manipulations. Cette approche est particulièrement étudiée dans le cadre de la théorie des choix sociaux [Gibbard, 1973,Satterthwaite, 1975,Barbera, 2001, Nehring et Puppe, 2007, Guo et Conitzer, 2010], mais aussi en théorie des jeux [Nash, 1951, Milgrom et Weber, 1985, Maskin, 1999]. Dans la théorie des choix sociaux, une fonction de choix social est considérée comme *robuste aux choix stratégiques* s'il n'existe pas d'agent qui, en fournissant un faux profil de préférence, peut obtenir un résultat préférable à celui obtenu s'il avait fourni son véritable profil de préférence. Cette définition consiste simplement à interdire les manipulations au sens de la définition de [Gibbard, 1973] (définition 1.2.1). Malheureusement, les approches par axiomatisation sont limitées par l'existence de théorèmes d'impossibilité entre plusieurs axiomes, comme nous allons l'illustrer sur les trois domaines que sont les systèmes de vote, les systèmes de recommandation et les systèmes de réputation.

Dans le cadre des systèmes de vote, le théorème d'impossibilité de Gibbard - Satterthwaite [Gibbard, 1973, Satterthwaite, 1975] montre que pour au minimum 3 options et 2 votants, il n'existe pas de règle de vote qui respecte simultanément les axiomes :

- de *non-dictature* signifiant que le résultat du vote ne dépend pas du profil de préférence d'un unique agent ;
- de *robustesse aux manipulations* signifiant qu'il n'existe pas d'agent a_i ayant un profil de préférence $\succeq_i$ tel que si a_i annonce un profil de préférence $\succeq'_i$, le candidat vainqueur soit préférable que celui obtenu si a_i avait énoncé ces véritables préférences.

Ce théorème d'impossibilité est du même ordre que celui de [Arrow, 1963] montrant qu'il n'existe pas de fonction d'agrégation définie sur l'ensemble de tous les profils et satisfaisant les axiomes d'unanimité, d'indifférence aux options non pertinentes et de non-dictature.

Si [Gibbard, 1973,Satterthwaite, 1975] ne considèrent que le cas du vote stratégique, [Tennenholtz, 2004,Altman et Tennenholtz, 2005,Altman et Tennenholtz, 2007b,Altman et Tennenholtz, 2007a] ont étudié l'axiomatisation des systèmes de réputation [1]. Ces travaux définissent un système de réputation comme robuste aux manipulations s'il n'existe pas de stratégie permettant à un agent de modifier l'ordre obtenu par la fonction de recommandation à son avantage. Les axiomes principalement considérés dans ce cadre sont :

- la *généralité* signifiant que la fonction de recommandation doit être définie pour tout graphe représentant la confiance (arêtes) entre les agents (nœuds) ;
- la *transitivité* signifiant que si les agents qui ont confiance dans un agent a_i ont un meilleur rang que ceux qui ont confiance dans un agent a_j alors a_i a un meilleur rang que a_j ;
- la *monotonie* signifiant que si un agent a_i a un meilleur rang qu'un agent a_j alors il existe au moins un agent a_k qui a confiance dans a_i et dont le rang est supérieur à tout agent qui a confiance dans a_j ;
- l'*indépendance des alternatives non-pertinentes* [2] signifiant que le rang d'un agent ne doit dépendre que du rang des agents qui ont confiance en lui ;
- l'*incitation à la vérité* signifiant qu'il n'est pas possible pour un agent d'augmenter son rang en communiquant de fausses préférences.

Comme dans le cas des fonctions de choix social, l'axiomatisation des systèmes de réputation se heurte à des théorèmes d'impossibilité [Altman et Tennenholtz, 2007b]. Par exemple, [Tennenholtz, 2004] a montré qu'il n'existe pas de systèmes de réputation respectant à la fois la généralité, la transitivité, la monotonie et l'incitation à la vérité. De leur côté, [Cheng et Friedman, 2005] ont axiomatisé les systèmes de réputation afin d'étudier leur robustesse aux attaques Sybil. Ce travail définit un système de réputation comme robuste aux attaques Sybil si aucun agent ne peut obtenir une meilleure valeur de réputation, en introduisant dans le système de fausses identités. Il y est montré que si une fonction de réputation est *symétrique*, c'est-à-dire que la réputation des agents ne dépend pas de leur position dans le graphe de confiance, elle ne peut pas être robuste aux attaques Sybil. Par ailleurs, une fonction de réputation *asymétrique* n'est robuste aux attaques Sybil que si les opérateurs d'agrégation et de propagation de la confiance sont *monotones*, *décroissants* et *sous-additifs*.

1.3.2 Complexité

Bien que la robustesse des systèmes aux manipulations soit mise en défaut par des théorèmes d'impossibilité, il peut tout de même être difficile pour un agent de calculer une stratégie lui permettant de manipuler un système. C'est pourquoi certaines approches venant principalement de la théorie du choix social étudient la robustesse d'un système aux manipulations en se fondant

1. Pour être précis, ces travaux se positionnent comme étudiant les systèmes de recommandation, mais leur modélisation est suffisamment générale pour représenter en fait des systèmes de réputation.
2. Malgré son nom, cet axiome diffère complètement de l'axiome défini par [Arrow, 1963] dans le cadre des fonctions de choix social.

sur la complexité algorithmique à calculer les manipulations en question [Bartholdi III *et al.*, 1989, Bartholdi III et Orlin, 1991, Elkind et Lipmaa, 2005, Xia *et al.*, 2009].

Par exemple dans le domaine de la formation de coalitions, [Conitzer et Sandholm, 2004] montrent qu'il est $NP-$difficile de trouver une manipulation lorsque la répartition des gains se fait en fonction de la valeur de Shapley. Dans les systèmes de vote, [Bartholdi III *et al.*, 1989, Bartholdi III et Orlin, 1991] ont montré que le vote majoritaire ou bien encore la règle de Borda sont faciles à manipuler, car il existe un algorithme glouton capable de calculer en un temps polynomial une manipulation. À l'inverse, la seconde règle de Copeland où le score de chaque option dépend d'une comparaison des profils de préférence deux à deux tout comme le scrutin à vote unique transférable (STV) sont difficiles à manipuler, car le problème de décision associé est NP-difficile.

Cependant, la complexité des manipulations diffère selon le type de manipulation, constructive, destructive, en collusion ou non, comme le montrent [Conitzer et Sandholm, 2002, Conitzer et Sandholm, 2006]. Si les manipulations constructives étudiées sont NP-difficiles pour la majorité des règles de votes (Borda, Copeland, Maximin), les manipulations destructives peuvent se calculer en temps polynomial. La table 1.5 résume les résultats de complexité prouvés par [Conitzer et Sandholm, 2006] en fonction de la règle de vote et du nombre d'options.

	Constructive				**Destructive**	
Nombre d'options	2	3	4,5,6	≥ 7	2	≥ 3
Borda	P	NP-c	NP-c	NP-c	P	P
Veto	P	NP-c	NP-c	NP-c	P	P
STV	P	NP-c	NP-c	NP-c	P	NP-c
Vote majoritaire à plusieurs tours	P	NP-c	NP-c	NP-c	P	NP-c
Copeland	P	P	NP-c	NP-c	P	P
Maximin	P	P	NP-c	NP-c	P	P
Vote binaire aléatoire	P	P	P	NP-c	P	?
Vote binaire régulier	P	P	P	P	P	P
Vote majoritaire	P	P	P	P	P	P

Tableau 1.5 – Complexité des manipulations [Conitzer et Sandholm, 2006]

Remarquons que de manière générale la complexité des manipulations est étudiée dans le pire cas. Cependant, [Walsh, 2009] montre que bien qu'une règle de vote soit difficile à manipuler dans le pire cas, il existe de nombreux cas pratiques dans lesquels calculer une manipulation se fait en temps polynomial. Ainsi, si la complexité algorithmique peut être considérée comme un frein aux manipulations, cela ne garantit en rien la robustesse d'un système dans le cas général.

1.3.3 Authentification

Le problème de l'authentification des agents est particulièrement important puisque l'attaque Sybil permet à un agent de manipuler le système sans risque en cas de détection. Par ailleurs, l'utilisation des agents Sybils simplifie la formation de collusion. C'est pourquoi des mécanismes de défense ont été spécifiquement conçus pour lutter contre des agents utilisant de fausses identités [Douceur, 2002, Levine *et al.*, 2006, Guo et Conitzer, 2010].

Par exemple, [Douceur, 2002] propose un mécanisme de validation permettant à un agent de décider s'il doit considérer deux identités comme distinctes. En effet, en faisant l'hypothèse

que les agents sont limités sur trois ressources, les capacités de communication, la mémoire et les capacités de calcul, [Douceur, 2002] propose différents défis qui ne peuvent être résolus par un agent en un temps borné que si cet agent est présent dans le système sous une seule et unique identité. Le premier défi consiste à diffuser une requête à l'ensemble des identités et à n'accepter comme valable que les agents ayant répondu en un temps limité. Cette requête permet ainsi de détecter de fausses identités partagées par un même agent ayant des contraintes de communication. Pour les agents ayant des capacités de mémoire limitées, il est possible de détecter les fausses identités en demandant aux agents de stocker une large quantité de données non compressibles. Enfin, pour détecter les identités ayant des capacités de calcul communes, le défi consiste à demander à chaque identité de résoudre un problème exponentiel en temps de résolution, mais vérifiable en temps constant. Ainsi, si plusieurs fausses identités reçoivent simultanément ce défi, elles ne peuvent y répondre chacune en un temps borné.

D'autres approches consistent à limiter le nombre d'identités que peuvent prendre les agents en associant un coût aux agents qui désirent rejoindre le système [Borisov, 2006]. Ce coût peut être monétaire [Hildrum *et al.*, 2004], comme algorithmique tel que la résolution de CAPTCHA [Von Ahn *et al.*, 2003]. [Margolin et Levine, 2008] proposent d'instaurer un coût à l'utilisation du système au cours du temps et non pas uniquement à la création des identités. Ils montrent alors que cette méthode réduit le gain des agents manipulateurs à utiliser un grand nombre de fausses identités comparé à l'instauration d'un coût à l'arrivée de l'agent dans le système.

Cependant, ces approches reposent sur l'hypothèse que les agents effectuant des attaques Sybils disposent de restriction sur leurs ressources similaires à celle des autres agents, ce qui n'est pas toujours le cas. Par exemple, l'utilisation de réseaux de machines-zombies permet à un agent manipulateur de disposer d'une puissance de calcul plus importante [Gu *et al.*, 2008]. Par ailleurs, l'introduction d'un coût vient réduire la propriété d'ouverture des systèmes multi-agents.

Une autre approche est de définir une autorité centrale d'authentification [Kent et Atkinson, 1998, Resnick *et al.*, 2001, Newsome *et al.*, 2004]. Dans ces approches, l'autorité centrale est chargée d'attribuer à chaque agent une identité unique et de vérifier que ce certificat ne peut pas être falsifié ou usurpé. [Chan *et al.*, 2003] proposent par exemple un mécanisme aléatoire de distribution d'identifiants aux différents agents du système.

1.4 Problématique générale

Dans ce premier chapitre d'état de l'art, nous avons présenté les systèmes multi-agents dans leur globalité. Si dans certains systèmes les agents sont conçus pour réaliser collectivement un même objectif, de nombreuses applications nécessitent de considérer des systèmes *décentralisés* et *ouverts* où un grand nombre d'agents *hétérogènes* et *pleinement autonomes* partagent le même environnement et prennent leurs décisions de manière *rationnelle*.

Or, dans ce type de systèmes, des *agents manipulateurs* peuvent user de *stratégies* leur permettant d'influencer les processus de décision des autres agents en biaisant leurs connaissances. Bien que la littérature consacrée à la lutte contre les manipulations soit abondante :

1. des théorèmes d'impossibilité montrent qu'il n'existe pas de systèmes parfaitement robustes aux manipulations ;
2. de nombreuses manipulations sont simples à calculer en pratique malgré des preuves de complexité au pire cas qui leurs sont défavorables ;
3. les solutions ad-hoc consistent à affaiblir les deux propriétés de décentralisation et d'ouverture désirables dans les systèmes multi-agents.

Au vu de ces résultats, il convient de se poser les questions suivantes :

1. Existe-t-il des conditions particulières qui rendent en pratique inefficaces les manipulations simples ?
2. Peut-on trouver des stratégies de défense à adjoindre à un système multi-agents qui n'affaiblissent pas ses propriétés fondamentales ?

Ainsi, **nous proposons d'étudier les manipulations au regard des propriétés dont elles ont besoin pour être mises en œuvre de manière efficace ainsi que les stratégies de défense qui viennent renforcer ou affaiblir ces propriétés en fonction de leur influence sur les manipulations**.

Pour répondre à ces questions, nous nous intéressons dans ce manuscrit aux problèmes des manipulations sur deux systèmes : les *jeux hédoniques* et les *systèmes de réputation*. Nous avons choisi ces deux systèmes, car ils permettent de couvrir des propriétés complémentaires : les jeux hédoniques sont des problèmes de décision collective et les systèmes de réputation de décision individuelle, les jeux hédoniques sont des systèmes statiques et les systèmes de réputation sont dynamiques, les jeux hédoniques se fondent sur des utilités ordinales et les systèmes de réputation sur des utilités cardinales. De plus, chaque famille nous permet d'aborder un point de notre problématique : caractériser les propriétés nécessaires à la mise en œuvre de manipulations et identifier des stratégies de défense au regard de ces propriétés.

Ainsi, dans le premier cas, nous considérons un problème d'*agents malhonnêtes* utilisant des agents *Sybil* afin de manipuler un *jeu hédonique*. Dans le second cas, nous considérons un problème d'*agents malveillants en collusion* manipulant un *système de réputation*.

Chapitre 2

Systèmes considérés

Sommaire

Résumé.
Si la problématique de la robustesse des systèmes multi-agents aux manipulations peut être étudiée dans de nombreux contextes, nous considérons dans ce manuscrit deux catégories de systèmes spécifiques : les *jeux hédoniques* et les *systèmes de réputation*. Ces deux familles de systèmes multi-agents s'intéressent aux processus de prise de décision des agents dans deux cadres différents. Dans le contexte des jeux de hédoniques, il s'agit pour un ensemble d'agents de décider collectivement quelle *structure de coalitions* ils vont former, et ce en essayant de satisfaire les *préférences* des différents agents. Dans le contexte des systèmes de réputation, les agents utilisent la dynamique du système pour décider quels autres agents du système sont dignes de *confiance* afin de leur déléguer des tâches à réaliser. Ce chapitre a pour objectif de présenter les concepts fondamentaux liés à ces deux types de systèmes.

2.1 Jeux de coalitions hédoniques

Des agents pleinement autonomes sont parfois amenés à coopérer temporairement dans le but de réaliser collectivement une tâche qu'ils ne peuvent pas faire seuls. Dans ce cas, les agents doivent se demander avec quels sous-ensembles des agents coopérer pour en tirer une utilité propre. Ce problème est appelé un problème de *formation de coalitions*.

2.1.1 Jeux de coalitions

Les *jeux de coalitions* ont été le sujet de très nombreuses publications dans la littérature [Nash, 1950, Shapley, 1952, Morgenstern et Von Neumann, 1953, Gamson, 1961, Kelso Jr et Crawford, 1982, Okada, 1996, Sandhlom et Lesser, 1997, Bloch, 1997, Ray et Vohra, 1999, Rahwan et Jennings, 2007, Elkind et Wooldridge, 2009, Génin, 2010, Hoefer *et al.*, 2014]. De tels jeux consistent, pour un ensemble $N = \{a_1, \ldots, a_n\}$ d'agents partageant le même environnement, à décider des sous-ensembles d'agents qui vont temporairement coopérer afin de réaliser collectivement une même tâche.

Définition 2.1.1 - *Coalition* : Soit N, l'ensemble des agents. Une *coalition* $C \subseteq N$ est un sous-ensemble non vide d'agents. La *coalition singleton* d'un agent $a_i \in N$ désigne la coalition $\{a_i\}$. La *grande coalition* est la coalition contenant l'ensemble des agents : $C = N$.

Dans la suite de ce manuscrit, nous désignons par $\mathcal{C}^N$ l'ensemble des coalitions possibles pour l'ensemble d'agents N, c'est-à-dire l'ensemble des sous-ensembles possibles de N. Nous désignons aussi par $\mathcal{C}^N_{a_i}$ l'ensemble des coalitions possibles contenant l'agent $a_i \in N$.

Exemple 2.1.1 - Considérons un système multi-agents où $N = \{a_1,a_2,a_3\}$. La Figure 2.1 présente $\mathcal{C}^N$ l'ensemble des coalitions possibles pour N.

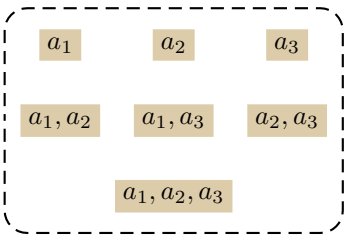

FIGURE 2.1 – Ensemble des coalitions possibles pour $N = \{a_1,a_2,a_3\}$

Dans cet exemple, la grande coalition est la coalition $C = \{a_1,a_2,a_3\}$. Parmi l'ensemble des coalitions possibles $\mathcal{C}^N$, l'ensemble des coalitions contenant l'agent a_1 est :

$$\mathcal{C}^N_1 = \{\ \{a_1\},\{a_1,a_2\},\{a_1,a_3\},\{a_1,a_2,a_3\}\ \}$$

De manière générique, [Génin, 2010] définit le problème de formation de coalitions pour l'ensemble des agents du système comme le fait de déterminer à un instant donné quelles coalitions ils vont former. Cela revient à trouver un partitionnement de l'ensemble des agents tel que chaque agent appartient à une seule et unique coalition. Une telle partition est appelée une *structure de coalitions*.

Définition 2.1.2 - *Structure de coalitions* : Soit N, l'ensemble des agents. Une *structure de coalitions* est un partitionnement de N, c'est-à-dire un ensemble de coalitions $\Pi = \{C_1, \ldots, C_k\}$ tel que les coalitions de Π sont :

1. non vides : $\forall i \in [1,k], C_i \neq \emptyset$;
2. deux à deux disjointes : $\forall i,j \in [1,k], i \neq j \implies C_i \cap C_j = \emptyset$;
3. couvrantes : $\forall a_i \in N, \exists C \in \Pi : a_i \in C$.

Notons que s'il est généralement considéré qu'un agent ne peut appartenir à un instant donné qu'à une seule et unique coalition, [Shehory et Kraus, 1998] étendent le problème aux cas des coalitions chevauchantes, c'est-à-dire où un agent peut appartenir simultanément à plusieurs coalitions. Cette généralisation leur permet de modéliser le problème d'affectation de tâches comme un problème de formation de coalitions afin d'obtenir une affectation qui maximise une fonction d'utilité.

Dans la suite, de ce manuscrit, nous dénotons par $C_{a_i}^{\Pi}$ la coalition de l'agent a_i dans la structure de coalitions Π. De même, nous désignons par $\mathcal{P}_N$ l'ensemble des structures de coalitions possibles à partir de l'ensemble d'agents N.

La Figure 2.2 montre l'ensemble des structures de coalitions possibles pour un ensemble d'agents $N = \{a_1,a_2,a_3\}$. Les arcs entre les différentes structures de coalitions représentent le passage d'une structure de coalitions à une autre lorsqu'un agent quitte sa coalition pour en rejoindre une autre.

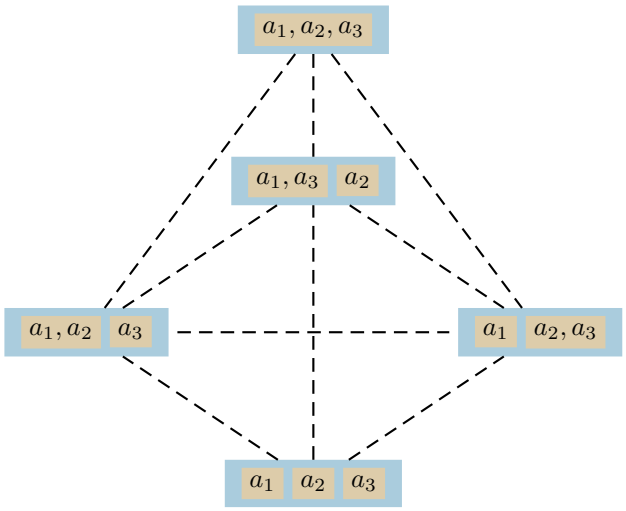

FIGURE 2.2 – Ensemble des structures de coalitions possibles pour $N = \{a_1,a_2,a_3\}$

Pour n agents, il existe $2^n - 1$ coalitions possibles, chaque agent étant présent dans 2^{n-1} de ces coalitions. Comme le montre [Wieder, 2008], le nombre de structures de coalitions possibles correspond au nombre de Bell :

$$B_n = \sum_{k=1}^{n} \left\{ \begin{array}{c} n \\ k \end{array} \right\}$$

Afin de donner un ordre de grandeur, nous présentons dans le tableau 2.1 les 10 premiers nombres de Bell. Cet ordre de grandeur nous donne un aperçu de la complexité d'énumérer l'ensemble des structures de coalitions possibles afin de décider laquelle former.

$\|N\|$	1	2	3	4	5	6	7	8	9	10
$B_{\|N\|}$	1	2	5	15	52	203	877	4140	21147	115975

Tableau 2.1 – Nombre de structures de coalitions possibles en fonction de $|N|$

Pour décider avec quels autres agents du système coopérer, les agents doivent disposer d'outils de comparaisons entre les différentes coalitions.

L'une des approches classiques de la théorie des jeux coopératifs [Shapley, 1952] est de considérer le gain que chaque agent va recevoir en formant cette coalition. Pour ce faire, les agents disposent d'une *fonction d'utilité* (aussi appelée *fonction caractéristique*) qui définit pour chaque coalition C une valeur réelle correspondant aux gains que reçoit un agent si la coalition C se forme.

Définition 2.1.3 - *Fonction d'utilité* : Soit $N = \{a_1, \ldots, a_n\}$ un ensemble d'agents. L'*utilité d'une coalition* $C \subseteq N$ pour l'agent $a_i \in C$ est définie par $u_{a_i} : 2^N \to \mathbb{R}$.

De nombreux travaux portent sur les propriétés de la fonction d'utilité des agents. On distingue notamment les *jeux de coalitions à utilité transférable* et *les jeux de coalitions à utilité non transférable*. Dans le premier cas, l'utilité d'une coalition est répartie entre les agents qui composent cette coalition [Shapley, 1952, Gamson, 1961, Hart et Kurz, 1983, Winter, 1989]. Dans le second cas, l'utilité d'une coalition dépend uniquement de l'évaluation que l'agent fait de la coalition à laquelle il appartient [McKelvey *et al.*, 1978, Harsanyi, 1963, Aumann, 1985, Winter, 1991, Suzuki *et al.*, 2015]. Notons qu'il existe des jeux de coalitions où l'utilité d'une coalition dépend également des autres coalitions de la structure : les *jeux de coalitions à externalité* [Ray et Vohra, 1999, Clippel et Serrano, 2005, Grabisch et Funaki, 2012].

Dans le cadre de jeux de coalitions coopératifs, des agents cherchent à former une structure de coalitions qui satisfait au maximum l'ensemble des agents [Aumann et Dreze, 1974]. Cette satisfaction commune se traduit par une fonction dite de *bien-être social* qui évalue la structure de coalitions [Rahwan, 2007]. Classiquement, le bien-être social est la somme des utilités des coalitions. Dans ce cas, si la fonction d'utilité des agents est super-additive ($u_{a_i}(C_1 \cup C_2) > u_{a_i}(C_1) + u_{a_i}(C_2)$) alors la structure de coalitions qui maximise le bien-être social est la grande coalition. D'autres bien-êtres sociaux, dits égalitaires, considèrent l'utilité de l'agent le moins satisfait :

$$u(\Pi) = \min_{a_i \in N} u_{a_i}(C^{\Pi}_{a_i})$$

L'utilisation des fonctions d'utilité permet une évaluation quantitative des coalitions auxquelles chaque agent peut appartenir. Une autre approche consiste a définir un opérateur de comparaison ordinale entre les structures de coalitions. Cette approche est celle des *jeux hédoniques* [Drèze et Greenberg, 1980, Bogomolnaia et Jackson, 2002, Elkind et Wooldridge, 2009].

Définition 2.1.4 - *Jeu hédonique* : Un *jeu hédonique* est défini par un couple $HG = \langle N, \succeq \rangle$ où N désigne l'ensemble des agents et $\succeq$ l'ensemble des profils de préférence des agents.

Le profil de préférence d'un agent désigne un ordre total sur l'ensemble des $2^{|N|-1}$ coalitions auquel il peut appartenir. Pour deux coalitions C_1 et C_2, $C_1 \succ_{a_i} C_2$ signifie que l'agent a_i préfère strictement la coalition C_1 à la coalition C_2. Le symbole $\sim_{a_i}$ représente quant à lui l'équivalence entre deux coalitions.

Définition 2.1.5 - *Profil de préférence* : Le *profil de préférence* d'un agent $a_i \in N$ (noté $\succeq_{a_i}$) désigne un ordre total sur $\mathcal{C}_{a_i}^N$, l'ensemble des coalitions contenant l'agent a_i.

Exemple 2.1.2 - Considérons un ensemble d'agents $N = \{a_1,a_2,a_3\}$. Chaque agent peut appartenir respectivement aux coalitions suivantes :

$$a_1 : \{a_1\},\{a_1,a_2\},\{a_1,a_3\},\{a_1,a_2,a_3\}$$
$$a_2 : \{a_2\},\{a_1,a_2\},\{a_2,a_3\},\{a_1,a_2,a_3\}$$
$$a_3 : \{a_3\},\{a_1,a_3\},\{a_2,a_3\},\{a_1,a_2,a_3\}$$

Considérons le profil de préférence de l'agent a_1 suivant :

$$\succeq_{a_1} = \{a_1,a_2\} \succ_{a_1} \{a_1,a_3\} \sim_{a_1} \{a_1,a_2,a_3\} \succ_{a_1} \{a_1\}$$

Ce profil de préférence signifie que l'agent a_1 préfère la coalition $\{a_1,a_2\}$ à la coalition $\{a_1,a_3\}$, qu'être en coalition avec l'agent a_3 est équivalent pour lui à former la grande coalition $\{a_1,a_2,a_3\}$ et que, dans tous les cas, il préfère coopérer avec un autre agent plutôt que d'être seul.

Le profil de préférence d'un agent est un ordre total, c'est-à-dire que pour tout agent $a_i \in N$, la relation binaire $\succeq_i$ respecte les propriétés de :

1. complétude : $\forall C_1,C_2 \in \mathcal{C}_{a_i}^N, C_1 \succeq_{a_i} C_2 \vee C_2 \succeq_{a_i} C_1$;
2. transitivité : $\forall C_1,C_2,C_3 \in \mathcal{C}_{a_i}^N, (\, C_1 \succeq_{a_i} C_2 \wedge C_2 \succeq_{a_i} C_3\,) \implies C_1 \succeq_{a_i} C_3$;
3. réflexivité : $\forall C_1 \in \mathcal{C}_{a_i}^N, C_1 \sim_{a_i} C_1$.

Si ces définitions classiques d'un jeu hédonique permettent de considérer tous les profils de préférence possibles, de nombreux travaux s'intéressent à des représentations compactes des profils de préférence des agents [Hajduková *et al.*, 2003, Ballester, 2004, Elkind et Wooldridge, 2009, Aziz *et al.*, 2014].

- les *listes de coalitions individuellement rationnelles*[3] modélisent des agents rationnels qui ne vont pas accepter de former une coalition moins préférée à leur coalition singleton [Ballester, 2004]. Ainsi, toute coalition $C \in \mathcal{C}_{a_i}^N$ telle que $\{a_i\} \succ_{a_i} C$ n'a pas besoin d'être modélisée dans le profil de préférence de l'agent $a_i \in N$;
- les *jeux à additivité séparable* modélisent les préférences des agents vis-à-vis des autres agents et non plus vis-à-vis de l'ensemble des coalitions. Chaque agent dispose d'une fonction $v_{a_i} : N \longrightarrow \mathbb{R}$ et la valeur d'une coalition est une agrégation (les opérateurs diffèrent selon les auteurs) des valeurs des agents qui la composent [Hajduková *et al.*, 2003, Hajduková *et al.*, 2004, Aziz *et al.*, 2011]. Des règles de départage comme un ordre lexicographique permettent d'obtenir un ordre strict sur les structures de coalitions lorsque la représentation ne permet pas de comparer deux structures.

Remarquons que la représentation en *réseaux de jeux hédoniques,* qui modélisent les préférences des agents par des règles en logique propositionnelle, généralise ces approches [Elkind et Wooldridge, 2009].

2.1.2 Concepts de solution

Dans la littérature, indépendamment du fait que les agents comparent les coalitions par une fonction d'utilité ou par des profils de préférence, les travaux portant sur les jeux de coalitions, s'intéressent principalement à deux questions : quelles sont les propriétés des structures

3. IRCL pour *individually rational coalition lists.*

de coalitions pour que celle-ci soit acceptable par les agents et comment former une structure de coalitions respectant ces propriétés ? Ces propriétés sont caractérisées par un *concept de solution*.

Bien que les jeux de coalitions fondés sur l'utilité disposent de concepts de solution spécifiques comme le nucléole [Schmeidler, 1969], nous ne présentons ici que les concepts spécifiques aux jeux hédoniques. Notons cependant que ces concepts peuvent être généralisées aux jeux à utilité transférable en considérant que pour deux coalitions C_1 et $C_2 : C_1 \succ_{a_i} C_2 \iff u_{a_i}(C_1) > u_{a_i}(C_2)$ et $C_1 \sim_{a_i} C_2 \iff u_{a_i}(C_1) = u_{a_i}(C_2)$.

D'un point de vue collectif, la meilleure structure de coalitions possible est celle qui satisfait parfaitement l'ensemble des participants, c'est-à-dire que chaque agent est dans la coalition qu'il préfère à toutes les autres. Une telle structure de coalitions est alors dite *individuellement optimale*.

Définition 2.1.6 - *Optimalité individuelle* : Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique. La structure de coalitions $\Pi \in \mathcal{P}_{HG}$ est *individuellement optimale* si :

$$\forall a_i \in N,\ \nexists C \in \mathcal{C}_{a_i}^N : C \succ_{a_i} C_{a_i}^{\Pi}$$

Exemple 2.1.3 - Considérons un jeu $HG = \langle N, \succeq \rangle$ tel que :

$$\begin{aligned} N &= \{a_1,a_2,a_3\} \\ \succeq_{a_1} &= \{a_1,a_2,a_3\} \succ_{a_1} \{a_1,a_2\} \succ_{a_1} \{a_1,a_3\} \succ_{a_1} \{a_1\} \\ \succeq_{a_2} &= \{a_1,a_2,a_3\} \succ_{a_2} \{a_1,a_2\} \succ_{a_2} \{a_2,a_3\} \succ_{a_2} \{a_2\} \\ \succeq_{a_3} &= \{a_1,a_2,a_3\} \succ_{a_3} \{a_1,a_3\} \succ_{a_3} \{a_2,a_3\} \succ_{a_3} \{a_3\} \end{aligned}$$

Dans un tel jeu, la structure de coalitions $\{\{a_1,a_2,a_3\}\}$ est individuellement optimale puisque, pour chaque agent, il n'existe pas de meilleure coalition.

Si l'optimalité individuelle d'une structure de coalitions correspond à un partitionnement parfait des agents, il est fréquent qu'une telle structure de coalitions n'existe pas, surtout lorsque si les agents sont hétérogènes et ont des profils de préférence opposés. En effet, il est fréquent qu'au moins un agent préfère une autre coalition que celle à laquelle il est affecté. Dans ce contexte, un moyen de comparer deux structures de coalitions est la dominance au sens de Pareto. Intuitivement, une structure de coalitions en domine une autre si tous les agents préfèrent leur coalition respective dans première structure comparée à celle de la seconde.

Définition 2.1.7 - *Dominance de Pareto* : Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique. La structure de coalitions $\Pi_1 \in \mathcal{P}_N$ *domine au sens de Pareto* la structure de coalitions $\Pi_2 \in \mathcal{P}_N$ si :

$$\begin{aligned} &\forall a_i \in N,\ C_{a_i}^{\Pi_1} \succeq_{a_i} C_{a_i}^{\Pi_2} \\ &\exists a_i \in N,\ C_{a_i}^{\Pi_1} \succ_{a_i} C_{a_i}^{\Pi_2} \end{aligned}$$

La dominance au sens de Pareto permet de définir un concept de solution optimal au sens de Pareto, garantissant que la structure de coalitions n'est pas dominée par une autre [Drèze et Greenberg, 1980, Pardalos *et al.*, 2008].

Définition 2.1.8 - *Optimalité au sens de Pareto* : Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique. La structure de coalitions $\Pi_1 \in \mathcal{P}_N$ est *optimale au sens de Pareto* si :

$$\nexists \Pi_2 \in \mathcal{P}_N : \forall a_i \in N, C_{a_i}^{\Pi_2} \succeq_{a_i} C_{a_i}^{\Pi_1}$$
$$\exists a_i \in N, C_{a_i}^{\Pi_2} \succ_{a_i} C_{a_i}^{\Pi_1}$$

Exemple 2.1.4 - Considérons un jeu $HG = \langle N, \succeq \rangle$ tel que :

$$N = \{a_1, a_2, a_3\}$$
$$\succeq_{a_1} = \{a_1,a_2\} \succ_{a_1} \{a_1,a_3\} \succ_{a_1} \{a_1,a_2,a_3\} \succ_{a_1} \{a_1\}$$
$$\succeq_{a_2} = \{a_1,a_2\} \succ_{a_2} \{a_2,a_3\} \succ_{a_2} \{a_1,a_2,a_3\} \succ_{a_2} \{a_2\}$$
$$\succeq_{a_3} = \{a_1,a_3\} \succ_{a_3} \{a_2,a_3\} \succ_{a_3} \{a_1,a_2,a_3\} \succ_{a_3} \{a_3\}$$

Considérons le profil de préférence de l'agent a_1 suivant :

$$\succeq_{a_1} = \{a_1,a_2\} \succ_{a_1} \{a_1,a_3\} \succ_{a_1} \{a_1,a_2,a_3\} \succ_{a_1} \{a_1\}$$

La structure de coalitions $\{ \{a_1\},\{a_2\},\{a_3\} \}$ est dominée au sens de Pareto par la structure $\{ \{a_1,a_2\},\{a_3\} \}$ car a_1 et a_2 préfèrent être ensembles et que a_3 ne change pas de coalition. La figure 2.3 montre les dominances au sens de Pareto pour les différentes structures de coalitions. Ici, les trois structures $\{ \{a_1,a_2\},\{a_3\} \}$, $\{ \{a_1,a_3\},\{a_2\} \}$, $\{ \{a_1\},\{a_2,a_3\} \}$ sont optimales au sens de Pareto.

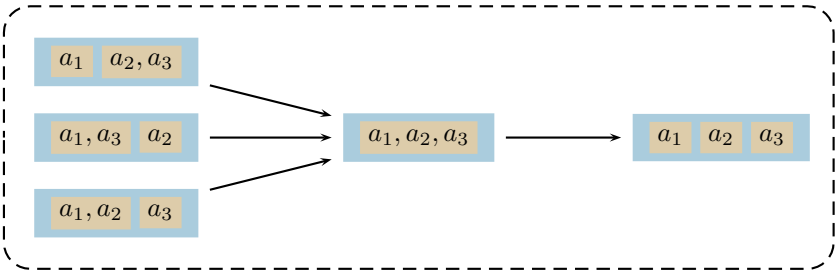

FIGURE 2.3 – Dominance au sens de Pareto des structures de coalitions pour le jeu HG

Intuitivement, l'optimalité au sens de Pareto signifie qu'il n'existe pas d'autre structure de coalitions telle que, pour tout agent, la seconde structure soit préférée. Ce concept de solution permet ainsi de garantir qu'un agent préférant former une autre coalition ne peut pas le faire sans que cela rende un autre agent moins satisfait. Si ce concept de solution est particulièrement intéressant pour maximiser le bien-être social, elle n'a que peu d'intérêt pour des agents égoïstes. C'est pourquoi la notion de *stabilité* d'une structure de coalitions considère les comportements individuels des agents [Nash, 1950, Morgenstern et Von Neumann, 1953]. La stabilité d'une structure de coalitions est définie en opposition à la volonté et la possibilité d'un agent seul ou en groupe à changer de coalition.

Le concept de solution le plus classique est celui de la *stabilité au sens de Nash*. Il regroupe toutes les structures de coalitions telles que, étant donné les coalitions présentes, aucun agent ne préfère quitter sa coalition actuelle pour en rejoindre une autre.

Définition 2.1.9 - *Nash stabilité* : Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique. La structure de coalitions $\Pi \in \mathcal{P}_N$ est *stable au sens de Nash* si et seulement si :

$$\forall a_i \in N,\ \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succeq_{a_i} C^{\Pi}_{a_i}$$

Exemple 2.1.5 - Reprenons l'exemple 2.1.4. La structure de coalitions $\{ \{a_1,a_2,a_3\} \}$ est stable au sens de Nash, car les trois agents préfèrent former la grande coalition plutôt que d'être dans leur coalition singleton. À l'opposé, la structure de coalitions $\{ \{a_1,a_2\},\{a_3\} \}$ n'est pas stable au sens de Nash puisque a_3 préfère quitter sa coalition singleton pour rejoindre les autres agents du jeu.

Dans la suite de ce manuscrit, nous notons NS_{HG} l'ensemble des structures de coalitions stables au sens de Nash du jeu HG. Remarquons que cet ensemble peut être vide (ou être composé de plusieurs structures distinctes). Si la non-stabilité au sens de Nash d'une structure de coalitions repose uniquement sur le désir d'un agent a_i de rejoindre une coalition C et ce indépendamment du fait que les agents de C l'accepte, la *stabilité individuelle* permet de considérer comme stables non seulement les structures de coalitions où aucun agent ne souhaite changer de coalition, mais également les structures où les agents désirant changer de coalition peuvent être refusés par ceux qu'ils désirent rejoindre.

Définition 2.1.10 - *Stabilité individuelle* : Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique. La structure de coalitions $\Pi \in \mathcal{P}_{HG}$ est *individuellement stable* si et seulement si :

$$\forall a_i \in N,\ \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_{a_i} C^{\Pi}_{a_i}$$
$$\forall a_j \in C, C \cup \{a_i\} \succeq_{a_j} C$$

Exemple 2.1.6 - Reprenons l'exemple 2.1.4. La structure de coalitions $\{ \{a_1,a_2\},\{a_3\} \}$ n'est pas stable au sens de Nash, car a_3 préfère rejoindre la coalition $\{a_1,a_2\}$. Par contre, comme ni a_1, ni a_2 ne préfèrent la grande coalition à leurs coalitions actuelles, ils peuvent tous deux rejeter a_3. Par conséquent, la structure de coalitions $\{ \{a_1,a_2\},\{a_3\} \}$ est individuellement stable. Inversement, la structure de coalitions $\{ \{a_1,a_3\},\{a_2\} \}$ n'est pas individuellement stable, car a_1 préfère quitter la coalition $\{a_1,a_3\}$ pour former la coalition $\{a_1,a_2\}$ où il est accepté par a_2.

Dans la suite de ce manuscrit, nous désignerons par IS_{HG} l'ensemble des structures de coalitions individuellement stables. Comme NS_{HG}, IS_{HG} peut être vide. Avec la stabilité individuelle, un agent désirant changer de coalition prend en compte les préférences des agents de la coalition qu'il souhaite rejoindre. Cependant, ceci se fait indépendamment des préférences des agents de la coalition quittée. La *stabilité individuelle contractuelle* définit l'ensemble des structures de coalitions où aucun agent ne peut changer de coalitions s'il est refusé dans la coalition qu'il souhaite rejoindre ou si son départ est refusé par au moins l'un des agents de la coalition qu'il souhaite quitter.

Définition 2.1.11 - *Stabilité individuelle contractuelle* : Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique. La structure de coalitions $\Pi \in \mathcal{P}_{HG}$ est *individuellement contractuellement stable* si

et seulement si :

$$\begin{aligned}\forall a_i \in N,\ \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} &\succ_{a_i} C^{\Pi}_{a_i} \\ \forall a_j \in C, C \cup \{a_i\} &\succeq_{a_j} C \\ \forall a_k \in C^{\Pi}_{a_i}, C^{\Pi}_{a_i} \setminus \{a_i\} &\succeq_{a_k} C^{\Pi}_{a_i}\end{aligned}$$

Exemple 2.1.7 - Reprenons l'exemple 2.1.4. Comme montré précédemment la structure de coalitions $\{\{a_1,a_3\},\{a_2\}\}$ n'est pas individuellement stable, car a_1 souhaite changer de coalition et est accepté par a_2. Par contre, comme a_3 refuse le départ de a_1, elle est individuellement contractuellement stable.

Si ces trois concepts de solution définissent la stabilité en ne considérant que les déviations individuelles des agents, la *stabilité au sens du cœur* permet de considérer comme stable toute structure de coalitions où il n'existe pas de sous-groupes d'agents préférant collectivement quitter leurs coalitions respectives afin de former ensemble une nouvelle coalition [Dimitrov *et al.*, 2006].

Définition 2.1.12 - *Stabilité au sens du cœur* : Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique. La structure de coalitions $\Pi \in \mathcal{P}_{HG}$ est *stable au sens du cœur* si :

$$\nexists N_2 \subseteq N : \forall a_j \in N_2, N_2 \succ_{a_j} C^{\Pi}_{a_j}$$

Exemple 2.1.8 - Reprenons l'exemple 2.1.4. Si la structure de coalitions $\{\{a_1,a_2,a_3\}$ est stable au sens de Nash, car individuellement aucun agent ne désire changer de coalition, elle n'est pas stable au sens du cœur, car les agents a_1 et a_2 préfèrent quitter la grande coalition pour former ensemble la coalition $\{a_1,a_2\}$. Par contre, la structure de coalitions $\{\{a_1,a_2\},\{a_3\}\}$ est stable au sens du cœur, car, même si l'agent a_3 préfère former la grande coalition, ce n'est pas le cas des agents a_1 et a_2.

Nous notons dans la suite CS_{HG} l'ensemble des structures de coalitions stables au sens du cœur pour le jeu hédonique HG. Comme pour NS_{HG}, CS_{HG} peut contenir 0, 1 ou plusieurs structures stables. Enfin, le fait de considérer des agents rationnels permet de définir un dernier concept de solution fondé sur le fait qu'un agent ne va accepter de former une coalition avec d'autres agents que si celle-ci est préférable à l'absence de coopération, c'est-à-dire former sa coalition singleton. Ce concept de solution est appelé la *rationalité individuelle* [Ballester, 2004].

Définition 2.1.13 - *Rationalité individuelle* : Soit $HG = \langle N, \succeq \rangle$ un jeu hédonique. La structure de coalitions $\Pi \in \mathcal{P}_{HG}$ est *individuellement rationnelle* si :

$$\forall a_i \in N, C^{\Pi}_{a_i} \succeq_{a_i} \{a_i\}$$

Exemple 2.1.9 - Soit le jeu hédonique $HG = \langle N, \succeq \rangle$ où :

$$
\begin{aligned}
N &= \{a_1,a_2,a_3\} \\
\succeq_{a_1} &= \{a_1,a_2\} \succ_{a_1} \{a_1,a_2,a_3\} \succ_{a_1} \{a_1\} \succ_{a_1} \{a_1,a_3\} \\
\succeq_{a_2} &= \{a_1,a_2\} \succ_{2} \{a_1,a_2,a_3\} \succ_{a_2} \{a_2\} \succ_{a_2} \{a_2,a_3\} \\
\succeq_{a_3} &= \{a_1,a_3\} \succ_{a_3} \{a_2,a_3\} \succ_{a_3} \{a_3\} \succ_{a_3} \{a_1,a_2,a_3\}
\end{aligned}
$$

Ici, la structure de coalitions $\{ \{a_1,a_2,a_3\} \}$ n'est pas individuellement rationnelle puisque a_3 préfère quitter la grande coalition et ne pas former de coalition. Les structures de coalitions $\{ \{a_1,a_2\},\{a_3\} \}$ et $\{ \{a_1\},\{a_2\},\{a_3\} \}$ sont, elles, individuellement rationnelles.

Remarquons que la structure de coalitions $\{ \{a_1\},\{a_2\},\ldots,\{a_n\} \}$ est toujours individuellement rationnelle.

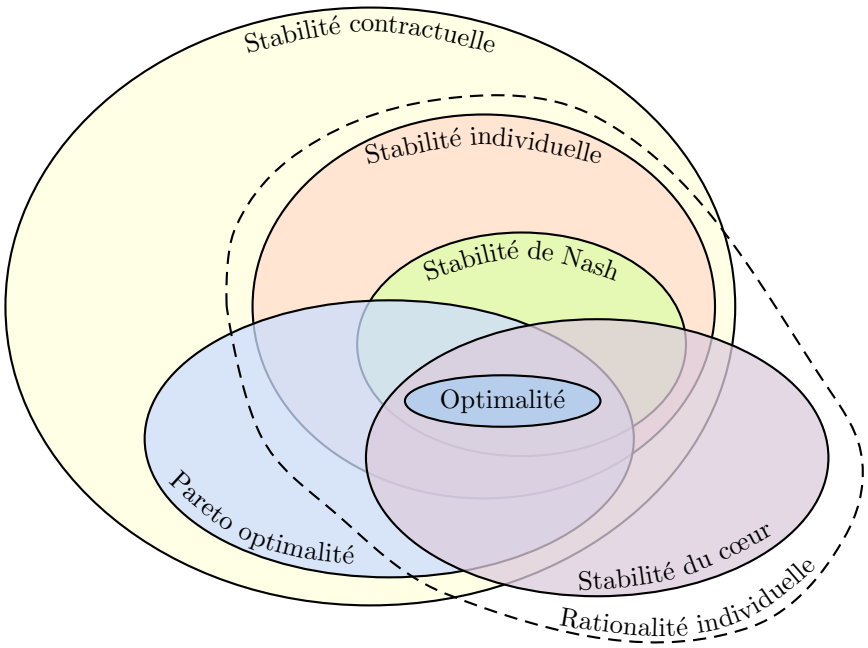

FIGURE 2.4 – Inclusions des différents concepts de solution

Par définition, certains de concepts sont des généralisations des autres : il existe une relation d'inclusion entre les ensembles stable au sens de Nash, individuellement stables et individuellement contractuellement stables [Génin, 2010]. La figure 2.4 résume les relations entre les différents concepts que nous avons présentés. Dans la suite de ce manuscrit, *nous ne considérons que les concepts de stabilité au sens de Nash, stabilité individuelle et stabilité au sens du cœur car les autres concepts de solution soit ne satisfont pas la rationalité individuelle (hypothèse modélisant des agents rationnels), soit sont individuellement optimales (n'incitant pas les agents à manipuler).*

Au-delà de la définition de concepts de solution, prouver l'existence d'une structure de coalitions appartenant à l'un de ces concepts est important [Yun Yeh, 1986, Rothkopf *et al.*, 1998, Ballester, 2004, Elkind et Wooldridge, 2009, Sung et Dimitrov, 2010, Peters et Elkind, 2015]. Le tableau 2.2 résume les résultats de complexité pour les différentes représentations des préférences et concepts de solution que nous considérons.

	CS	NS	IS
IRCL de taille $\leq n^3$	NP-c	NP-c	NP-c
Réseaux de coalitions hédoniques	NP-h	NP-c	NP-c
$\mathcal{W}$-préférences (avec règle de départage)	P	NP-c	?
$\mathcal{W}$-préférences	NP-c	NP-c	NP-c
$\mathcal{W}\beta$-préférences (avec règle de départage)	P	NP-c	?
$\mathcal{W}\beta$-préférences	NP-c	NP-c	NP-c
Jeux à additivité séparable	NP-h	NP-c	NP-c

Tableau 2.2 – Complexité des problèmes d'existence des différents concepts de solution selon les modèles de jeux hédoniques [Peters et Elkind, 2015]

Malgré ces résultats de complexité, de nombreux algorithmes de formation de coalitions ont été proposés [Sandholm, 1999, Larson et Sandholm, 2000, Génin, 2010]. De manière abstraite, ces algorithmes sont des *protocoles de sélection* qui pour un jeu retournent une unique solution.

Définition 2.1.14 - *Protocole de sélection* : Un *protocole de sélection* $\mathbb{P}$ est une fonction qui prend en entrée un jeu de coalitions HG et retourne une structure de coalitions unique $\mathbb{P}(HG)$, appelée la solution du jeu HG.

Dans la littérature en dehors des approches naïves qui consiste à énumérer toutes les structures de coalitions possibles [Yun Yeh, 1986,Rothkopf *et al.*, 1998], nous distinguons deux grandes familles de protocoles de sélection : ceux cherchant à maximiser le bien-être social en considérant les agents du système comme coopératif et ceux cherchant à définir une structure de coalitions acceptable par tous les agents.

- les *algorithmes de génération des structures de coalitions* sont des algorithmes *any-time* centralisés et, bien souvent, distribués permettant de trouver une structure de coalitions proche de l'optimum en termes de bien-être social sans avoir à parcourir nécessairement l'ensemble de structures de coalitions existantes [Sandholm, 1999, Larson et Sandholm, 2000, Dang et Jennings, 2004, Rahwan, 2007, Rahwan et Jennings, 2007, Rahwan *et al.*, 2007, Rahwan et Jennings, 2008, Keinänen, 2010, Michalak *et al.*, 2010] ;
- les *protocoles de négociation* sont des algorithmes décentralisés dans lesquels les agents échangent des propositions de coalitions à former avec, selon les protocoles, la possibilité de quitter une coalition en cours de formation ou non [Kelso Jr et Crawford, 1982, Vauvert et El Fallah-Seghrouchni, 2001, Aknine *et al.*, 2004, Génin, 2010, Génin et Aknine, 2011].

2.2 Systèmes de réputation

Dans un système multi-agents de grande taille, chaque agent dispose de ressources privées et de capacités d'action propres. Si les agents peuvent mettre en commun leurs ressources en formant une coalition, ils peuvent également déléguer la réalisation d'une tâche à un autre agent du système. L'une des problématiques est alors de décider à quel agent faire *confiance* pour réaliser correctement cette tâche. Ainsi, les *systèmes de réputation* désignent des systèmes multi-agents où les agents disposent d'un processus leur permettant de prendre une telle décision.

2.2.1 Confiance et réputation

De manière générale, les systèmes de réputation sont des systèmes où les agents interagissent, collectent, partagent et agrègent les résultats de leurs interactions passées afin de décider à quels agents ils peuvent faire confiance pour de futures interactions. [Sabater et Sierra, 2005, Marti et Garcia-Molina, 2006, Artz et Gil, 2007, Jøsang *et al.*, 2007, Hoffman *et al.*, 2009, Pinyol et Sabater-Mir, 2013] ont proposé différentes classifications des systèmes de réputation et [Resnick *et al.*, 2000] ont identifié les trois axiomes qui définissent ce type de système :

1. les agents doivent être persistants dans le temps ;
2. les résultats des interactions doivent communiqués aux autres agents ;
3. le processus de décision doit être guidé par ces résultats d'interaction.

Il existe de nombreuses définitions de la confiance dans les systèmes multi-agents [Stephen, 1994, Grandison et Sloman, 2000, Mui *et al.*, 2002, Xiu et Liu, 2005, Golbeck, 2006, Jøsang *et al.*, 2007] allant d'une mesure de la capacité d'un agent à accepter de dépendre d'un autre, à une estimation de la capacité d'un agent à réaliser de manière fiable une tâche dans un contexte donné, en passant par une mesure du risque pris en déléguant une tâche. Dans toutes ces définitions, l'agent utilise le résultat de ces interactions passées obtenir cette estimation. C'est pourquoi, de manière générique, nous considérons dans ce manuscrit la définition de la confiance de [Mui *et al.*, 2002] :

Définition 2.2.1 - *Confiance* : La confiance est une estimation subjective du comportement futur d'un agent fondé sur l'historique des interactions passées.

Les modèles de confiance sont nombreux :
- des valeurs uniques associées à un seuil indiquant s'il y a confiance ou non [Stephen, 1994] ;
- des valeurs uniques définies dans des ensembles continus [Stephen, 1994, Kamvar *et al.*, 2003, Cheng et Friedman, 2005, Jøsang *et al.*, 2007, Zhou et Hwang, 2007] ;
- des tuples de valeurs continues représentant différents critères d'évaluation [Sabater et Sierra, 2001, Jøsang et Ismail, 2002, Theodorakopoulos et Baras, 2006, Jin *et al.*, 2007] ou modélisant une appartenance à des ensembles flous [Sabater *et al.*, 2006] ;
- des valeurs discrètes par exemple très mauvais, mauvais, neutre, bon, très bon [Abdul-Rahman et Hailes, 2000].

La confiance peut être *dépendante au contexte*, signifiant qu'il y a une valeur de confiance par agent pour chaque type d'interaction possible [Page *et al.*, 1999, Sabater *et al.*, 2006, Danezis et Mittal, 2009]. De manière générale, les agents ne disposant pas obligatoirement des informations nécessaires au calcul d'une confiance, les systèmes de réputation leur permettent de partager une partie de leurs informations privées. Lorsque cela arrive, l'agent partageant une information fournit un *témoignage*. Dans la littérature, les témoignages des agents sont étudiés selon deux axes : la dissémination et la représentation.
- La dissémination porte sur la méthode utilisée par les agents pour partager les témoignages. Nous distinguons principalement deux approches : *centralisée* lorsqu'une autorité centrale est chargée de collecter les témoignages des agents pour ensuite les partager avec les agents du système [Schafer *et al.*, 1999, Carbo *et al.*, 2002, Srivatsa *et al.*, 2005, Zhou et Hwang, 2007, Jøsang et Haller, 2007] et *décentralisée* lorsque les agents fournissent leurs témoignages directement aux autres agents du système [Sabater et Sierra, 2001, Xiong et Liu, 2004, Sabater *et al.*, 2006, Zhou et Hwang, 2007] ;
- La représentation des témoignages correspond au fait que les agents partagent directement le résultat de chaque interaction comme dans eBay [Schafer *et al.*, 1999], ou une agrégation

des confiances et des témoignages déjà reçus [Jøsang et Ismail, 2002, Kamvar *et al.*, 2003, Cheng et Friedman, 2005].

Notons que les agents peuvent partager leurs *observations personnelles,* mais également des témoignages que d'autres agents leur ont eux-mêmes fournis. C'est ainsi que nous pouvons parler de témoignages *directs* et de témoignages *indirects* [Sabater et Sierra, 2005]. À partir de ses observations personnelles et des témoignages (directs ou indirects) reçus, un agent (ou une autorité centrale dans le cas de système centralisé) peut les agréger afin de calculer la *réputation* des agents. Si la confiance est définie comme une estimation du comportement des agents fondés sur ses propres observations, [Wang et Vassileva, 2003] définissent la réputation d'un agent comme la croyance d'un agent en la capacité, de l'honnêteté et la fiabilité d'un autre agent en se fondant sur les témoignages qu'il a reçus.

De manière générique, nous considérons dans ce manuscrit la définition suivante de la réputation :

Définition 2.2.2 - *Réputation* : La *réputation* d'un agent est une agrégation des témoignages des autres agents envers lui, représentant une estimation collective de son comportement futur.

La *fonction de réputation* désigne l'algorithme utilisé par les agents (ou l'autorité centrale) pour calculer les valeurs de réputation. Elles peuvent être *globales* ou *personnalisées* [Sabater et Sierra, 2005, Pinyol et Sabater-Mir, 2013]. Dans le premier cas, la réputation d'un agent est définie indépendamment de celui l'évalue et est donc la même du point de vue de tout agent [Schafer *et al.*, 1999, Kamvar *et al.*, 2003, Zhou et Hwang, 2007]. Dans le second cas, la réputation est différente du point de vue de chaque agent [Jøsang et Ismail, 2002, Sabater et Sierra, 2001, Cheng et Friedman, 2005, Srivatsa *et al.*, 2005]. Les fonctions de réputation personnalisée modélisent le fait que l'évaluation d'une interaction est subjective et que ce qui peut paraître comme une bonne interaction pour un agent ne l'est pas nécessairement pour un autre.

Il existe aussi deux autres catégories de fonction de réputation selon qu'elles sont *symétriques* ou *asymétriques* [Cheng et Friedman, 2005] . Cette distinction repose sur une représentation de la confiance des agents sous forme d'un graphe orienté valué G où les nœuds désignent les agents, les arcs des interactions passés et leurs poids sont la valeur de confiance. Ce graphe est appelé le *graphe de confiance* et une fonction de réputation est asymétrique si, pour tout graphe G' isomorphe au graphe G, la réputation de l'agent a_i est identique à celle de son image sur G'.

Remarquons qu'il y a un lien entre ces catégories : une fonction personnalisée est nécessairement asymétrique tandis qu'une fonction globale peut être symétrique ou asymétrique. De plus, une fonction de réputation peut donner un *rang de réputation* ou une *valeur de réputation.* Un rang de réputation permet d'ordonner qualitativement les agents entre eux. Une valeur de réputation permet non de les ordonner, mais y associe un sens quantitatif.

Les fonctions de réputation se fondent sur différents critères [Xiong et Liu, 2004, Sabater et Sierra, 2005, Srivatsa *et al.*, 2005, Hoffman *et al.*, 2009] : une distinction entre les témoignages directs et indirects, une distinction entre les rôles des agents dans le système ou leur appartenance de l'agent à un groupe, une distinction entre les types d'interaction, un facteur d'oubli des informations anciennes. Par exemple, [Kamvar *et al.*, 2003] proposent l'utilisation d'*agents de confiance* qui disposent d'une valeur de confiance minimale par défaut : leur témoignage a donc initialement plus de poids que les témoignages des autres agents.

2.2.2 Exemples de systèmes de réputation

À titre d'exemple, nous présentons ici trois systèmes de réputation fortement étudiés dans la littérature : BetaReputation [Jøsang et Ismail, 2002], EigenTrust [Kamvar *et al.*, 2003] et FlowTrust [Cheng et Friedman, 2005].

BetaReputation : Système de réputation est fondé sur une approche bayésienne et produisant une valeur de réputation. La confiance est modélisée par un couple $\langle r_{ij},s_{ij}\rangle$ correspondant respectivement à la partie positive et négative de l'évaluation d'un agent a_i des interactions qu'il a eues avec un agent a_j. Ces deux valeurs doivent appartenir à un même domaine de définition fini et doivent correspondre au gain réel positif et négatif obtenu alors d'une interaction. La réputation d'un agent est alors modélisée par une fonction de densité beta. Sémantiquement, la réputation correspond à la valeur espérée de la qualité d'une future interaction avec cet agent. Lorsque l'agent a_i reçoit un témoignage $\langle r_{jk},s_{jk}\rangle$ de l'agent a_j vis-à-vis de l'agent a_k, il l'agrège avec ses propres observations comme suit :

$$r_k^{i:j} = \frac{2r_{ij}r_{jk}}{(s_{ij}+2)(r_{jk}+s_{jk}+2)+2r_{ij}}$$
$$s_k^{i:j} = \frac{2r_{ij}s_{jk}}{(s_{ij}+2)(r_{jk}+s_{jk}+2)+2r_{ij}}$$

Intuitivement, lorsque l'agent a_i reçoit un témoignage provenant de l'agent a_j, le témoignage est pondéré par la confiance que a_i à envers l'agent a_j. De ce fait, BetaReputation utilise une fonction de réputation personnalisée. En effet, la réputation de a_k (notée $Rep(r_k,s_k)$) est calculée à partir de l'agrégation de l'ensemble des témoignages reçus par la fonction :

$$Rep(r_k,s_k) = \frac{r_k - s_k}{r_k + s_k + 2}$$

Des extensions de systèmes existent en y intégrant un facteur d'oubli $\lambda \in [0,1]$ pondérant l'importance des interactions les plus anciennes ou un une troisième composante de la confiance u_{ij} représentant un degré d'incertitude lors de l'évaluation des interactions.

EigenTrust : Système de réputation global asymétrique inspiré du Google PageRank [Page *et al.*, 1999] qui utilise la matrice d'adjacente C du graphe de confiance et produit un rang de réputation. La confiance y est modélisée par la somme des interactions satisfaisantes, notée $sat(i,j)$, moins la somme des interactions non satisfaisantes, notée $unsat(i,j)$. Cette valeur de confiance, notée $s_{ij} \in \mathbb{Z}$, est ensuite normalisée en une valeur $c_{i,j} \in [0,1]$. Il s'agit donc d'une répartition d'un poids entre tous les agents. De plus, EigenTrust attribue sous forme d'un vecteur $\vec{p}$ une valeur minimale par défaut à des agents de confiance. Pour un paramètre d'exploration $a \in [0,1]$, la réputation d'agent (représenté au sein d'un vecteur $\vec{t}$) est la probabilité qu'un marcheur aléatoire sur le graphe de confiance partant de l'agent a_i s'arrête sur l'agent a_j. C'est le point fixe de la fonction ci-dessous lorsque k est incrémenté :

$$\vec{t}^{k+1} = (1-a)C^T\vec{t}^{k} + a\vec{p}$$

FlowTrust : Système de réputation asymétrique personnalisé se fondant sur le graphe de confiance. La confiance c_{ij} est une valeur réelle unique abstraite. La réputation de a_k est le flot maximal pour l'ensemble des collections de chemins disjoints $\mathbb{P}_i$ allant de a_i vers a_k sur le graphe de confiance G :

$$f(G,k)_i = \max_{\mathcal{P}_{i,k}\in\mathbb{P}_i} \sum_{P\in\mathcal{P}_{i,k}} \min\{c_{xy}|(x,y)\in P\}$$

Ce système de réputation a la particularité d'être robuste à un certain nombre de manipulations sous des contraintes très restrictives.

2.2.3 Manipulations et stratégies de défense

S'il existe de nombreux systèmes de réputation, l'une des principales problématiques au-delà de la formalisation et de l'algorithmique est l'étude de la robustesse de ces systèmes aux manipulations [Levien et Aiken, 1998, Adar et Huberman, 2000, Sabater et Sierra, 2001, Cheng et Friedman, 2005, Srivatsa *et al.*, 2005, Feldman *et al.*, 2006, Levine *et al.*, 2006, Sabater *et al.*, 2006, Altman et Tennenholtz, 2007b, Danezis et Mittal, 2009, Koutrouli et Tsalgatidou, 2011]. Bien que certaines études s'intéressent à des manipulations de type positionnement stratégique (attaque éclipse, déni de service, prolifération, resquillage [Adar et Huberman, 2000]), ces manipulations ne sont pas spécifiques aux systèmes de réputation, car elles dépendent du contexte applicatif.

Dans le contexte des systèmes de réputation, [Hoffman *et al.*, 2009, Jøsang et Golbeck, 2009, Tavakolifard et Almeroth, 2012] ont proposé des taxonomies des différentes manipulations. Cependant, dans ce manuscrit et au regard de notre problématique, nous les classons en fonction des axiomes fondamentaux identifiés par [Resnick *et al.*, 2000] pour caractériser un système de réputation.

Axiome 1 (les agents doivent être persistants dans le temps) : Trois manipulations s'attaquent à cet axiome, le *blanchiment* [Feldman *et al.*, 2006], l'*attaque Sybil* [Douceur, 2002] et la *discrimination* [Jøsang et Golbeck, 2009]. Le blanchiment consiste à quitter le système lorsqu'une valeur de réputation est trop faible afin de s'y réintroduire sous une nouvelle identité ayant la même réputation qu'un agent nouvel entrant. Les attaques Sybil consistent simplement à s'introduire dans le système sous de multiples fausses identités. La discrimination consiste à cibler les interactions : toujours bien interagir avec les agents persistants dans le temps (comme des agents de confiance) et toujours mal interagir avec les agents identifiés comme peu persistants.

Axiome 2 (les résultats des interactions doivent être communiqués aux autres agents et sont accessibles dans le futur) : Trois manipulations s'attaquent à cet axiome, la *promotion*, l'*autopromotion* et la *diffamation* qui consistent à partager avec les autres agents du système des *faux témoignages* afin que le calcul des valeurs de réputation soit à l'avantage de l'agent malhonnête [Jin *et al.*, 2007]. Nous distinguons ici la promotion qui est effectuée par des agents en collusion, de l'autopromotion qui consiste à utiliser des agents Sybil.

Axiome 3 (le processus de décision doit être guidé par ces résultats d'interaction) : Une manipulation s'attaque à cet axiome, la *traîtrise* ou attaque planifiée [Jøsang et Golbeck, 2009], qui consiste à augmenter la valeur de réputation de l'agent malveillant par une succession de bonnes interactions avant d'effectuer volontairement une mauvaise interaction à un instant précis.

Toutes ces manipulations peuvent être combinées entre elles et s'attaquent simultanément à tous les axiomes. Par exemple, l'*attaque oscillante* combine promotion, diffamation et traîtrise, voire dans certains cas du blanchiment : lorsqu'un agent voit sa réputation devenir trop faible, il change d'identité tout en changeant de rôle.

Pour lutter contre les manipulations, de nombreuses stratégies de défense ont été proposées dans la littérature [Levine *et al.*, 2006, Hoffman *et al.*, 2009]. Comme pour les manipulations, nous proposons ici une classification de ces stratégies en fonction des axiomes du système (soit

de [Resnick *et al.*, 2000], soit [Altman et Tennenholtz, 2007b]) qu'elles renforcent ou remettent en cause.

Axiome 1 (les agents doivent être persistants dans le temps) : Certaines techniques renforcent la persistance faisant l'hypothèse de l'existence d'agents toujours de confiance [Kamvar *et al.*, 2003] tandis que d'autres s'assurent qu'aucune identité ne persiste dans le temps, par exemple en forçant l'utilisation d'identifiant jetable à court terme, pour éviter les discriminations [Dellarocas, 2000, Singh et Liu, 2003].

Axiome 2 (les résultats des interactions sont communiqués aux autres agents et sont accessibles dans le futur) : Ces techniques consistent soit à introduire une notion d'oubli pour éviter la persistance des informations et détecter rapidement les traîtrises (bien que cela rende plus sensible le système aux attaques oscillantes, car il suffit d'un faible nombre de bonnes interactions pour retrouver une bonne valeur de réputation) [Jøsang et Ismail, 2002], soit certifier l'origine et le contenu des messages (bien que cela ne fournisse aucune protection contre les promotions où plusieurs agents malveillants se mettent d'accord pour fabriquer de fausses interactions) [Srivatsa *et al.*, 2005].

Axiome d'ouverture (le système multi-agents est ouvert et décentralisé) : Ces techniques consistent à rendre coûteuse l'utilisation du système pour limiter les attaques Sybil et le blanchiment. Ces techniques sont fondées sur des défis cryptographiques [Douceur, 2002, Borisov, 2006] ou des coûts monétaires [Von Ahn *et al.*, 2003, Margolin et Levine, 2008].

Axiome de rationalité (les agents prennent leurs décisions rationnellement) : Certaines techniques font une hypothèse de rationalité sur le comportement des agents malveillants pour les forcer à changer de stratégie. Par exemple, [Miller *et al.*, 2005, Friedman *et al.*, 2007] proposent des mécanismes de paiement incitant les agents à fournir de vrais témoignages et [Bonnet, 2012] se sert d'une fonction d'agrégation de témoignages qui force les agents malveillants à adopter une stratégie stochastique pour maximiser leur gain.

Axiome de réponse positive (un témoignage positif doit augmenter la réputation de l'agent recommandé) : Quelques techniques proposent des fonctions de réputation dans lesquelles, pour limiter la promotion et la diffamation, les opérateurs d'agrégation des témoignages sont sous-additifs [Cheng et Friedman, 2005].

Axiome de structure (le graphe confiance est un réseau social) : Ces techniques utilisent toutes les propriétés des réseaux sociaux pour extraire les groupes suspects dans leurs interactions [Newsome *et al.*, 2004, Yu *et al.*, 2006, Staab et Engel, 2009]. Si ces propriétés sont vérifiées alors ces techniques sont très efficaces contre des agents en collusion effectuant de la promotion. En revanche, un positionnement stratégique permet de mettre en défaut cette défense.

Une approche orthogonale est l'utilisation d'une notion de crédibilité. Dans de nombreux systèmes de réputation [Mui *et al.*, 2002, Kamvar *et al.*, 2003, Cheng et Friedman, 2005, Yu *et al.*, 2006], les témoignages d'un agent a_j sont pondérés par la confiance que l'agent a_i à envers a_j. Ainsi, la confiance joue un double rôle : mesurer la fiabilité de l'agent a_j lors de ses interactions et mesurer sa fiabilité lorsqu'il communique un témoignage. La crédibilité consiste à définir une seconde mesure de confiance spécifique à la production de témoignage. Ces mesures de crédibilité peuvent prendre plusieurs formes : un degré de similarité entre le témoignage reçu et les observations directes de l'agent [Sabater et Sierra, 2001, Srivatsa *et al.*, 2005, Koutrouli et Tsalgatidou, 2011], un degré de similarité entre le témoignage reçu et le résultat de l'interaction suivante [Zhao et Li, 2009], une quantité d'inconsistance entre plusieurs témoignages reçus par

plusieurs agents [Muller et Vercouter, 2004, Muller et Vercouter, 2005], le gain d'information produit par l'acceptation du témoignage reçu [Whitby *et al.*, 2004].

Comme des valeurs de confiance, les valeurs de crédibilité viennent affecter l'agrégation des témoignages, soit en pondérant les témoignages [Sabater et Sierra, 2001, Srivatsa *et al.*, 2005, Koutrouli et Tsalgatidou, 2011], soit en les filtrant et les retirant directement du processus d'agrégation [Muller et Vercouter, 2004, Whitby *et al.*, 2004, Zhao et Li, 2009]. Remarquons que certaines techniques de filtrage sont drastiques : un témoignage filtré implique de filtrer tous les témoignages de l'agent qui l'a produit [Muller et Vercouter, 2004].

2.3 Positionnement de ce manuscrit

Dans ce chapitre, nous avons présenté les deux systèmes multi-agents que nous considérons dans ce manuscrit.

Le premier est un jeu de formation de coalitions et, plus précisément, un *jeu de coalitions hédoniques*. Dans ce jeu, chaque agent définit des préférences vis-à-vis des différentes coalitions qu'il pourrait rejoindre. L'objectif est de calculer une *structure de coalitions* satisfaisante pour l'ensemble des agents à partir de leur *profil de préférence*. À notre connaissance, les travaux portant sur la robustesse des jeux de coalitions se concentrent uniquement sur des jeux à utilité transférable. C'est pourquoi nous avons retenu les jeux hédoniques (à utilité non transférable) et nous proposons, comme étape préalable à tout travail de conception de protocole, d'étudier les caractéristiques et les propriétés des manipulations sur ce type de jeu. En particulier, nous proposons dans le chapitre 3 un modèle de manipulation générique (cumulant attaques Sybil et faux témoignages). Nous étudions ensuite dans les chapitres 4 et 5 la robustesse des jeux de coalitions hédoniques en fonction des concepts de solution associés et de certaines hypothèses que nous pouvons faire sur les agents.

Le second système que nous considérons dans ce manuscrit est un *système de réputation*. Il s'agit pour les agents de partager le résultat de leurs interactions passées et de les agréger en une valeur de réputation. Les agents utilisent ensuite cette valeur pour décider avec qui interagir dans le futur. Contrairement aux jeux de coalitions hédoniques, la question de leur robustesse a été largement étudiée. Toutefois, nous avons remarqué qu'aucune technique proposée ne s'intéresse au troisième axiome de Resnick, c'est-à-dire à l'utilisation des valeurs de réputation dans le processus de décision de l'agent comme technique de défense. C'est pourquoi, dans le chapitre 6, nous proposons un modèle d'interaction entre agents utilisant un système de réputation et les différentes manipulations que nous considérons et, dans le chapitre 7, nous étudions comment l'utilisation de la valeur réputation dans le processus de décision influe sur la robustesse du système. De plus, une seconde approche nous a paru intéressante : l'utilisation d'une notion de crédibilité. En effet, la crédibilité ne repose pas sur les axiomes classiquement utilisés dans les systèmes de réputation et peut être a priori appliquée sur toute fonction de réputation. Parmi les approches proposées dans la littérature, considérer le gain d'information à l'acceptation des témoignages a été peu étudié. Dans le chapitre 8, nous proposons alors une nouvelle mesure de crédibilité générique fondée sur la *divergence de Kullback-Leibler*, de l'information perdue lorsqu'un faux témoignage est accepté. Cette approche n'a que peu de sens lorsque la crédibilité est un facteur de pondération, c'est pourquoi, nous proposons des fonctions de filtrage pour retirer du processus d'agrégation les témoignages jugés comme faux.

Deuxième partie

Manipulations d'un modèle de préférence - les jeux hédoniques

Chapitre 3

Un modèle de manipulation

Sommaire

Résumé.
Comme nous l'avons vu dans la section 2.1, le problème de la formation de coalitions a été étudié sous de nombreuses formes. Dans ce chapitre, nous nous intéressons plus spécifiquement aux *jeux hédoniques* où les agents cherchent à former des structures de coalitions respectant leurs préférences vis-à-vis des différentes coalitions possibles. À partir d'un modèle générique de jeu hédonique, nous présentons comment des agents rationnels peuvent individuellement calculer leurs *degrés de satisfaction* quant à la solution du jeu. Si le respect de certains concepts de solution permet de garantir certaines propriétés vis-à-vis de la solution, celle-ci est généralement sous-optimale pour au moins un agent. Un agent malhonnête peut se demander s'il existe une stratégie lui permettant d'altérer le jeu afin d'augmenter sa *probabilité de satisfaction*. Nous définissons donc dans la seconde partie de ce chapitre un modèle générique de *manipulation* se basant à la fois sur de *faux témoignages* et sur l'introduction dans le système d'*agents Sybil*. Nous montrons comment à partir d'hypothèses sur les autres agents du jeu, il est possible pour un agent malhonnête de calculer s'il est *rationnel* de mettre en œuvre une telle manipulation.

Afin de simplifier la lecture de ce chapitre, le tableau 3.1 résume les principales notations utilisées. Nous donnons en annexe A.1 les notations portant sur les jeux de coalitions hédoniques utilisés dans ce manuscrit.

Notation d'un jeu hédonique	
$HG = \langle N, \succeq, \mathbb{P} \rangle$	Un jeu hédonique
$N = \{a_1, \dots a_n\}$	Ensemble des agents du jeu
$\succeq = \{\succeq_{a_1}, \dots \succeq_{a_n}\}$	Ensemble des profils de préférence
$\mathcal{P}_N$	Ensemble des structures de coalitions possibles de N
$C_{a_i}^{\Pi}$	Coalition de l'agent a_i dans la structure $\Pi \in \mathcal{P}_N$
$\mathcal{C}_{a_i}^N \subseteq 2^N$	Sous-ensemble des coalitions auxquelles l'agent a_i appartient
Concepts de solution	
$S_{HG} \subseteq \mathcal{P}_N$	Structures de coalitions satisfaisant le concept de solution S
$NS_{HG} \subseteq \mathcal{P}_N$	Structures de coalitions stables au sens de Nash
$IS_{HG} \subseteq \mathcal{P}_N$	Structures de coalitions individuellement stables
$CS_{HG} \subseteq \mathcal{P}_N$	Structures de coalitions stables au sens du cœur
Solution du jeu	
$\mathbb{P}(HG)$	Solution du jeu hédonique HG
$AP_{a_i}^x(HG) \subseteq \mathcal{P}_N$	Structures de coalitions acceptables par a_i à une profondeur x,
$\mathrm{P}^*(a_i, x \mid HG)$	Probabilité que $\mathbb{P}(HG)$ appartienne à $AP_{a_i}^x(HG)$.

Tableau 3.1 – Principales notations des jeux hédoniques

3.1 Coopération entre agents égoïstes

3.1.1 Éléments du modèle

Dans ce chapitre, nous considérons le cadre des jeux hédoniques que nous avons présenté en section 2.1. Il s'agit pour un ensemble d'agents de construire une structure de coalitions, à partir des profils de préférence des agents vis-à-vis des coalitions. Comme nous ne nous intéressons pas à la définition des profils de préférence des agents, nous considérons que chaque agent dispose préalablement d'un tel profil à l'initialisation d'un jeu hédonique. Afin d'illustrer notre propos, nous prenons l'exemple suivant :

Exemple 3.1.1 - Soit le jeu hédonique $HG = \langle N, \succeq \rangle$ avec :

$$
\begin{aligned}
N &= \{a_1,a_2,a_3,a_4\} \\
\succeq_{a_1} &= \{a_1,a_2\} \succ_{a_1} \{a_1,a_3,a_4\} \succ_{a_1} \{a_1,a_3\} \sim_{a_1} \{a_1,a_2,a_4\} \succ_{a_1} \{a_1,a_2,a_3,a_4\} \sim_{a_1} \{a_1\} \\
&\quad \succ_{a_1} \{a_1,a_4\} \sim_{a_1} \{a_1,a_2,a_3\} \\
\succeq_{a_2} &= \{a_1,a_2\} \sim_{a_2} \{a_2,a_3,a_4\} \succ_{a_2} \{a_1,a_2,a_3\} \sim_{a_2} \{a_2,a_4\} \succ_{a_2} \{a_1,a_2,a_4\} \succ_{a_2} \{a_2,a_3\} \\
&\quad \succ_{a_2} \{a_1,a_2,a_3,a_4\} \sim_{a_2} \{a_2\} \\
\succeq_{a_3} &= \{a_1,a_3\} \sim_{a_3} \{a_2,a_3,a_4\} \succ_{a_3} \{a_3,a_4\} \succ_{a_3} \{a_1,a_2,a_3\} \succ_{a_3} \{a_2,a_3\} \\
&\quad \succ_{a_3} \{a_1,a_2,a_3,a_4\} \sim_{a_3} \{a_3\} \succ_{a_3} \{a_1,a_3,a_4\} \\
\succeq_{a_4} &= \{a_1,a_4\} \succ_{a_4} \{a_2,a_4\} \succ_{a_4} \{a_3,a_4\} \succ_{a_4} \{a_4\} \succ_{a_4} \{a_1,a_2,a_4\} \succ_{a_4} \{a_1,a_3,a_4\} \\
&\quad \sim_{a_4} \{a_2,a_3,a_4\} \succ_{a_4} \{a_1,a_2,a_3,a_4\}
\end{aligned}
$$

Ce jeu possède 15 structures de coalitions possibles :

$$\Pi_1 = \{ \{a_1\},\{a_2\},\{a_3\},\{a_4\} \}$$
$$\Pi_2 = \{ \{a_1,a_2\},\{a_3\},\{a_4\} \}$$
$$\Pi_3 = \{ \{a_1,a_3\},\{a_2\},\{a_4\} \}$$
$$\Pi_4 = \{ \{a_1,a_4\},\{a_2\},\{a_3\} \}$$
$$\Pi_5 = \{ \{a_1\},\{a_2,a_3\},\{a_4\} \}$$
$$\Pi_6 = \{ \{a_1\},\{a_2,a_4\},\{a_3\} \}$$
$$\Pi_7 = \{ \{a_1\},\{a_2\},\{a_3,a_4\} \}$$
$$\Pi_8 = \{ \{a_1,a_2\},\{a_3,a_4\} \}$$
$$\Pi_9 = \{ \{a_1,a_3\},\{a_2,a_4\} \}$$
$$\Pi_{10} = \{ \{a_1,a_4\},\{a_2,a_3\} \}$$
$$\Pi_{11} = \{ \{a_1,a_2,a_3\},\{a_4\} \}$$
$$\Pi_{12} = \{ \{a_1,a_2,a_4\},\{a_3\} \}$$
$$\Pi_{13} = \{ \{a_1,a_3,a_4\},\{a_2\} \}$$
$$\Pi_{14} = \{ \{a_1\},\{a_2,a_3,a_4\} \}$$
$$\Pi_{15} = \{ \{a_1,a_2,a_3,a_4\} \}$$

En échangeant tout ou partie de leurs profils de préférence, les agents vont définir collectivement la solution du jeu, c'est-à-dire la structure de coalitions qu'ils vont former. Ces échanges de profils de préférence sont définis par un protocole de sélection associé au jeu hédonique. Comme nous l'avons vu dans la section 2.1.2, de nombreux protocoles de sélection ont été proposés. Nous ne nous intéressons pas ici à la définition d'un nouveau protocole, mais aux propriétés qui sont respectées par la solution qu'il retourne. Dans la suite de ce manuscrit, nous considérons la définition suivante pour un protocole de sélection :

Définition 3.1.1 - *Protocole de sélection* : Un *protocole de sélection* $\mathbb{P}$ est une fonction qui est associée à un jeu HG et retourne une structure de coalitions unique, éventuellement vide, notée $\mathbb{P}(HG)$.

Dans la suite de ce manuscrit, nous définissons un jeu hédonique par le triplet $HG = \langle N, \succeq, \mathbb{P} \rangle$ où N désigne l'ensemble des agents, $\succeq$ leurs profils de préférence et $\mathbb{P}$ le protocole de sélection utilisé par les agents pour décider collectivement de la solution. Les propriétés de la solution d'un jeu sont caractérisées par le concept de solution auquel la solution appartient. Il existe de nombreux concepts de solution que nous avons détaillés en section 2.1.2. Un protocole de sélection garantissant que la solution du jeu appartient à un concept de solution donné est dit satisfaisant ce concept.

Définition 3.1.2 - *Satisfaction d'un concept de solution* : Soit $S_{HG} \subseteq \mathcal{P}_N$ l'ensemble des structures de coalitions de N appartenant au concept de solution S. Un protocole sélection $\mathbb{P}$ *satisfait le concept de solution* S si :

$$\forall HG = \langle N, \succeq, \mathbb{P} \rangle, S_{HG} \neq \emptyset \implies \mathbb{P}(HG) \in S_{HG}$$

Exemple 3.1.2 - Reprenons l'exemple 3.1.1. Si le protocole de sélection $\mathbb{P}$ satisfait la stabilité au sens de Nash (définition 2.1.9), la solution du jeu appartient nécessairement à l'ensemble des structures de coalitions de la figure 3.1 :

$$\mathbb{P}(HG) \in NS_{HG} = \{\Pi_8, \Pi_9\}$$

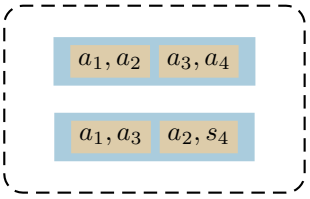

FIGURE 3.1 – Structures de coalitions stables au sens de Nash de l'exemple 3.1.1

3.1.2 Satisfaction individuelle

Dans ce manuscrit, nous considérons que les agents sont rationnels, c'est-à-dire qu'un agent ne va accepter la solution d'un jeu que si celle-ci est satisfaisante de son point de vue, indépendamment des préférences des autres agents. De ce fait, un agent ne va accepter de coopérer avec d'autres agents en formant une coalition que s'il considère celle-ci comme préférable à l'absence de coopération, c'est-à-dire former la coalition singleton. Cette propriété est une propriété de rationalité individuelle (définition 2.1.13). Nous considérons donc que le protocole de sélection doit satisfaire au minimum cette rationalité individuelle.

Hypothèse 3.1.1 - *Rationalité individuelle minimale* : Le protocole de sélection $\mathbb{P}$ *satisfait au minimum la rationalité individuelle* :

$$\forall HG = \langle N, \succeq, \mathbb{P} \rangle, S_{HG} = \emptyset \implies \mathbb{P}(HG) = \{ \{a_1\}, \ldots, \{a_n\} \}$$

Cette hypothèse signifie que si le concept de solution est vide alors le protocole de sélection doit renvoyer la structure de coalitions uniquement composée de coalitions singletons. Le fait de considérer que le protocole de sélection satisfait au minimum la rationalité individuelle permet de représenter les profils de préférence des agents de manière plus compacte, car toutes les coalitions qu'un agent considère comme moins préférées à la coalition singleton ne peuvent pas se former et n'ont donc pas d'intérêt à être représentées. C'est pourquoi, dans la suite, nous utilisons une représentation *IRCL*[4] des profils de préférence.

Exemple 3.1.3 - Reprenons l'exemple 3.1.1. Les profils de préférence des agents peuvent désormais être représentés de manière plus compacte par :

$$\begin{aligned}
\succeq_{a_1} &= \{a_1,a_2\} \succ_{a_1} \{a_1,a_3,a_4\} \succ_{a_1} \{a_1,a_3\} \sim_{a_1} \{a_1,a_2,a_4\} \succ_{a_1} \{a_1,a_2,a_3,a_4\} \sim_{a_1} \{a_1\} \\
\succeq_{a_2} &= \{a_1,a_2\} \sim_{a_2} \{a_2,a_3,a_4\} \succ_{a_2} \{a_1,a_2,a_3\} \sim_{a_2} \{a_2,a_4\} \succ_{a_2} \{a_1,a_2,a_4\} \succ_{a_2} \{a_2,a_3\} \\
&\quad \succ_{a_2} \{a_1,a_2,a_3,a_4\} \sim_{a_2} \{a_2\} \\
\succeq_{a_3} &= \{a_1,a_3\} \sim_{a_3} \{a_2,a_3,a_4\} \succ_{a_3} \{a_3,a_4\} \succ_{a_3} \{a_1,a_2,a_3\} \succ_{a_3} \{a_2,a_3\} \\
&\quad \succ_{a_3} \{a_1,a_2,a_3,a_4\} \sim_{a_3} \{a_3\} \\
\succeq_{a_4} &= \{a_1,a_4\} \succ_{a_4} \{a_2,a_4\} \succ_{a_4} \{a_3,a_4\} \succ_{a_4} \{a_4\}
\end{aligned}$$

4. Listes de coalitions individuellement rationnelles (cf. section 2.1.1).

De nombreux concepts de solution, comme l'optimalité au sens de Pareto (définition 2.1.8), satisfont l'optimalité de la solution d'un point de vue collectif, mais ne garantissent pas la rationalité individuelle de la solution. De ce fait, nous ne considérons dans ce manuscrit que les trois concepts de solution canoniques qui satisfont également la rationalité individuelle : la stabilité au sens de Nash (définition 2.1.9), la stabilité individuelle (définition 2.1.10) et la stabilité du cœur (définition 2.1.12).

Ces trois concepts de solution permettent de garantir qu'étant donné la solution d'un jeu, aucun agent n'aura individuellement ou collectivement intérêt à changer de coalition. Cependant, ils ne garantissent pas l'optimalité individuelle de la solution (définition 2.1.6). Afin de mesurer le degré de satisfaction de chaque agent, nous définissons un concept d'acceptation des structures de coalitions. Intuitivement, dans une structure de coalitions Π, l'agent a_i est satisfait avec un degré x si au plus x coalitions sont préférées à la coalition $C^{\Pi}_{a_i}$.

Définition 3.1.3 - *Concept d'acceptation* : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique. Soit $a_i \in N$ un agent ayant pour profil de préférence $C_{a_i,1} \succeq_{a_i} \ldots \succeq_{a_i} C_{a_i,x} \succeq_{a_i} \ldots \succeq_{a_i} C_{a_i,2^{n-1}}$. Le *concept d'acceptation* $AP^x_{a_i}(HG)$ de profondeur $x \in [1,2^{n-1}]$ pour l'agent a_i désigne l'ensemble des structures de coalitions tel que :

$$AP^x_{a_i}(HG) = \{\Pi \in \mathcal{P}_N | C^{\Pi}_{a_i} \succeq_{a_i} C_{a_i,x}\}$$

Ainsi, le concept d'acceptation de profondeur x désigne l'ensemble des structures de coalitions de N tel que a_i est dans $C_{a_i,x}$ ou dans une coalition préférée à $C_{a_i,x}$.

Exemple 3.1.4 - Soit l'agent a_4 de l'exemple 3.1.1 ayant le profil de préférence :

$$\succ_{a_4} = \{a_1,a_4\} \succ_{a_4} \{a_2,a_4\} \succ_{a_4} \{a_3,a_4\} \succ_{a_4} \{a_4\}$$

Comme $C_{4,1} = \{a_1,a_4\}$, le concept d'acceptation de a_4 de profondeur 1 est :

$$\begin{aligned} AP^1_{a_4}(HG) &= \{\Pi_4,\Pi_{10}\} \\ &= \{\{\ \{a_1,a_4\},\{a_2\},\{a_3\}\ \},\{\ \{a_1,a_4\},\{a_2,a_3\}\ \}\} \end{aligned}$$

Remarquons que par définition, le concept d'acceptation d'un agent a_i à une profondeur $x \in [1,2^{n-1}[$ inclut toutes les structures de coalitions appartenant au concept d'acceptation de profondeur $x+1$. De même, le concept d'acceptation d'un agent à la profondeur 2^{n-1} correspond à l'ensemble des structures de coalitions possibles, soit $\mathcal{P}_N$.

Propriété 3.1.1 : Par définition des concepts d'acceptation (définition 3.1.3), pour tout jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et pour tout agent $a_i \in N$, on a :

$$AP^1_{a_i}(HG) \subseteq AP^2_{a_i}(HG) \subseteq \ldots \subseteq AP^{2^{n-1}}_{a_i}(HG) = \mathcal{P}_N$$

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et un agent $a_i \in N$ ayant le profil de préférence $C_{a_i,1} \succeq_{a_i} \ldots \succeq_{a_i} C_{a_i,x} \succeq_{a_i} \ldots \succeq_{a_i} C_{a_i,2^{n-1}}$.

Montrons dans un premier temps que $\forall x \in [1,2^{n-1}[$, $AP^x_{a_i}(HG) \subseteq AP^{x+1}_{a_i}(HG)$. Soit une structure de coalitions $\Pi \in AP^x_{a_i}(HG)$. Par définition du concept d'acceptation (définition 3.1.3), $C^{\Pi}_{a_i} \succeq_{a_i} C_{a_i,x}$. Comme $C_{a_i,x} \succeq_{a_i} C_{a_i,x+1}$, nous avons par transitivité $C^{\Pi}_{a_i} \succeq_{a_i} C_{a_i,x+1}$. Comme $\forall \Pi \in AP^x_{a_i}(HG), C^{\Pi}_{a_i} \succeq_{a_i} C_{a_i,x+1}$, nous avons $\Pi \in AP^{x+1}_{a_i}(HG)$. Ainsi, $\forall x \in [1,2^{n-1}[$, $AP^x_{a_i}(HG) \subseteq AP^{x+1}_{a_i}(HG)$.

Montrons maintenant que $AP^{2^{n-1}}_{a_i}(HG) = \mathcal{P}_N$. Par définition du profil de préférence de l'agent a_i, $\forall C \in \mathcal{C}^N_{a_i}$, on a $C \succeq_{a_i} C_{a_i,2^{n-1}}$. Ainsi, $\forall \Pi \in \mathcal{P}_N, C^{\Pi}_{a_i} \succeq_{a_i} C_{a_i,2^{n-1}}$. Or, par définition du concept d'acceptation, $AP^{2^{n-1}}_{a_i}(HG) = \{\Pi \in \mathcal{P}_N | C^{\Pi}_{a_i} \succeq_{a_i} C_{a_i,2^{n-1}}\}$. Par conséquent, $AP^{2^{n-1}}_{a_i}(HG) = \mathcal{P}_N$. □

Exemple 3.1.5 - Dans l'exemple 3.1.1, les concepts d'acceptation de l'agent a_4 sont :

$AP^1_{a_4}(HG) = \{\Pi_4, \Pi_{10}\}$
$AP^2_{a_4}(HG) = AP^1_{a_4}(HG) \cup \{\Pi_6, \Pi_9\}$
$AP^3_{a_4}(HG) = AP^2_{a_4}(HG) \cup \{\Pi_7, \Pi_8\}$
$AP^4_{a_4}(HG) = AP^3_{a_4}(HG) \cup \{\Pi_1, \Pi_2, \Pi_3, \Pi_5, \Pi_{11}\}$
$AP^5_{a_4}(HG) = AP^4_{a_4}(HG) \cup \{\Pi_{12}\}$
$AP^6_{a_4}(HG) = AP^5_{a_4}(HG) \cup \{\Pi_{13}, \Pi_{14}\}$
$AP^7_{a_4}(HG) = AP^6_{a_4}(HG) \cup \{\emptyset\}$
$AP^8_{a_4}(HG) = AP^7_{a_4}(HG) \cup \{\Pi_{15}\}$

Les concepts d'acceptation d'un agent permettent de définir un degré de satisfaction de cet agent vis-à-vis de la solution du jeu. Intuitivement, le degré de satisfaction d'un agent correspond au plus petit concept d'acceptation auquel la solution du jeu appartient.

Définition 3.1.4 - *Degré de satisfaction* : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique. Soit $\mathbb{P}(HG)$ la solution de HG. L'agent $a_i \in N$ a un *degré de satisfaction* de x si :

$$\mathbb{P}(HG) \in AP^x_{a_i}(HG) \setminus AP^{x-1}_{a_i}(HG)$$

Notons que notre définition du degré de satisfaction implique que plus le degré est petit, plus l'agent est satisfait. Ainsi, si un protocole de sélection satisfait la rationalité individuelle, il n'existe pas d'agent $a_i \in N$ dont le degré de satisfaction est supérieur à x_i tel que $C_{a_i,x_i} = \{a_i\}$.

Exemple 3.1.6 - Reprenons l'exemple 3.1.1 et considérons un protocole $\mathbb{P}$ satisfaisant la stabilité au sens de Nash. La solution du jeu appartient donc à l'ensemble $NS_{HG} = \{\Pi_8, \Pi_9\}$. Le tableau 3.2 montre alors le degré de satisfaction des agents en fonction de la solution du jeu. Le degré de satisfaction de a_1 est de 1 si la solution du jeu est la structure de coalitions Π_8 car $C^{\Pi}_{a_1} = \{a_1, a_2\}$ est la coalition qu'il préfère à toutes les autres. À l'inverse, l'agent a_4 aurait un degré de satisfaction de 3 puisqu'il préférerait former les coalitions $\{a_1, a_4\}$ et $\{a_2, a_4\}$.

	a_1	a_2	a_3	a_4
$\Pi_8 = \{\{a_1, a_2\}, \{a_3, a_4\}\}$	1	1	2	3
$\Pi_9 = \{\{a_1, a_3\}, \{a_2, a_4\}\}$	3	2	1	2

Tableau 3.2 – Degrés de satisfaction des agents en fonction de la solution du jeu HG

Notons que pour un protocole de sélection satisfaisant le concept de solution S, une structure de coalitions $\Pi \in \mathcal{P}_N$ est à la fois solution de HG et satisfaisante pour a_i avec un degré x si $\Pi \in S_{HG} \cap AP^x_{a_i}(HG)$.

3.1.3 Une approche stochastique

Lorsqu'il existe plusieurs structures de coalitions appartenant à un concept de solution, le protocole de sélection doit opérer un choix et ne retourner qu'une unique structure de coalitions. Il existe de nombreuses techniques pour cela. Par exemple, le protocole peut retourner la première structure de coalitions trouvée, ou celle qui minimise le nombre de coalitions, ou encore dépendre de facteurs externes aux préférences tels que l'identité de l'agent initiant le protocole et l'ordre d'échange des préférences. Afin d'être le plus génériques possible, nous représentons ce choix par une distribution de probabilité sur les structures de coalitions retournées par le concept de solution.

Définition 3.1.5 - *Probabilité de sélection* : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu de coalitions hédonique et $\Pi \in \mathcal{P}_N$ une structure de coalitions. La *probabilité de sélection* $\mathrm{P}(\Pi|HG)$ désigne la probabilité que la structure de coalitions Π soit retournée comme solution du jeu HG par le protocole $\mathbb{P}$.

Exemple 3.1.7 - Soit le jeu hédonique HG de l'exemple 3.1.1. Supposons que le protocole de sélection $\mathbb{P}$ définisse la solution d'un jeu en la sélectionnant aléatoirement de manière uniforme parmi l'ensemble des structures de coalitions stables au sens de Nash. Comme $NS_{HG} = \{\Pi_8, \Pi_9\} = \{ \{\{a_1,a_2\},\{a_3,a_4\}\} , \{\{a_1,a_3\},\{a_2,a_4\}\} \}$, nous avons :

$$\mathrm{P}(\Pi_8|HG) = 1/2$$

$$\mathrm{P}(\Pi_9|HG) = 1/2$$

$$\forall \Pi_i \in \mathcal{P}_N \setminus NS_{HG} : \mathrm{P}(\Pi_i|HG) = 0$$

Une connaissance des probabilités de sélection des structures de coalitions d'un jeu permet à un agent de calculer la probabilité d'être satisfait avec un degré d'au maximum x. Intuitivement, cela correspond à la probabilité que la solution du jeu HG appartienne au concept d'acceptation de profondeur x :

Définition 3.1.6 - *Probabilité de satisfaction* : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique. La *probabilité de satisfaction* $\mathrm{P}^*(a_i,x|HG)$ de l'agent a_i à une profondeur x est définie par :

$$\mathrm{P}^*(a_i,x|HG) = \sum_{\Pi \in AP^x_{a_i}(HG)} \mathrm{P}(\Pi|HG)$$

Intuitivement, la probabilité de satisfaction de l'agent désigne la probabilité que la solution du jeu appartienne au concept d'acceptation de profondeur x. Notons que la propriété d'inclusion des concepts d'acceptation (propriété 3.1.1) permet de déduire un ordre sur les probabilités de satisfaction d'un agent.

Propriété 3.1.2 : Pour tout jeu hédonique HG et pour tout agent $a_i \in N$, nous avons :

$$0 \leq \mathrm{P}^*(a_i,1|HG) \leq \mathrm{P}^*(a_i,2|HG) \leq \ldots \leq \mathrm{P}^*(a_i,2^{n-1}|HG) = 1$$

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et un agent $a_i \in N$ ayant le profil de préférence $C_{a_i,1} \succeq_{a_i} \ldots \succeq_{a_i} C_{a_i,x} \succeq_{a_i} \ldots \succeq_{a_i} C_{a_i,2^{n-1}}$. Comme $AP_{a_i}^{2^{n-1}}(HG) = \mathcal{P}_N$ (propriété 3.1.1) et comme la solution de HG appartient à $\mathcal{P}_N$, nous avons $\mathrm{P}^*(a_i,2^{n-1}|HG) = 1$. Par définition de la probabilité de satisfaction (définition 3.1.6) et comme $\forall x \in [1,2^{n-1}[: AP_{a_i}^x(HG) \subseteq AP_{a_i}^{x+1}(HG)$ (propriété 3.1.1) alors :

$$\begin{aligned}
\mathrm{P}^*(a_i,x+1|HG) &= \sum_{\Pi \in AP_{a_i}^{x+1}(HG)} \mathrm{P}(\Pi|HG) \\
&= \sum_{\Pi \in AP_{a_i}^{x}(HG)} \mathrm{P}(\Pi|HG) + \sum_{\Pi \in AP_{a_i}^{x+1}(HG) \setminus AP_{a_i}^{x}(HG)} \mathrm{P}(\Pi|HG) \\
&= \mathrm{P}^*(a_i,x|HG) + \sum_{\Pi \in AP_{a_i}^{x+1}(HG) \setminus AP_{a_i}^{x}(HG)} \mathrm{P}(\Pi|HG)
\end{aligned}$$

Ainsi, comme $\forall \Pi \in \mathcal{P}_N, \mathrm{P}(\Pi|HG) \geq 0$, $\forall x \in [1,2^{n-1}[, \mathrm{P}^*(a_i,x|HG) \leq \mathrm{P}^*(a_i,x+1|HG)$, nous obtenons l'inégalité $\mathrm{P}^*(a_i,x|HG) \leq \mathrm{P}^*(a_i,x+1|HG)$. □

Exemple 3.1.8 - Reprenons l'exemple 3.1.7 où le protocole de sélection $\mathbb{P}$ définit la solution d'un jeu en le sélectionnant aléatoirement de manière uniforme parmi l'ensemble des structures de coalitions stables au sens de Nash. Rappelons que l'ensemble des structures de coalitions stables au sens de Nash est :

$$NS_{HG} = \{\{\{a_1,a_2\},\{a_3,a_4\}\},\{\{a_1,a_3\},\{a_2,a_4\}\}\}$$

Le profil de préférence de l'agent a_4 est :

$$\succeq_{a_4} = \{a_1,a_4\} \succ_{a_4} \{a_2,a_4\} \succ_{a_4} \{a_3,a_4\} \succ_{a_4} \{a_4\}$$

Nous obtenons donc pour l'agent a_4 :

$$\mathrm{P}^*(a_4,1|HG) = 0$$

$$\mathrm{P}^*(a_4,2|HG) = 1/2$$

$$\forall i \in [3;8] : \mathrm{P}^*(a_4,i|HG) = 1$$

3.2 Manipulation d'un jeu hédonique

La satisfaction des concepts de solution permet de garantir que la solution d'un jeu respecte certaines propriétés vis-à-vis des préférences individuelles des agents. Par exemple, la stabilité au sens de Nash (définition 2.1.9) permet de garantir qu'aucun agent n'aura individuellement intérêt

à changer de coalition pour en rejoindre une autre. Cependant, la satisfaction de ces concepts de solution ne permet pas de garantir à un agent que la solution du jeu soit individuellement optimale (définition 2.1.6).

Comme nous considérons dans ce manuscrit des agents hétérogènes, ceux-ci cherchent à obtenir une solution ayant le meilleur degré de satisfaction, c'est-à-dire le plus petit possible. Ainsi, si un agent $a_i \in N$ n'est pas certain que la solution du jeu sera optimale pour lui (c'est-à-dire que $\mathrm{P}^*(a_i,1|HG) \neq 1$), il peut se demander s'il n'existe pas une *manipulation* lui permettant d'augmenter ses probabilités de satisfaction. Un tel agent est malhonnête au sens de la définition 1.2.3 puisque son objectif est uniquement de maximiser la probabilité que la solution lui soit satisfaisante.

3.2.1 Définition d'une manipulation

Dans la suite de ce chapitre, nous désignons par a_m un agent malhonnête et par M sa manipulation. L'agent malhonnête doit alors décider quelle manipulation mettre en œuvre, et cela avant que le protocole de sélection ne soit initié. Ainsi, pour un jeu HG, l'agent malhonnête doit considérer le jeu tel qu'il serait sans mettre en œuvre de manipulation et le jeu tel qu'il sera lorsque la manipulation M est mise en œuvre. Nous notons alors :

- $HG = \langle N, \succeq, \mathbb{P}\rangle$ le jeu hédonique en l'absence de manipulation ;
- $HG^M = \langle N^M, \succeq^M, \mathbb{P}\rangle$ le jeu hédonique HG subissant la manipulation M.

Pour définir une manipulation, il convient de considérer les paramètres du jeu sur lesquels un agent malhonnête peut agir. Le premier de ces paramètres est son profil de préférence. Un agent $a_i \in N$ peut manipuler un jeu hédonique en mentant sur ses préférences vis-à-vis des différentes coalitions possibles. Nous parlons alors de *faux profils de préférence*.

Définition 3.2.1 - *Faux profil de préférence* : Soit $a_i \in N$ un agent ayant le profil de préférence $C_{a_i,1} \succeq_{a_i} \ldots \succeq_{a_i} C_{a_i,x} \succeq_{a_i} \ldots \succeq_{a_i} C_{a_i,2^{n-1}}$. L'agent a_i fournit un *faux profil de préférence* s'il définit dans le jeu HG^M le profil $\succeq_{a_i}^M$ tel que :

$$\exists j,k \in [1,2^{n-1}] : C_{a_i,j} \succeq_{a_i} C_{a_i,k} \wedge C_{a_i,k} \succ_{a_i}^M C_{a_i,j}$$

Notons que notre définition de faux profils de préférence est similaire à la définition d'une manipulation selon [Gibbard, 1973] dans le cadre des systèmes de votes (définition 1.2.1). L'utilisation de faux profils de préférence a pour but d'altérer le jeu afin que son résultat augmente les probabilités de satisfaction de l'agent.

Exemple 3.2.1 - Considérons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$ où :

$$\begin{aligned} N &= \{a_1,a_2,a_3\} \\ \succeq_{a_1} &= \{a_1,a_2\} \succ_{a_1} \{a_1,a_2,a_3\} \succ_{a_1} \{a_1\} \\ \succeq_{a_2} &= \{a_1,a_2\} \succ_{a_2} \{a_1,a_2,a_3\} \succ_{a_2} \{a_2\} \\ \succeq_{a_3} &= \{a_1,a_2,a_3\} \succ_{a_3} \{a_3\} \end{aligned}$$

Dans ce jeu, si le protocole de sélection satisfait la stabilité individuelle, la solution du jeu peut être la partition $\{\{a_1,a_2,a_3\}\}$ ou la partition $\{\{a_1,a_2\},\{a_3\}\}$. L'agent a_1 peut cependant fournir un faux profil de préférence où la coalition $\{a_1,a_2,a_3\}$ est supprimée de la liste :

$$\succeq_{a_1}^M = \{a_1,a_2\} \succ_{a_1}^M \{a_1\}$$

Par cette manipulation, l'agent a_1 s'assure que la solution du jeu sera satisfaisante pour lui, car il définit le jeu hédonique HG^M où seule la partition $\{\{a_1,a_2\},\{a_3\}\}$ est individuellement stable et donc $\mathrm{P}^*(a_1,1|HG^M) = 1$.

Le second paramètre sur lequel un agent malhonnête peut agir est l'ensemble des agents du jeu en y introduisant des agents Sybil. En effet, l'introduction des fausses identités transforme le jeu HG où $N = \{a_1,\ldots,a_n\}$ en un jeu HG^M tel que $N^M = \{a_1,\ldots,a_n,s_1,\ldots,s_x\}$ où s_i désigne le i-ème agent Sybil.

Comme le problème associé à un jeu hédonique est un problème de partitionnement d'agents, augmenter le nombre d'agents présents dans le jeu augmente également le nombre de coalitions possibles et donc le nombre de structures de coalitions à considérer. Rappelons que le nombre de structures de coalitions possibles est donné par le nombre de Bell [Wieder, 2008]. Ainsi, si pour $|N| = 6$ il existe $B_6 = 203$ structures de coalitions possibles, le fait d'introduire un seul agent Sybil dans le jeu fait passer ce nombre de structures à $B_7 = 877$. De même, le fait d'augmenter le nombre d'agents présents dans le jeu a des conséquences sur les profils de préférence des agents. Selon la définition 2.1.5, le profil de préférence d'un agent a_i est un ordre total sur l'ensemble des de $\mathcal{C}_{a_i}^N$. Par conséquent, le fait d'introduire m fausses identités dans le jeu implique que le profil de préférence de chaque agent (honnête ou non) doit être défini sur 2^{n+m-1} coalitions.

Ainsi, en considérant qu'un agent malhonnête peut altérer le résultat d'un jeu par l'intermédiaire de ces deux paramètres, nous définissons une manipulation comme suit :

Définition 3.2.2 - *Manipulation d'un jeu hédonique* : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$ où $N = \{a_1,\ldots,a_{n-1},a_m\}$. Une *manipulation* sur HG mise en œuvre par l'agent a_m est un triplet $M = \langle\{a_m,s_1,\ldots,s_x\},\{\succeq_{a_m}^M, \succeq_{s_1}^M, \ldots, \succeq_{s_x}^M\}, \succeq_{a_m}\rangle$ où :
- $\{s_1,\ldots,s_x\}$ est un ensemble de x agents Sybil introduits par a_m ;
- $\{\succeq_{a_m}^M, \succeq_{s_1}^M, \ldots, \succeq_{s_x}^M\}$ est l'ensemble des profils de préférence révélés par a_m et s_1, …, s_x ;
- $\succeq_{a_m}$ désigne le véritable profil de préférence de a_m.

Cette définition générique nous permet de considérer autant les manipulations n'utilisant que les faux profils de préférence (définition 3.2.1) que celles nécessitant l'introduction d'agents Sybil. Remarquons que nous considérons qu'un agent malhonnête ne peut pas altérer le protocole de sélection utilisé pour résoudre le jeu. En effet, ce protocole est supposé être préalablement fixé et connu de tous les agents.

3.2.2 Des hypothèses de manipulation

Mettre en œuvre une manipulation M a pour objectif d'altérer le jeu afin d'augmenter les probabilités de satisfaction de a_m. Pour construire une telle manipulation, un agent malhonnête doit disposer de certaines connaissances sur le jeu. Par ailleurs, l'utilisation d'agents Sybil nécessite que l'agent malhonnête puisse déterminer comment les autres agents du jeu vont les intégrer dans leurs profils de préférence. Ainsi, l'agent malhonnête doit formuler certaines hypothèses sur le jeu et les agents honnêtes. Nous présentons ici ces hypothèses en nous plaçant du point de vue du malhonnête.

Premièrement, nous faisons l'hypothèse qu'un agent malhonnête connaît l'ensemble des paramètres d'un jeu.

Hypothèse 3.2.1 - *Connaissance initiale* : L'agent malhonnête a_m a une *connaissance initiale* de N et de $\succeq$.

Si la connaissance des agents qui vont participer au jeu est nécessaire à l'ensemble des agents pour définir leurs profils de préférence, la connaissance des profils de préférence des autres agents ne l'est pas. Par exemple, certains protocoles de sélection tels que ceux proposés par [Génin et Aknine, 2011] sont fondés sur la négociation entre agents et non sur l'échange des profils de préférence. Cependant, un agent malhonnête peut les connaître en les ayant préalablement demandés aux autres agents, en les déduisant de jeux précédents, ou encore par des connaissances a priori sur des catégories d'agents.

Deuxièmement, nous faisons l'hypothèse de l'unicité de l'agent malhonnête. Nous entendons par *unicité* le fait qu'un agent malhonnête considère les autres agents du jeu comme honnêtes et qu'ils ne vont pas eux-mêmes effectuer de manipulation.

Hypothèse 3.2.2 - *Unicité du malhonnête* : Un agent malhonnête $a_m \in N$ considère que les autres agents de N sont honnêtes.

Cette hypothèse repose sur le fait que s'il existe un autre agent malhonnête a_{m2} dans le jeu, soit a_{m2} est en collusion avec a_m, auquel cas a_{m2} peut être considéré comme un agent Sybil dans la définition de M, soit a_m considère a_{m2} comme tous les autres agents du jeu.

Si l'agent malhonnête redéfinit dans M son profil de préférence et les profils de préférence de ses agents Sybil, les préférences des agents honnêtes dans HG^M sont elles aussi affectées par l'introduction de nouveaux agents. L'agent malhonnête peut cependant émettre des hypothèses sur la définition de ces nouveaux profils de préférence. Nous considérons tout d'abord une adaptation de l'*indépendance des alternatives non-pertinentes* classiquement utilisée en théorie du choix social [Arrow, 1963]. Elle stipule que si pour deux coalitions C_1 et C_2 telles que C_1 est préférée à C_2 dans le jeu HG, C_1 reste préférée à C_2 malgré l'introduction de nouveaux agents dans le jeu.

Hypothèse 3.2.3 - *Indépendance des alternatives non-pertinentes* : Un agent $a_i \in N$ est *indépendant des alternatives non-pertinentes* si :

$$\forall C_1,C_2 \in \mathcal{C}_{a_i}^N, C_1 \succeq_{a_i} C_2 \Leftrightarrow C_1 \succeq_{a_i}^M C_2$$

Intuitivement, cette hypothèse représente le fait que l'utilisation d'agents Sybil ne change pas les préférences des agents honnêtes vis-à-vis des coalitions de N. L'indépendance des alternatives non-pertinentes repose sur le fait que les agents sont rationnels et que l'utilité associée à un choix ne dépend pas des autres choix possibles. Nous formulons une seconde hypothèse portant sur les préférences des agents honnêtes et correspondant à une indifférence a priori quant à la présence d'agents qu'ils ne connaissent pas dans une coalition.

Hypothèse 3.2.4 - *Bénéfice du doute* : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et un ensemble d'agents U tels que $U \cap N = \emptyset$. Un agent $a_i \in N$ accorde le *bénéfice du doute* aux agents de U si dans un jeu $HG' = \langle N \cup U, \succeq' \rangle$, a_i définit $\succeq'_{a_i}$ telle que :

$$\forall C \subseteq \mathcal{C}_{a_i}^N, \forall C' \subseteq U, C \sim'_{a_i} C \cup C'$$

L'hypothèse de bénéfice du doute satisfait la propriété d'ouverture du système en permettant à un nouvel agent d'être accepté dans une coalition qu'il désire rejoindre bien que les autres

membres de cette coalition ne le connaissent pas. *Si cette dernière hypothèse peut paraître particulièrement forte, nous montrons dans le chapitre 4 que la restriction à l'acceptation d'un seul et unique nouvel agent est suffisante.*

Ces quatre hypothèses permettent à l'agent a_m de déduire les préférences des autres agents lors de la mise en œuvre de M.

Exemple 3.2.2 - Considérons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ où $N = \{a_1,a_2,a_m\}$. Supposons que a_1 a le profil de préférence $\succeq_{a_1} = \{a_1,a_2\} \succ_{a_1} \{a_1,a_2,a_m\} \succ_{a_1} \{a_1\}$ et qu'un agent malhonnête a_m met en œuvre la manipulation suivante : $M = \langle \{a_m,s_1,s_2\}, \{\succeq^M_{a_m}, \succeq^M_{s_1}, \succeq^M_{s_2}\}, \succeq_{a_m} \rangle$. Par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute, nous avons :

$$\begin{aligned}\succeq^M_{a_1} = \{a_1,a_2\} &\sim^M_{a_1} \{a_1,a_2,s_1\} \sim^M_{a_1} \{a_1,a_2,s_2\} \sim^M_{a_1} \{a_1,a_2,s_1,s_2\} \succ^M_{a_1} \{a_1,a_2,a_3\} \\ &\sim^M_{a_1} \{a_1,a_2,a_3,s_1\} \sim^M_{a_1} \{a_1,a_2,a_3,s_2\} \sim^M_{a_1} \{a_1,a_2,a_3,s_1,s_2\} \succ^M_{a_1} \{a_1\} \\ &\sim^M_{a_1} \{a_1,s_1\} \sim^M_{a_1} \{a_1,s_2\} \sim^M_{a_1} \{a_1,s_1,s_2\}\end{aligned}$$

À partir de ceci, l'agent malhonnête peut calculer le résultat d'un jeu HG^M et de décider de mettre en œuvre M si cette dernière lui permet d'améliorer sa probabilité de satisfaction. Notons que ces hypothèses sont optimistes du point de vue de l'agent malhonnête. *Cette approche nous permettra de montrer dans le chapitre 4 que si, dans des conditions favorables, il est difficile pour un agent malhonnête de manipuler le jeu alors cela est encore plus difficile dans un contexte défavorable.*

3.2.3 Rationalité d'une manipulation

Pour définir si un agent a un intérêt à effectuer une manipulation, nous nous intéressons à la rationalité de sa mise en œuvre. Une manipulation M est dite *rationnelle* sur un jeu HG si elle permet à l'agent malhonnête d'augmenter ces probabilités de satisfaction. Nous proposons de quantifier cette rationalité en nous fondant sur le degré de satisfaction des agents (définition 3.1.4) et de leur probabilité de satisfaction (définition 3.1.6).

Rappelons que la satisfaction d'un agent dépend du concept d'acceptation (définition 3.1.3) auquel la solution du jeu appartient. Dans un jeu HG^M, l'agent malhonnête a_m et ses agents Sybil $s_1, \dots s_x$ sont en réalité le même agent. Par ailleurs, les profils de préférence fournis par a_m et ses agents Sybil ne correspondent pas nécessairement aux véritables préférences de a_m. Par conséquent, nous devons redéfinir le concept d'acceptation de l'agent malhonnête dans le jeu qu'il manipule. Nous nommons ce nouveau concept, *concept d'acceptation malhonnête.*

Définition 3.2.3 - ***Concept d'acceptation malhonnête*** **:** Soit un jeu $HG = \langle N, \succeq, \mathbb{P} \rangle$ et $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \dots \succeq_{a_m} C_{a_m,x} \succeq_{a_m} \dots \succeq_{a_m} C_{a_m,2^{n-1}}$. Soit $HG^M = \langle N^M, \succeq^M, \mathbb{P} \rangle$ le jeu résultant de la manipulation M sur HG. Le *concept d'acceptation malhonnête* de a_m sur HG^M à une profondeur $x \in [1,2^{n-1}]$ (noté $AP^x_{a_m}(HG^M)$) désigne l'ensemble des structures de coalitions $\Pi \in \mathcal{P}_{N^M}$ tel qu'il existe un agent $a_i \in \{a_m,s_1,\dots,s_x\}$ dont la coalition $C^\Pi_{a_i}$ vérifie :

$$(\, C^\Pi_{a_i} \setminus \{a_i\}\,) \cup \{a_m\} \succeq_{a_m} C_{a_m,x}$$

Intuitivement, une partition $\Pi \in \mathcal{P}_{N^M}$ appartient au concept d'acceptation malhonnête de profondeur x si au moins l'une des deux conditions suivantes est satisfaite :

1. l'agent malhonnête est dans une coalition préférée à $C_{a_m,x}$: $C^{\Pi}_{a_m} \succeq_{a_m} C_{a_m,x}$;
2. il existe un agent $s_i \in \{s_1,\ldots,s_x\}$ est dans une coalition préférée à $C_{a_m,x}$.

Comme le concept d'acceptation malhonnête est défini à partir des véritables préférences de a_m ($\succeq_{a_m}$), a_m n'a pas d'intérêt à être en coalition avec lui-même (sous la forme de l'un des agents Sybil). De ce fait, les structures de coalitions où l'agent malhonnête et au moins un de ses agents Sybils sont membres d'une même coalition n'appartiennent pas au concept d'acceptation malhonnête. Ces partitions sont dites *non acceptables*.

Définition 3.2.4 - *Structure de coalitions non acceptable* : Soit un jeu $HG^M = \langle N^M, \succeq^M, \mathbb{P}\rangle$ le jeu hédonique résultant de la manipulation M mise en œuvre par un agent a_m. La structure de coalitions $\Pi \in \mathcal{P}_{N^M}$ est *non acceptable* si :

$$\forall a_i \in \{a_m,s_1,\ldots,s_x\}, \exists a_j \in \{a_m,s_1,\ldots,s_x\} \setminus \{a_i\} : C^{\Pi}_{a_i} = C^{\Pi}_{a_j}$$

Exemple 3.2.3 - Soit le jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$ avec $N = \{a_1,a_2,a_m\}$ où l'agent a_m est un agent malhonnête ayant le profil de préférence $\{a_1,a_m\} \succ_{a_m} \{a_1,a_2,a_m\} \succ_{a_m} \{a_2,a_m\} \succ_{a_m} \{a_m\}$. Soit HG^M le jeu résultant de la manipulation $M = \langle \{a_m,s\}, \{\succeq^M_{a_m}, \succeq^M_{s}\}, \succeq_{a_m}\rangle$ sur HG. La figure 3.2 représente les différents concepts d'acceptation de a_m sur le jeu HG^M. La figure 3.3 représente, quant à elle, les 4 structures de coalitions non acceptables pour a_m car il est en coalition avec s.

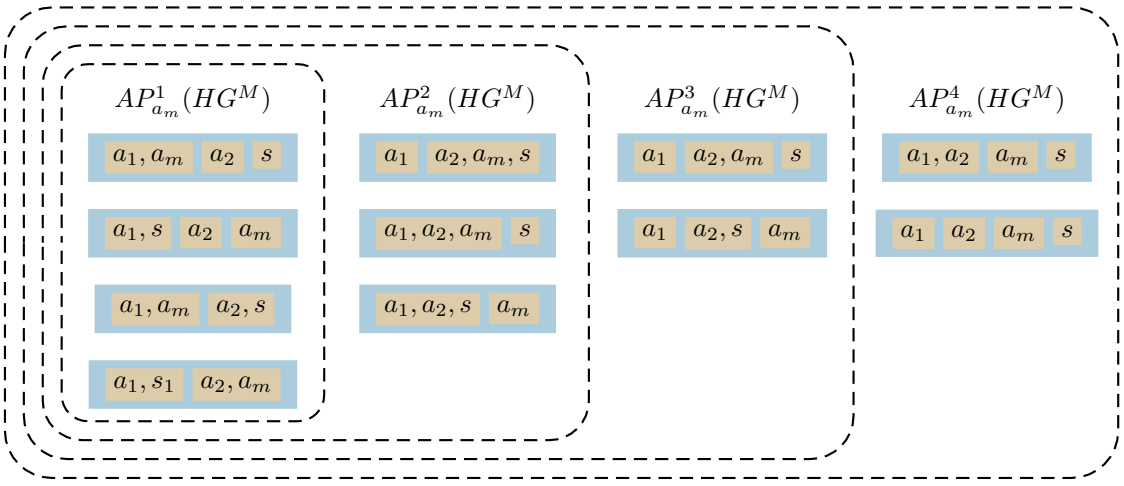

FIGURE 3.2 – Concepts d'acceptation de l'agent a_m dans le jeu HG^M

Le concept d'acceptation malhonnête permet de calculer la probabilité de satisfaction (définition 3.1.6) de l'agent malhonnête. Comme son objectif est d'augmenter cette probabilité de satisfaction, nous définissons la rationalité d'une manipulation [5] :

5. Cette définition de la rationalité d'une manipulation est une généralisation de celle que nous proposée dans [Vallée *et al.*, 2014c, Vallée *et al.*, 2013] et qui était fondée l'existence d'une coalition seuil C_θ.

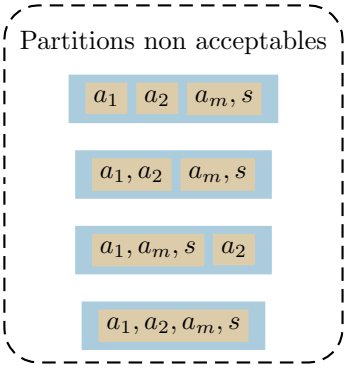

FIGURE 3.3 – Partitions de HG^M non acceptables pour a_m

Définition 3.2.5 - *Manipulation k-rationnelle* : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et un agent malhonnête $a_m \in N$. Soit HG^M le jeu hédonique résultant d'une manipulation M mise en œuvre par a_m sur HG. La manipulation M est *k-rationnelle* si :

$$\forall i : 1 \leq i < k, \mathrm{P}^*(a_m, i|HG) = \mathrm{P}^*(a_m, i|HG^M)$$
$$\mathrm{P}^*(a_m, k|HG) < \mathrm{P}^*(a_m, k|HG^M)$$

La manipulation M est *rationnelle* sur HG s'il existe un $k \in [1, 2^{n-1}]$ tel que M soit k-rationnelle.

Intuitivement, une manipulation est k-rationnelle si elle permet d'améliorer la probabilité de l'agent malhonnête (sous sa véritable identité ou celle d'un de ses Sybil) d'être dans la coalition $C_{a_m,k}$ sans diminuer pour autant la probabilité d'être dans toutes les coalitions $C_{a_m,i}$ préférées à $C_{a_m,k}$. Notons que, par définition et par l'hypothèse de rationalité individuelle minimale (hypothèse 3.1.1), si $\{a_m\} \succeq_{a_m} C_{a_m,k}$ alors la manipulation ne peut pas être k-rationelle. Dans toute la suite, nous considérons trivialement que $C_{a_m,k} \succeq_{a_m} \{a_m\}$: aucun agent ne désire être dans une coalition qu'il ne désire pas.

Exemple 3.2.4 - Soit le jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ suivant :

$$N = \{a_1, a_2, a_m\}$$
$$\succeq_{a_1} = \{a_1, a_2, a_m\} \succ_{a_1} \{a_1\}$$
$$\succeq_{a_2} = \{a_2, a_m\} \succ_{a_2} \{a_1, a_2, a_m\} \succ_{a_2} \{a_2\}$$
$$\succeq_{a_m} = \{a_2, a_m\} \succ_{a_m} \{a_1, a_2, a_m\} \succ_{a_m} \{a_m\}$$

Supposons que le protocole de sélection satisfait la stabilité individuelle. Ainsi, nous avons :

$$\mathrm{P}(\{\{a_1, a_2, a_m\}\}|HG) = 1/2$$
$$\mathrm{P}(\{\{a_1\}, \{a_2, a_m\}\}|HG) = 1/2$$

Par conséquent, la probabilité de satisfaction de a_m est :

$$\mathrm{P}^*(a_m, 1|HG) = 1/2$$
$$\mathrm{P}^*(a_m, 2|HG) = 1$$

Considérons la manipulation $M_1 = \langle\{a_m\},\{\succeq_{a_m}^{M_1}\}, \succeq_{a_m}\rangle$ où :

$$\succeq_{a_m}^{M_1} = \{a_2,a_m\} \succ_{a_m}^{M_1} \{a_m\} \succ_{a_m}^{M_1} \{a_1,a_m\} \succ_{a_m}^{M_1} \{a_1,a_2,a_m\}$$

Dans HG^{M_1}, seule la partition $\{ \{a_1\},\{a_2,a_m\} \}$ est individuellement stable. Comme le montre l'inégalité ci-dessous, la manipulation M_1 est 1-rationnelle.

$$\mathrm{P}^*(a_m,1|HG) < \mathrm{P}^*(a_m,1|HG^M) = 1$$

Considérons maintenant la manipulation $M_2 = \langle\{a_m,s\},\{\succeq_{a_m}^{M_2}, \succeq_s^{M_2}\}, \succeq_{a_m}\rangle$ avec :

$$\succeq_{a_m}^{M_2} = \{a_2,a_m\} \succ_{a_m}^{M_2} \{a_1,a_2,a_m\} \succ_{a_m}^{M_2} \{a_m\}$$
$$\succeq_s^{M_2} = \{a_2,s\} \succ_s^{M_2} \{a_2,a_3,s\} \succ_s^{M_2} \{s\}$$

Si les hypothèses de bénéfice du doute et d'indépendance des alternatives non-pertinentes (hypothèses 3.2.4 et 3.2.3) sont satisfaites, HG^{M_2} a 4 structures de coalitions individuellement stables (figure 3.4). Les probabilités de sélection des solutions sont :

$$\mathrm{P}(\{\{a_1\},\{a_2,a_m\},\{s\}\}|HG^{M_2}) = 1/4$$
$$\mathrm{P}(\{\{a_1,s\},\{a_2,a_m\}\}|HG^{M_2}) = 1/4$$
$$\mathrm{P}(\{\{a_1,a_2,a_m\},\{s\}\}|HG^{M_2}) = 1/4$$
$$\mathrm{P}(\{\{a_1,a_2,s\},\{a_m\}\}|HG^{M_2}) = 1/4$$

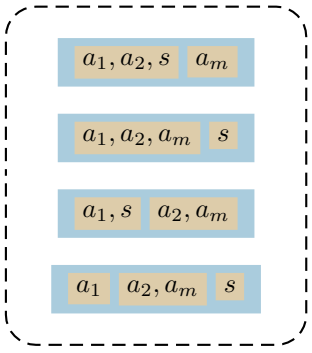

FIGURE 3.4 – Ensemble des structures de coalitions individuellement stables du jeu HG^{M_2}

Par conséquent, la probabilité de satisfaction de a_m dans HG^{M_2} est :

$$\mathrm{P}^*(a_m,1|HG^{M_2}) = 1/2$$
$$\mathrm{P}^*(a_m,2|HG^{M_2}) = 1$$

Ainsi, M_2 n'améliore pas les probabilités de satisfaction de a_m et n'est donc pas rationnelle.

3.3 Conclusion

Dans ce chapitre, nous nous sommes intéressés à la définition d'un modèle générique de manipulation sur les jeux hédoniques. Nous considérons un modèle canonique de jeu hédonique où les agents définissent un *profil de préférence* sur l'ensemble des coalitions auquel ils peuvent appartenir. Nous associons à ce modèle un *protocole de sélection* abstrait qui retourne la solution du jeu.

Si ce protocole garantit que la solution d'un jeu satisfait certaines propriétés comme l'appartenance à un *concept de solution* donné (stabilité au sens de Nash ou stabilité au sens du cœur), la solution n'est généralement pas optimale pour l'ensemble des agents. C'est pourquoi nous mesurons la qualité d'une solution du point de vue d'un agent en définissant un concept de solution spécifique à l'agent, appelé *concept d'acceptation*. Cette mesure est appelée son *degré de satisfaction*. Pour un degré fixé et sachant certaines données du jeu (profils de préférence des autres agents et protocole de sélection), un agent peut calculer la probabilité que la solution du jeu soit satisfaisante. Nous appelons cette estimation la *probabilité de satisfaction*.

Nous avons alors proposé un modèle de *manipulation*, c'est-à-dire une stratégie permettant à un agent d'altérer le jeu avant qu'il ne soit résolu afin d'augmenter ses probabilités de satisfaction. La définition d'une manipulation que nous proposons repose à la fois sur l'utilisation de *faux profils de préférence* et l'introduction dans le jeu de faux agents appelés *agents Sybil*. À partir d'hypothèses sur les autres agents, nous définissons une mesure de *rationalité d'une manipulation* correspondant au fait qu'appliquer la manipulation lui permet d'augmenter effectivement son degré de satisfaction. Cette définition de la rationalité d'une manipulation est l'élément central nous permettant d'étudier les conditions particulières qui rendent inefficaces des manipulations pourtant simples en pratique. Dans le chapitre 4, nous étudions les caractéristiques nécessaires aux manipulations pour être rationnelles lorsque le concept de solution satisfait la stabilité au sens de Nash. Dans le chapitre 5, nous poursuivons cette étude en relâchant un certain nombre d'hypothèses et en considérant d'autres concepts de solution.

Chapitre 4

Stabilité au sens de Nash : un concept de solution robuste

Sommaire

Résumé.
La *stabilité au sens de Nash* est un concept de solution classique de la théorie des jeux permettant de garantir qu'individuellement aucun agent n'a d'intérêt à changer de coalition. À partir du modèle de jeu hédonique et de manipulation proposé dans le chapitre 3, nous montrons qu'utiliser la stabilité au sens de Nash permet de lutter efficacement contre les *manipulations*. Pour cela, nous considérons deux manipulations spécifiques : la *manipulation constructive* et la *manipulation destructive* utilisant toutes deux un unique *agent Sybil*. D'une part, nous caractérisons les conditions nécessaires à la *rationalité* de ces deux manipulations. D'autre part, nous montrons que considérer ces deux manipulations est suffisant pour estimer la robustesse d'un jeu hédonique, car toute manipulation doit au moins en satisfaire les conditions. Nous démontrons ensuite que calculer une manipulation rationnelle est un *problème difficile* pour les agents malhonnêtes. Enfin, nous présentons une étude empirique illustrant le fait que la stabilité au sens de Nash permet, dans la majorité des cas, de garantir la robustesse d'un jeu hédonique face aux manipulations.

Nous considérons dans un ce chapitre unique concept de solution : la stabilité au sens de Nash (définition 2.1.9). Dans la suite de cette section, nous écrivons par mesure de simplicité *stabilité* en lieu et place de stabilité au sens de Nash. Rappelons qu'une structure de coalitions est stable si aucun agent n'a individuellement intérêt à changer de coalition. De ce fait, nous considérons ici que le protocole de sélection $\mathbb{P}$ (définition 3.1.1) retourne la solution du jeu en la sélectionnant aléatoirement uniformément parmi l'ensemble des structures de coalitions stables.

Hypothèse 4.0.1 - *Satisfaction de la Nash stabilité* : Le protocole de sélection $\mathbb{P}$ utilisé par les agents du jeu HG définit la solution en la sélectionnant aléatoirement uniformément parmi l'ensemble des structures de coalitions stables au sens de Nash du jeu HG, et satisfait l'hypothèse de rationalité individuelle minimale en l'absence de structures de coalitions stables au sens de Nash.

Afin de simplifier la lecture de ce chapitre, le tableau 4.1 résume les principales notations utilisées.

Manipulations	
HG^M	Jeu résultant de l'application de la manipulation M sur HG
M_C et M_D	Manipulations constructive et destructive
$NS_{HG} \subseteq \mathcal{P}_N$	Structures de coalitions stables au sens de Nash
$UR^{HG}_{a_m} \subseteq \mathcal{P}_N$	Structures de coalitions dont a_m est l'unique responsable de la non-stabilité
$f(\Pi,s,C_0) \in \mathcal{P}_{N\cup\{s\}}$	Structure de coalitions construite par l'ajout de l'agent s dans la coalition $C_0 \in \Pi \cup \{\emptyset\}$
$f^{-1}(\Pi') \in \mathcal{P}_N$	Structure de coalitions servant à construire $\Pi' \in \mathcal{P}_{N\cup\{s\}}$
$card_M(\Pi \vert HG^M)$	Nombre de structures de coalitions stables dans HG^M construites à partir de $\Pi \in \mathcal{P}_N$

Tableau 4.1 – Principales notations des manipulations sur les jeux hédoniques

4.1 Manipulation constructive

4.1.1 Une manipulation qui augmente le nombre de solutions

La première manipulation que nous proposons d'étudier consiste à augmenter artificiellement le nombre de structures de coalitions étant à la fois satisfaisantes pour l'agent a_m et stables. C'est pour cela que nous appelons cette manipulation, une *manipulation constructive*. Pour ce faire, l'agent a_m fournit un faux profil de préférence (définition 3.2.1) où il se dit indifférent vis-à-vis de toutes les coalitions de $\mathcal{C}^N_{a_m}$ et introduit un agent Sybil présentant ses véritables préférences.

Définition 4.1.1 - *Manipulation constructive* : Soit $HG = \langle N, \succeq, \mathbb{P}\rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête ayant le profil de préférence $\succeq_{a_m} = C_{a_m,1} \succeq_{a_m} C_{a_m,2} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. La *manipulation constructive* M_C mise en œuvre par a_m sur HG est $M_C = \langle\{a_m,s\},\{\succeq^{M_C}_{a_m}, \succeq^{M_C}_{s}\}, \succeq_{a_m}\rangle$ où :

$$\succeq^{M_C}_{a_m} = C_{a_m,1} \sim^{M_C}_{a_m} C_{a_m,2} \sim^{M_C}_{a_m} \ldots \sim^{M_C}_{a_m} C_{a_m,2^{n-1}}$$
$$\succeq^{M_C}_{s} = C_{a_m,1} \cup \{s\} \setminus \{a_m\} \succeq^{M_C}_{s} C_{a_m,2} \cup \{s\} \setminus \{a_m\} \succeq^{M_C}_{s} \ldots \succeq^{M_C}_{s} C_{a_m,2^{n-1}} \cup \{s\} \setminus \{a_m\}$$

Remarquons que dans cette définition, l'agent a_m et l'agent Sybil s considèrent leur coalition singleton respective comme préférée à toutes les coalitions auxquels ils peuvent appartenir conjointement. Cela permet de garantir que la solution du jeu après manipulation soit acceptable (définition 3.2.4).

Exemple 4.1.1 - Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ avec :

$$
\begin{aligned}
N &= \{a_1,a_2,a_3,a_m\} \\
\succeq_{a_1} &= \{a_1,a_2\} \succ_{a_1} \{a_1,a_3,a_m\} \succ_{a_1} \{a_1,a_3\} \sim_{a_1} \{a_1,a_2,a_m\} \succ_{a_1} \{a_1,a_2,a_3,a_m\} \sim_{a_1} \{a_1\} \\
\succeq_{a_2} &= \{a_1,a_2\} \sim_{a_2} \{a_2,a_3,a_m\} \succ_{a_2} \{a_1,a_2,a_3\} \sim_{a_2} \{a_2,a_m\} \succ_{a_2} \{a_1,a_2,a_m\} \succ_{a_2} \{a_2,a_3\} \\
&\quad \succ_{a_2} \{a_1,a_2,a_3,a_m\} \sim_{a_2} \{a_2\} \\
\succeq_{a_3} &= \{a_1,a_3\} \sim_{a_3} \{a_2,a_3,a_m\} \succ_{a_3} \{a_3,a_m\} \succ_{a_3} \{a_1,a_2,a_3\} \succ_{a_3} \{a_2,a_3\} \\
&\quad \succ_{a_3} \{a_1,a_2,a_3,a_m\} \sim_{a_3} \{a_3\} \\
\succeq_{a_m} &= \{a_1,a_m\} \succ_{a_m} \{a_2,a_m\} \succ_{a_m} \{a_3,a_m\} \succ_{a_m} \{a_m\}
\end{aligned}
$$

La manipulation constructive mise en œuvre par a_m est :

$$
\begin{aligned}
\succeq_{a_m}^{M_C} &= \{a_1,a_m\} \sim_{a_m}^{M_C} \{a_2,a_m\} \sim_{a_m}^{M_C} \{a_3,a_m\} \sim_{a_m}^{M_C} \{a_1,a_2,a_m\} \sim_{a_m}^{M_C} \{a_1,a_3,a_m\} \\
&\quad \sim_{a_m}^{M_C} \{a_2,a_3,a_m\} \sim_{a_m}^{M_C} \{a_1,a_2,a_3,a_m\} \sim_{a_m}^{M_C} \{a_m\} \\
\succeq_{s}^{M_C} &= \{a_1,s\} \succ_{s}^{M_C} \{a_2,a_m\} \succ_{s}^{M_C} \{a_3,s\} \succ_{s}^{M_C} \{s\}
\end{aligned}
$$

L'application de M_C sur HG donne donc le jeu $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ où :

$$N^{M_C} = \{a_1,a_2,a_3,a_m,s\}$$

$$
\begin{aligned}
\succeq_{1}^{M_C} &= \{a_1,a_2\} \sim_{a_1}^{M_C} \{a_1,a_2,s\} \succ_{a_1}^{M_C} \{a_1,a_3,a_m\} \sim_{a_1}^{M_C} \{a_1,a_3,a_m,s\} \\
&\quad \succ_{a_1}^{M_C} \{a_1,a_3\} \sim_{a_1}^{M_C} \{a_1,a_3,s\} \sim_{a_1}^{M_C} \{a_1,a_2,a_m\} \sim_{a_1}^{M_C} \{a_1,a_2,a_m,s\} \\
&\quad \succ_{a_1}^{M_C} \{a_1,a_2,a_3,a_m\} \sim_{a_1}^{M_C} \{a_1,a_2,a_3,a_m,s\} \sim_{a_1}^{M_C} \{a_1\} \sim_{a_1}^{M_C} \{a_1,s\} \\
\succeq_{2}^{M_C} &= \{a_1,a_2\} \sim_{a_2}^{M_C} \{a_1,a_2,s\} \sim_{a_2}^{M_C} \{a_2,a_3,a_m\} \sim_{a_2}^{M_C} \{a_2,a_3,a_m,s\} \\
&\quad \succ_{a_2}^{M_C} \{a_1,a_2,a_3\} \sim_{a_2}^{M_C} \{a_1,a_2,a_3,s\} \sim_{a_2}^{M_C} \{a_2,a_m\} \sim_{a_2}^{M_C} \{a_2,a_m,s\} \\
&\quad \succ_{a_2}^{M_C} \{a_1,a_2,a_m\} \sim_{a_2}^{M_C} \{a_1,a_2,a_m,s\} \succ_{a_2}^{M_C} \{a_2,a_3\} \sim_{a_2}^{M_C} \{a_2,a_3,s\} \\
&\quad \succ_{a_2}^{M_C} \{a_1,a_2,a_3,a_m\} \sim_{a_2}^{M_C} \{a_1,a_2,a_3,a_m,s\} \sim_{a_2}^{M_C} \{a_2\} \sim_{a_2}^{M_C} \{a_2,s\} \\
\succeq_{3}^{M_C} &= \{a_1,a_3\} \sim_{a_3}^{M_C} \{a_1,a_3,s\} \sim_{a_3}^{M_C} \{a_2,a_3,a_m\} \sim_{a_3}^{M_C} \{a_2,a_3,a_m,s\} \\
&\quad \succ_{a_3}^{M_C} \{a_3,a_m\} \sim_{a_3}^{M_C} \{a_3,a_m,s\} \succ_{a_3}^{M_C} \{a_1,a_2,a_3\} \sim_{a_3}^{M_C} \{a_1,a_2,a_3,s\} \\
&\quad \succ_{a_3}^{M_C} \{a_2,a_3\} \sim_{a_3}^{M_C} \{a_2,a_3,s\} \succ_{a_3}^{M_C} \{a_1,a_2,a_3,a_m\} \sim_{a_3}^{M_C} \{a_1,a_2,a_3,a_m,s\} \\
&\quad \sim_{a_3}^{M_C} \{a_3\} \sim_{a_3}^{M_C} \{a_3,s\}
\end{aligned}
$$

$$
\begin{aligned}
\succeq_{a_m}^{M_C} &= \{a_1,a_m\} \sim_{a_m}^{M_C} \{a_2,a_m\} \sim_{a_m}^{M_C} \{a_3,a_m\} \sim_{a_m}^{M_C} \{a_1,a_2,a_m\} \sim_{a_m}^{M_C} \{a_1,a_3,a_m\} \\
&\quad \sim_{a_m}^{M_C} \{a_2,a_3,a_m\} \sim_{a_m}^{M_C} \{a_1,a_2,a_3,a_m\} \sim_{a_m}^{M_C} \{a_m\} \\
\succeq_{s}^{M_C} &= \{a_1,s\} \succ_{s}^{M_C} \{a_2,a_m\} \succ_{s}^{M_C} \{a_3,s\} \succ_{s}^{M_C} \{s\}
\end{aligned}
$$

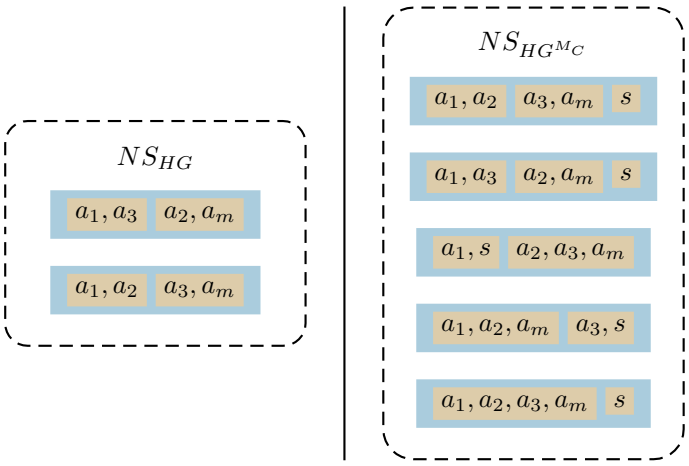

FIGURE 4.1 – Structures de coalitions stables au sens de Nash de HG et de HG^{M_C}

La figure 4.1 présente les structures de coalitions stables pour les jeux HG et HG^{M_C}.

Comme nous le montre la figure 4.1, la manipulation constructive permet d'augmenter le nombre de structures de coalitions stables. Cette augmentation est due au fait que la manipulation constructive rend stable les structures de coalitions où l'agent a_m est l'unique agent de N préférant rejoindre une autre coalition. L'agent malhonnête est alors appelé l'*unique responsable de la non-stabilité* de ces structures de coalitions.

Définition 4.1.2 - *Agent responsable de la non-stabilité* : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $\Pi \notin NS_{HG}$ une structure de coalitions. L'agent $a_i \in N$ est l'*unique responsable de la non-stabilité* de Π si :

1. $\exists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succeq_{a_i} C^{\Pi}_{a_i}$;
2. $\forall a_j \in N \setminus \{a_i\}, \forall C \in \Pi \cup \{\emptyset\}, C^{\Pi}_{a_j} \succeq_{a_j} C \cup \{a_j\}$.

Nous notons dans toute la suite $UR^{HG}_{a_i}$ l'ensemble des structures de coalitions dont l'agent a_i est l'unique responsable de la non-stabilité.

Exemple 4.1.2 - Reprenons l'exemple 4.1.1. La figure 4.2 présente les trois structures de coalitions de $\mathcal{P}_N$ dont l'agent a_m est l'unique responsable de la non-stabilité.

Dans la suite de cette section, nous étudions la stabilité des structures de coalitions résultantes de l'ajout de l'agent s dans l'une des coalitions de Π ou de la coalition singleton $\{s\}$. Nous montrons alors que l'ensemble des structures de coalitions stables de HG^{M_C} peut être calculé à partir de HG. Pour simplifier la lecture, nous notons $f(\Pi, s, C_0) = \Pi' \in \mathcal{P}_{N^{M_C}}$ la fonction qui ajoute l'agent s à la coalition $C_0 \in \Pi \cup \{\emptyset\}$ et $f^{-1}(\Pi')$ la fonction qui supprime l'agent s de Π'. Remarquons que pour toute structure de coalitions $\Pi' \in \mathcal{P}_{N^{M_C}}$, il existe une est une seule structure $\Pi \in \mathcal{P}_N$ telle que $f^{-1}(\Pi') = \Pi$.

Exemple 4.1.3 - Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ où $N = \{a_1, a_2, a_3, a_m\}$. Considérons la structure de coalitions $\Pi = \{\{a_1, a_2\}, \{a_3\}, \{a_m\}\}$. L'ajout de l'agent s à Π produit 4 structures de coalitions possibles dans $\mathcal{P}_N$:

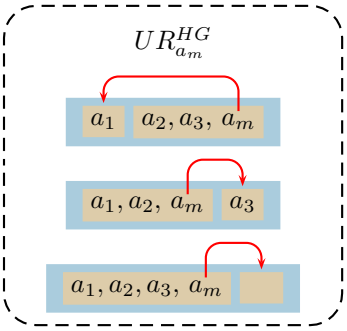

FIGURE 4.2 – Partitions de $\mathcal{P}_N$ dont a_m est l'unique responsable de la non-stabilité

- $f(\Pi,s,\emptyset) = \{ \{a_1,a_2\},\{a_3\},\{a_m\},\{s\} \}$;
- $f(\Pi,s,\{a_1,a_2\}) = \{ \{a_1,a_2,s\},\{a_3\},\{a_m\} \}$;
- $f(\Pi,s,\{a_3\}) = \{ \{a_1,a_2\},\{a_3,s\},\{a_m\} \}$;
- $f(\Pi,s,\{a_m\}) = \{ \{a_1,a_2\},\{a_3\},\{a_m,s\} \}$.

Afin d'étudier la rationalité de M_C, nous devons définir l'ensemble des structures de coalitions stables de HG^{M_C}. Dans la suite, pour toute partition $\Pi \in \mathcal{P}_N$, nous désignons par $card_{M_C}(\Pi|HG)$ le nombre de structures de coalitions $\Pi' \in NS_{HG^{M_C}}$ construites à partir de Π :

$$card_{M_C}(\Pi|HG) = |\{ \Pi' \in NS_{HG^{M_C}} | f^{-1}(\Pi') = \Pi \}|$$

Par extension, pour tout ensemble de structures de coalitions $\mathcal{P} \subseteq \mathcal{P}_N$, nous désignons par $card_{M_C}(\mathcal{P}|HG)$ le nombre de structures de coalitions $\Pi' \in NS_{HG^{M_C}}$ construites à partir des structures de coalitions de l'ensemble $\mathcal{P}$:

$$card_{M_C}(\mathcal{P}|HG) = \sum_{\Pi \in \mathcal{P}} card_{M_C}(\Pi|HG)$$

Regardons dans un premier temps l'effet de la manipulation constructive sur les structures de coalitions stables pour le jeu HG.

Propriété 4.1.1 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Soit $\Pi \in NS_{HG}$ une structure de coalitions stable au sens de Nash de HG. La structure de coalitions $\Pi' = f(\Pi,s,C_0)$ de HG^{M_C} est stable au sens de Nash si et seulement si les deux conditions suivantes sont satisfaites :

1. $C_0 \neq C^{\Pi}_{a_m}$;
2. $\forall C \in (\Pi \setminus \{C^{\Pi}_{a_m}\}) \cup \{\emptyset\} : C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}$.

Démonstration : Fixons une structure de coalitions $\Pi \in NS_{HG}$ et une coalition $C_0 \in (\Pi \setminus \{C^{\Pi}_{a_m}\}) \cup \{\emptyset\}$. Montrons dans un premier temps que la condition (1) est nécessaire pour que la structure de coalitions $\Pi' = f(\Pi,s,C_0)$ de HG^{M_C} soit stable au sens de Nash. Par définition du profil de préférence de a_m dans la manipulation constructive (définition 4.1.1), nous avons $\{a_m\} \succ^{M_C}_{a_m} C^{\Pi}_{a_m} \cup \{s\}$. Par construction de Π', si $C_0 = C^{\Pi}_{a_m}$, nous avons $C^{\Pi'}_{a_m} = C^{\Pi}_{a_m} \cup \{s\}$. Par conséquent, par définition de la stabilité (définition 2.1.9), si $C_0 = C^{\Pi}_{a_m}$ alors la structure de

coalitions Π' n'est pas stable pour le jeu HG^{M_C}.
Supposons maintenant que la condition (1) est satisfaite et montrons que la condition (2) est également nécessaire à la stabilité de la structure de coalitions Π'. Par définition des préférences de a_m dans la manipulation constructive, nous avons :

$$\forall C \in (\Pi' \setminus \{C_0\}) \cup \{\emptyset\}, C^{\Pi}_{a_m} \sim^{M_C}_{a_m} C \cup \{a_m\}$$

Par conséquent, pour la structure de coalitions Π', l'agent a_m ne préfère pas changer de coalition. Comme $\Pi \in NS_{HG}$, nous avons $\forall a_i \in N \setminus \{a_m\}, \forall C \in \Pi \cup \{\emptyset\}, C^{\Pi}_{a_i} \succeq_{a_i} C \cup \{a_i\}$. Par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute (hypothèses 3.2.3 et 3.2.4), $\forall a_i \in N \setminus \{a_m\}, \forall C \in \Pi' \cup \{\emptyset\}$, nous avons :

$$C^{\Pi}_{a_i} \sim^{M_C}_{a_i} C^{\Pi}_{a_i} \cup \{s\} \succeq^{M_C}_{a_i} C \cup \{a_i\} \sim^{M_C}_{a_i} C \cup \{a_i, s\}$$

Par construction de Π', $\forall a_i \in N \setminus \{a_m\}$, nous avons soit $C^{\Pi'}_{a_i} = C^{\Pi}_{a_i}$, soit $C^{\Pi'}_{a_i} = C^{\Pi}_{a_i} \cup \{s\}$. Dans les deux cas, les agents honnêtes ne préfèrent pas changer de coalition pour la structure de coalitions Π'. Supposons que $\exists C \in (\Pi \setminus \{C^{\Pi}_{a_m}\}) \cup \{\emptyset\}$ telle que $C \cup \{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$. Par définition de $\succ^{M_C}_s$, nous avons alors $C \cup \{s\} \succ^{M_C}_s C_0 \cup \{s\}$. Par conséquent, la structure de coalitions Π' n'est pas stable sous cette hypothèse. Supposons maintenant que $\forall C \in (\Pi \setminus \{C^{\Pi}_{a_m}\}) \cup \{\emptyset\}$, $C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}$. Par définition des préférences de s dans la manipulation constructive, nous avons :

$$\forall C \in (\Pi' \setminus \{C^{\Pi}_{a_m}\}) \cup \{\emptyset\}, C^{\Pi'}_{s} \succeq^{M_C}_{s} C \cup \{s\}$$

En conséquence, si la condition (2) est satisfaite, la structure de coalitions Π' est stable. □

Ainsi, selon la propriété 4.1.1, l'agent Sybil peut rejoindre toute coalition qu'il ne préfère pas à une autre au sein d'une structure de coalitions Π stable de HG et garantir la stabilité de cette nouvelle structure. Remarquons que, comme la partition Π est stable, nous avons $C^{\Pi}_{a_m} \succeq_{a_m} C_0 \cup \{a_m\}$. Le corollaire de cette propriété est qu'il existe au moins une structure de coalitions $\Pi' \in NS_{HG^{M_C}}$ pour chaque structure de coalitions Π stable dans le jeu HG.

Corollaire 4.1.1 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Pour toute structure de coalitions $\Pi \in NS_{HG}$, nous avons :

$$\begin{aligned} card_{M_C}(\Pi | HG) &= |\{C_0 \in \Pi | \forall C \in (\Pi \setminus \{C^{\Pi}_{a_m}\}) \cup \{\emptyset\}, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}\}| \\ &\geq 1 \end{aligned}$$

L'intuition derrière ce corollaire est qu'il existe nécessairement au moins une coalition $C_0 \in (\Pi \setminus \{C^{\Pi}_{a_m}\})$ respectant la seconde condition de la propriété 4.1.1 (et elle respecte aussi la première condition par définition de C_0). Remarquons que si le profil de préférence de l'agent malhonnête est strict, c'est-à-dire qu'il n'existe pas de coalition également préférée à une autre, alors $card_{M_C}(\Pi | HG) = 1$ pour toute structure de coalitions $\Pi \in NS_{HG}$.

Exemple 4.1.4 - Reprenons l'exemple 4.1.1. Il existe deux structures de coalitions stables : $\{\{a_1, a_2\}, \{a_3, a_m\}\}$ et $\{\{a_1, a_3\}, \{a_2, a_m\}\}$. Considérons la structure de coalitions $\{\{a_1, a_2\}, \{a_3, a_m\}\}$. Si s rejoint la coalition $\{a_1, a_2\}$, la partition $\Pi' = \{ \{a_1, a_2, s\}, \{a_3, a_m\} \} \notin NS_{HG}$ car $\{s\} \succ^{M_C}_s \{a_1, a_2, s\}$. De même, comme $C^{\Pi}_{a_m} \in \{a_3, a_m\}$, $\{ \{a_1, a_2\}, \{a_3, a_m, s\} \} \notin NS_{HG}$. En revanche, la

Π	C_0	$f(\Pi,s,C_0)$	$\Pi' \in NS_{HG^{M_C}}$?
$\{ \{a_1,a_2\},\{a_3,a_m\} \}$	$\{a_1,a_2\}$	$\{ \{a_1,a_2,s\},\{a_3,a_m\} \}$	×
	$\{a_3,a_m\}$	$\{ \{a_1,a_2\},\{a_3,a_m,s\} \}$	×
	$\emptyset$	$\{ \{a_1,a_2\},\{a_3,a_m\},\{s\} \}$	√
$\{ \{a_1,a_3\},\{a_2,a_m\} \}$	$\{a_1,a_3\}$	$\{ \{a_1,a_3,s\},\{a_2,a_m\} \}$	×
	$\{a_2,a_m\}$	$\{ \{a_1,a_3\},\{a_2,a_m,s\} \}$	×
	$\emptyset$	$\{ \{a_1,a_3\},\{a_2,a_m\},\{s\} \}$	√

Tableau 4.2 – Stabilité des structures de coalitions en fonction de C_0 lorsque $\Pi \in NS_{HG}$

structure de coalitions $\{ \{a_1,a_2\},\{a_3,a_m\},\{s\} \}$ est stable. Le tableau 4.2 récapitule la propriété de stabilité pour les structures de coalitions construites à partir de NS_{HG} en fonction de la coalition C_0 que rejoint l'agent Sybil s. Nous avons donc :

$$card_{M_C}(\{ \{a_1,a_2\},\{a_3,a_m\} \}|HG) = 1$$
$$card_{M_C}(\{ \{a_1,a_3\},\{a_2,a_m\} \}|HG) = 1$$
$$card_{M_C}(NS_{HG}|HG) = 2$$

Étudions maintenant l'influence de la manipulation constructive sur les structures de coalitions de HG qui ne sont pas stables, car au moins un agent honnête désire changer de coalition.

Propriété 4.1.2 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Suppons $\Pi \notin NS_{HG}$ telle qu'il existe un agent $a_i \in N \setminus \{a_m\}$ et une coalition $C \in \Pi \cup \{\emptyset\}$ vérifiant $C \cup \{a_i\} \succ_{a_i} C^{\Pi}_{a_i}$. Pour toute $C_0 \in \Pi \cup \{\emptyset\}$, nous avons :

$$\Pi' = f(\Pi,s,C_0) \notin NS_{HG^{M_C}}$$

Démonstration : Fixons une structure de coalitions $\Pi \notin NS_{HG}$ car il existe un agent $a_i \in N \setminus \{a_m\}$ tel que $\exists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_{a_i} C^{\Pi}_{a_i}$. Par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute (hypothèse 3.2.3 et 3.2.4), nous avons :

$$C \cup \{a_i\} \sim^{M_C}_{a_i} C \cup \{a_i,s\} \succ^{M_C}_{a_i} C^{\Pi}_{a_i} \sim^{M_C}_{a_i} C^{\Pi}_{a_i} \cup \{s\}$$

Par construction de Π', soit $C \in \Pi'$, soit $C \cup \{s\} \in \Pi'$. Dans les deux cas, il existe une coalition $C \in \Pi' \cup \{\emptyset\}$ telle que $C \cup \{a_i\} \succ^{M_C}_{a_i} C^{\Pi'}_{a_i}$. Par conséquent, la structure de coalitions Π' n'est pas stable. □

Ainsi selon la propriété 4.1.2, toute structure de coalitions résultante de l'ajout de s dans l'une des coalitions de Π n'est pas stable au sens de Nash pour le jeu HG^{M_C}.

Corollaire 4.1.2 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Pour toute structure de coalitions $\Pi \notin NS_{HG} \cup UR^{HG}_{a_m}$, $card_{M_C}(\Pi|HG) = 0$.

Étudions maintenant le dernier cas, c'est-à-dire l'influence de la manipulation constructive sur les structures de coalitions qui ne sont pas stables, car l'agent malhonnête a_m est l'unique responsable de la non-stabilité (définition 4.1.2).

Propriété 4.1.3 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Soit Π une structure de coalitions dont l'agent malhonnête a_m est l'unique responsable de la non-stabilité. La structure de coalitions $\Pi' = f(\Pi, s, C_0)$ est stable pour le jeu HG^{M_C} si et seulement si :

$$\forall C_1 \in \Pi \cup \{\emptyset\}, C_0 \cup \{a_m\} \succeq_{a_m} C_1 \cup \{a_m\}$$

Démonstration : Fixons une structure de coalitions $\Pi \in UR^{HG}_{a_m}$ et une coalition $C_0 \in (\Pi \setminus \{C^{\Pi}_{a_m}\}) \cup \{\emptyset\}$. Soit $\Pi' = f(\Pi, s, C_0)$ la structure de coalitions construite par l'ajout de s dans la coalition C_0.
Montrons dans un premier temps que s'il existe une coalition $C_1 \in \Pi \cup \{\emptyset\}$: $C_1 \cup \{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$ alors la structure de coalitions Π' ne peut pas être stable au sens de Nash. Par définition des préférences de l'agent Sybil s dans la manipulation constructive (définition 4.1.1), nous avons $C_1 \cup \{s\} \succ^{M_C}_{s} C_0 \cup \{s\}$. Par ailleurs, par construction de Π', soit $C_1 \in \Pi'$, soit $C_1 = \emptyset$. Dans les deux cas, par définition de la stabilité, $,\Pi' \notin NS_{HG^{M_C}}$ car l'agent s souhaite rejoindre la coalition C_1.
Montrons maintenant que si $\forall C_1 \in \Pi \cup \{\emptyset\}, C_0 \cup \{a_m\} \succeq_{a_m} C_1 \cup \{a_m\}$ alors la structure de coalitions Π' est stable au sens de Nash. Par définition d'un agent responsable de la non-stabilité d'une structure de coalitions (définition 4.1.2), cela vrai si et seulement si $C_0 \neq C^{\Pi}_m$. Par définition des préférences de a_m et de s dans la manipulation constructive $\forall C \in \Pi' \cup \{\emptyset\}$, nous avons :

$$C^{\Pi'}_{a_m} \succeq^{M_C}_{a_m} C \cup \{a_m\}$$
$$C^{\Pi'}_{s} \succeq^{M_C}_{s} C \cup \{s\}$$

Par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute (hypothèses 3.2.3 et 3.2.4), $\forall a_i \in N \setminus \{a_m\}, \forall C \in \Pi \cup \{\emptyset\}$, nous avons :

$$C \cup \{a_i\} \sim^{M_C}_{a_i} C \cup \{a_i, s\} \succ^{M_C}_{a_i} C^{\Pi}_{a_i} \sim^{M_C}_{a_i} C^{\Pi}_{a_i} \cup \{s\}$$

Par construction de Π', $\forall a_i \in N \setminus \{a_m\}$, soit $C^{\Pi'}_{a_i} = C^{\Pi}_{a_i}$, soit $C^{\Pi'}_{a_i} \cup \{s\}$. Dans les deux cas, $\forall a_i \in N \setminus \{a_m\}$, nous avons $\forall C \in \Pi' \cup \{\emptyset\}, C^{\Pi'}_{a_i} \succeq^{M_C}_{a_i} C \cup \{a_i\}$. Par définition, la partition Π' est donc stable dans le jeu HG^{M_C}. □

Remarquons que, par définition d'agent responsable de la non-stabilité de Π (définition 4.1.2), il existe nécessairement au moins une coalition $C_0 \in \Pi \cup \{\emptyset\}$ telle que $\forall C_1 \in \Pi \cup \{\emptyset\}, C_0 \cup \{a_m\} \succeq_{a_m} C_1 \cup \{a_m\}$. Nous pouvons ainsi déduire le nombre de structures de coalitions stables dans le jeu HG^{M_C} construit à partir d'une structure de coalitions où l'agent malhonnête est l'unique responsable de la non-stabilité.

Corollaire 4.1.3 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Pour

toute structure de coalitions $\Pi \in UR_{a_m}^{HG}$, nous avons :

$$\begin{aligned} card_{M_C}(\Pi|HG) &= |\{C_0 \in \Pi | \forall C \in \Pi \cup \{\emptyset\}, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}\}| \\ &\geq 1 \end{aligned}$$

Exemple 4.1.5 - Reprenons l'exemple 4.1.1. L'ensemble des structures de coalitions où l'agent malhonnête est l'unique responsable de la non-stabilité sont : $\{\{a_1\},\{a_2,a_3,a_m\}\}$, $\{\{a_1,a_2,a_m\},\{a_3\}\}$ et $\{\{a_1,a_2,a_3,a_m\}\}$. La table 4.3 indique si une structure de coalitions Π' est stable en fonction de la coalition C_0 que rejoint l'agent Sybil s.

Π	C_0	$f(\Pi,s,C_0)$	$\Pi' \in NS_{HG^{M_C}}$?
$\{\{a_1\},\{a_2,a_3,a_m\}\}$	$\{a_1\}$	$\{\{a_1,s\},\{a_2,a_3,a_m\}\}$	$\surd$
	$\{a_2,a_3,a_m\}$	$\{\{a_1\},\{a_2,a_3,a_m,s\}\}$	$\times$
	$\emptyset$	$\{\{a_1\},\{a_2,a_3,a_m\},\{s\}\}$	$\times$
$\{\{a_1,a_2,a_m\},\{a_3\}\}$	$\{a_1,a_2,a_m\}$	$\{\{a_1,a_2,a_m,s\},\{a_3\}\}$	$\times$
	$\{a_3\}$	$\{\{a_1,a_2,a_m\},\{a_3,s\}\}$	$\surd$
	$\emptyset$	$\{\{a_1,a_2,a_m\},\{a_3\},\{s\}\}$	$\times$
$\{\{a_1,a_2,a_3,a_m\}\}$	$\{a_1,a_2,a_3,a_m\}$	$\{\{a_1,a_2,a_3,a_m,s\}\}$	$\times$
	$\emptyset$	$\{\{a_1,a_2,a_3,a_m\},\{s\}\}$	$\surd$

Tableau 4.3 – Stabilité au sens de Nash de $f(\Pi,s,C_0)$ en fonction de C_0 et de $\Pi \in UR_{a_m}^{HG}$

De ce fait, nous avons :

$$\begin{aligned} &card_{M_C}(\{\{a_1\},\{a_2,a_3,a_m\}\}|HG) = 1 \\ &card_{M_C}(\{\{a_1,a_2,a_m\},\{a_3\}\}|HG) = 1 \\ &card_{M_C}(\{\{a_1,a_2,a_3,a_m\}\}|HG) = 1 \\ &card_{M_C}(UR_{a_m}^{HG}|HG) = 3 \end{aligned}$$

Remarquons que si le profil de préférence de a_m est strict comme dans l'exemple 4.1.5 alors il existe une seule et unique structure de coalitions $\Pi' \in NS_{HG^{M_C}}$ construite de toute structure de coalitions $\Pi \in UR_{a_m}^{HG}$.

Grâce aux corollaires 4.1.1, 4.1.2 et 4.1.3, nous avons montré que la manipulation constructive permet d'augmenter le nombre de structures de coalitions stables sans pour autant rendre instables dans HG^{M_C} celles qui l'étaient dans HG. En effet, comme le montrent les exemples 4.1.1, 4.1.4 et 4.1.5, la manipulation constructive permet de passer d'un jeu hédonique ayant deux structures de coalitions stables à un jeu HG^{M_C} où il en a cinq.

4.1.2 Conditions de rationalité

Augmenter le nombre de structures de coalitions ne suffit pas à augmenter les probabilités de satisfaction des agents. En effet, il est possible de créer des structures stables, mais non satisfaisantes. Si ces dernières sont plus nombreuses que les structures stables satisfaisantes créées alors certaines probabilités de satisfaction peuvent diminuer. Ainsi, la manipulation constructive

est k-rationnelle si ces nouvelles structures de coalitions appartiennent au concept d'acceptation de profondeur k.

Exemple 4.1.6 - Prenons le jeu $HG = \langle N, \succeq, \mathbb{P} \rangle$ suivant :

$$\begin{aligned} N &= \{a_1,a_2,a_3,a_m\} \\ \succeq_{a_1} &= \{a_1,a_2,a_3\} \succ_{a_1} \{a_1,a_2,a_3,a_m\} \succ_{a_1} \{a_1,a_2\} \succ_{a_1} \{a_1\} \\ \succeq_{a_2} &= \{a_1,a_2,a_3\} \succ_{a_2} \{a_1,a_2,a_3,a_m\} \succ_{a_2} \{a_1,a_2\} \succ_{a_2} \{a_2\} \\ \succeq_{a_3} &= \{a_1,a_2,a_3\} \sim_{a_3} \{a_3,a_m\} \succ_{a_3} \{a_1,a_2,a_3,a_m\} \succ_{a_3} \{a_3\} \\ \succeq_{a_m} &= \{a_3,a_m\} \succ_{a_m} \{a_1,a_2,a_3,a_m\} \succ_{a_m} \{a_m\} \end{aligned}$$

Parmi les 15 structures de coalitions possibles, seulement 2 d'entre elles sont stables $\{\{a_1,a_2\},\{a_3,a_m\}\}$ et $\{\{a_1,a_2,a_3,a_m\}\}$. Ainsi, les probabilités de satisfaction de l'agent malhonnête sont :

$$\begin{aligned} \mathrm{P}^*(a_m,1|HG) &= 1/2 \\ \forall i \in [2,8] : \mathrm{P}^*(a_m,i|HG) &= 1 \end{aligned}$$

Comme l'agent malhonnête a_m est l'unique responsable de la non-stabilité de la structure de coalitions $\{\{a_1,a_2,a_3\},\{a_m\}\}$, nous avons :

$$NS_{HG^{M_C}} = \{\{\{a_1,a_2\},\{a_3,a_m\},\{s\}\},\{\{a_1,a_2,a_3,a_m\},\{s\}\},\{\{a_1,a_2,a_3,s\},\{a_m\}\}\}$$

Ainsi, les probabilités de satisfaction du jeu HG^{M_C} sont :

$$\begin{aligned} \mathrm{P}^*(a_m,1|HG^{M_C}) &= 1/3 \\ \forall i \in [2,8] : \mathrm{P}^*(a_m,i|HG^{M_C}) &= 1 \end{aligned}$$

La manipulation constructive mise en œuvre par a_m réduit ainsi la probabilité de satisfaction de degré 1 et n'est donc pas rationnelle.

Cet exemple nous montre qu'effectuer une manipulation constructive n'est pas toujours rationnel. L'agent malhonnête doit calculer si la manipulation est rationnelle avant sa mise en œuvre. Dans la suite de cette section, nous montrons quelles sont les conditions minimales nécessaires à la k-rationalité de la manipulation constructive. Par les corollaires 4.1.1, 4.1.2 et 4.1.3, nous avons :

$$\begin{aligned} \forall \Pi \in NS_{HG}, card_{M_C}(\Pi|HG) &\geq 1 \\ \forall \Pi \in UR_{a_m}^{HG}, card_{M_C}(\Pi|HG) &\geq 1 \\ \forall \Pi \notin NS_{HG} \cup UR_{a_m}^{HG}, card_{M_C}(\Pi|HG) &= 0 \end{aligned}$$

Par extension aux ensembles de structures de coalitions, nous avons donc :

$$\begin{aligned} card_{M_C}(NS_{HG}|HG) &= \sum_{\Pi \in NS_{HG}} card_{M_C}(\Pi|HG) \geq |NS_{HG}| \\ card_{M_C}(UR_{a_m}^{HG}|HG) &= \sum_{\Pi \in UR_{a_m}^{HG}} card_{M_C}(\Pi|HG) \geq |UR_{a_m}^{HG}| \end{aligned}$$

Ces inégalités nous permettent de déduire le nombre de structures de coalitions stables dans le jeu HG^{M_C} à partir de la connaissance des structures de coalitions stables dans HG et des structures de coalitions dont l'agent malhonnête est l'unique responsable de la non-stabilité.

Propriété 4.1.4 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Le nombre de structures de coalitions stables au sens de Nash dans HG^{M_C} est :

$$\begin{aligned} |NS_{HG^{M_C}}| &= card_{M_C}(NS_{HG}|HG) + card_{M_C}(UR_{a_m}^{HG}|HG) \\ &\geq |NS_{HG}| + |UR_{a_m}^{HG}| \end{aligned}$$

À partir des propriétés 4.1.1, 4.1.2, 4.1.3 et 4.1.4, nous pouvons déduire les conditions nécessaires à la rationalité de la manipulation constructive. Rappelons que, de par la définition 3.2.5, la manipulation constructive est k-rationnelle si elle permet d'améliorer la probabilité de l'agent malhonnête d'être dans la coalition $C_{a_m,k}$ sans diminuer pour autant la probabilité d'être dans toutes les coalitions $C_{a_m,i}$ préférées à $C_{a_m,k}$. De manière générique, nous avons la propriété suivante.

Propriété 4.1.5 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique tel que $NS_{HG} \neq \emptyset$ et $a_m \in N$ un agent malhonnête ayant le profil de préférence $\succeq_{a_m} = C_{a_m,1} \succeq_{a_m} C_{a_m,2} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. La manipulation constructive mise en œuvre par a_m est k-rationnelle seulement si :

$$\begin{aligned} \forall i \in [1,k[, \frac{|\{\Pi \in NS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,i}\}|}{|NS_{HG}|} &= \frac{(1)_i + (2)_i}{(3)} \\ \text{et } \frac{|\{\Pi \in NS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,k}\}|}{|NS_{HG}|} &< \frac{(1)_k + (2)_k}{(3)} \end{aligned}$$

où :

- $(1)_k = card_{M_C}(\{\Pi \in NS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,k}\}|HG)$;
- $(2)_k = card_{M_C}(\{\Pi \in UR_{a_m}^{HG} | \exists C_0 \in \Pi \cup \{\emptyset\} : (2.1)_k \wedge (2.2)\}|HG)$;
- $(2.1)_k = C_0 \cup \{a_m\} \sim_{a_m} C_{a_m,k}$;
- $(2.2) = \forall C \in \Pi, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}$;
- $(3) = card_{M_C}(NS_{HG}|HG) + card_{M_C}(UR_{a_m}^{HG}|HG)$.

Pour des raisons de lisibilité, la démonstration de cette propriété peut être trouvée en annexe B. Intuitivement, ces conditions correspondent au fait que l'ajout de l'agent s créé un plus grand nombre de structures de coalitions contenant la coalition $C_{a_m,k}$ que de structures de coalitions ne la contenant pas. Dans cette propriété,

- $(1)_k$ est le nombre de structures de coalitions de $NS_{HG^{M_C}}$ contenant $C_{a_m,k}$ construites à partir des coalitions de NS_{HG} ;
- $(2)_k$ est le nombre de structures de coalitions de $NS_{HG^{M_C}}$ contenant $(C_{a_m,k} \setminus \{a_m\}) \cup \{s\}$ construites à partir des coalitions de $UR_{a_m}^{HG}$;
- (3) est la cardinalité de l'ensemble $NS_{HG^{M_C}}$.

Cette propriété caractérise les conditions nécessaires à la k-rationalité d'une manipulation constructive dans le cas où $NS_{HG} \neq \emptyset$. Si $NS_{HG} = \emptyset$ ces conditions changent. En effet, lorsque $NS_{HG} = \emptyset$, $\frac{|\{\Pi \in NS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,k}\}|}{|NS_{HG}|}$ est indéfini. Ainsi,

Propriété 4.1.6 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique tel que $NS_{HG} = \emptyset$ et $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. La manipulation

constructive mise en œuvre par a_m est k-rationnelle si :

$$\forall i \in [1,k[: \forall \Pi \in UR^{HG}_{a_m}, C_{a_m,i} \setminus \{a_m\} \notin \Pi$$
$$\exists \Pi \in UR^{HG}_{a_m} : C_{a_m,k} \setminus \{a_m\} \in \Pi$$

Intuitivement, en l'absence de structures de coalitions stables pour HG, la manipulation est k-rationnelle s'il existe une structure de coalitions dont l'agent malhonnête est l'unique responsable de la non-stabilité, car il désire rejoindre la coalition $C_{a_m,k} \setminus \{a_m\}$.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ tel que $NS_{HG} = \emptyset$ et un agent malhonnête a_m ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Soit $x \in]k,2^{n-1}]$ tel que $C_{a_m,x} = \{a_m\}$. Nous supposons ,$k < x$ car, dans le cas contraire, l'agent malhonnête n'a pas d'intérêt à manipuler le jeu. Par l'hypothèse de rationalité minimale (hypothèse 3.1.1), la solution du jeu HG^{M_C} est la structure de coalitions $\{\{a_1\},\ldots,\{a_{n-1}\},\{a_m\}\}$. Nous avons donc $\forall i \in [1,x[, \mathrm{P}^*(a_m,i|HG^{M_C}) = 0$.
Prouvons la première condition. Par définition de la k-rationalité (définition 3.2.5), la manipulation constructive est k-rationnelle si $\forall i : 1 \leq i < k, \mathrm{P}^*(a_m,i|HG) = \mathrm{P}^*(a_m,i|HG^{M_C}) = 0$. Or, comme la solution du jeu est supposée tirée aléatoirement uniformément parmi l'ensemble des structures de coalitions stables, $\mathrm{P}^*(a_m,i|HG^{M_C}) = 0$ si et seulement si $\nexists \Pi' \in NS_{HG^{M_C}}$ telle que $C^{\Pi'}_{a_m} \sim_{a_m} Ca_m,k$ ou $(C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\} \sim_{a_m} Ca_m,k$. Comme $NS_{HG} = \emptyset$, par la propriété 4.1.3, nous avons $f^{-1}(\Pi') \in UR^{HG}_{a_m}$ pour toute structure de coalitions $\Pi' \in NS_{HG^{M_C}}$. Par conséquent, $\forall \Pi \in UR^{HG}_{a_m}, C_{a_m,i} \setminus \{a_m\} \notin \Pi$ et donc $\nexists \Pi' \in NS_{HG^{M_C}}$ telle que $C^{\Pi'}_{a_m} \sim_{a_m} Ca_m,k$ ou $(C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\} \sim_{a_m} Ca_m,k$.
Prouvons la seconde condition. Pour qu'il existe une structure de coalitions $\Pi' \in NS_{HG^{M_C}}$ afin que l'inégalité $\mathrm{P}^*(a_m,k|HG^{M_C}) > 0$ soit satisfaite, il est nécessaire que $\exists \Pi \in UR^{HG}_{a_m} : C_{a_m,k} \setminus \{a_m\} \in \Pi$. □

Exemple 4.1.7 - Soit le jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ avec :

$$N = \{a_1,a_2,a_m\}$$
$$\succeq_{a_1} = \{a_1,a_2\} \succ_{a_1} \{a_1,a_m\} \succ_{a_1} \{a_1\}$$
$$\succeq_{a_2} = \{a_1,a_2\} \succ_{a_2} \{a_2,a_m\} \succ_{a_2} \{a_2\}$$
$$\succeq_{a_m} = \{a_1,a_2,a_m\} \succ_{a_m} \{a_1,a_m\} \succ_{a_m} \{a_m\}$$

Ce jeu ne dispose pas de structure de coalitions stable. Par contre, $UR^{HG}_{a_m} = \{\{\{a_1,a_2\},\{a_m\}\}\}$. Ainsi, l'agent malhonnête a_m peut mettre en œuvre la manipulation constructive M_C en étant sûr que cette dernière est 1-rationnelle. En effet, comme nous avons $NS_{HG} = \emptyset$ et $NS_{HG^{M_C}} = \{\{\{a_1,a_2,s\},\{a_m\}\}\}$, les probabilités de satisfaction de a_m à une profondeur 1 dans HG et dans HG^{M_C} sont respectivement :

$$\mathrm{P}^*(a_m,1|HG) = 0$$
$$\mathrm{P}^*(a_m,1|HG^{M_C}) = 1$$

Notons que l'une des conséquences de la propriété 4.1.6 est que l'existance d'au moins une structure de coalitions dont l'agent malhonnête est l'unique responsable de la non-stabilité et une condition nécessaire à la rationalité de la manipulation constructive.

Corollaire 4.1.4 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête. Si $NS_{HG} = \emptyset$ et $UR^{HG}_{a_m} = \emptyset$, alors la manipulation constructive ne peut pas être k-rationnelle.

Il existe un cas particulier intéressant qui nous permet de simplifier les caractérisations précédentes : celui où l'agent malhonnête a un profil de préférence strict.

Propriété 4.1.7 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succ_{a_m} \ldots \succ_{a_m} C_{a_m,2^{n-1}}$. Notons par $(1)_k$ l'expression $\forall i \in [1,k[, C_{a_m,i} \setminus \{a_m\} \notin \Pi$. La manipulation constructive est k-rationnelle si pour tout $i \in [1,k[$ les deux conditions suivantes sont vérifiées :

$$\frac{|\{\Pi \in NS_{HG} | C_{a_m,i} \in \Pi\}|}{|NS_{HG}|} = \frac{|\{\Pi \in UR^{HG}_{a_m} | Ca_m,i \setminus \{a_m\} \in \Pi \wedge (1)_i\}|}{|UR^{HG}_{a_m}|} \tag{4.1}$$

$$\frac{|\{\Pi \in NS_{HG} | C_{a_m,k} \in \Pi\}|}{|NS_{HG}|} < \frac{|\{\Pi \in UR^{HG}_{a_m} | Ca_m,k \setminus \{a_m\} \in \Pi \wedge (1)_k\}|}{|UR^{HG}_{a_m}|} \tag{4.2}$$

Cette propriété signifie que si le profil de préférence de l'agent malhonnête est strict, la manipulation constructive est k-rationnelle si la proportion de structures de coalitions stables contenant la coalition Ca_m,k est supérieure à la proportion de structures de coalitions de $UR^{HG}_{a_m}$ contenant la coalition $Ca_m,k \setminus \{a_m\}$. Encore pour raison de lisibilité, la démonstration de cette propriété peut être trouvée en annexe B. Intuitivement, la démonstration repose sur le fait que les préférences strictes induisent une simplification des conditions de la propriété 4.1.5. Le corollaire de cette propriété est que, si l'agent malhonnête a un profil de préférence strict, la manipulation constructive ne peut être rationnelle que s'il existe aux moins une structure de coalitions dont il est l'unique responsable de la non-stabilité.

Corollaire 4.1.5 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succ_{a_m} \ldots \succ_{a_m} C_{a_m,2^{n-1}}$. Si $UR^{HG}_{a_m} = \emptyset$ alors, pour tout $k \in [1,2^{n-1}]$, la manipulation constructive n'est pas k-rationnelle.

Ces propriétés nous permettent d'affirmer qu'un agent malhonnête a_m ayant une connaissance des préférences des autres agents du système (hypothèse 3.2.1) peut calculer s'il est rationnel de mettre en œuvre une manipulation constructive.

4.2 Manipulation destructive

4.2.1 Une manipulation qui réduit le nombre de solutions

Un agent malhonnête qui effectue une manipulation constructive cherche à augmenter sa probabilité de satisfaction (définition 3.1.6) en rendant stables les structures de coalitions dont il est l'unique responsable de la non-stabilité (définition 4.1.2). Nous proposons ici une seconde manipulation où l'agent malhonnête réduit le nombre de structures de coalitions stables au sens de Nash. Cette manipulation est appelée la manipulation *destructive*. L'intuition de cette manipulation est d'ajouter un agent Sybil s qui sera l'unique responsable de la non-stabilité des structures de coalitions n'appartenant pas au concept d'acceptation désiré. Pour cela, l'agent

malhonnête révèle son véritable profil de préférence et introduit un agent Sybil s désirant rejoindre toutes coalitions de $\mathcal{C}^N_{a_m}$ hormis une coalition $C_{a_m,k}$ préalablement fixée.

Définition 4.2.1 - *Manipulation destructive* : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$ et un agent malhonnête $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. La *manipulation destructive* mise en œuvre par a_m sur HG est $M_D = \langle \{a_m,s\}, \{\succeq^{M_D}_{a_m}, \succeq^{M_D}_{s}\}, \succeq_{a_m}\rangle$ où :

$$\begin{aligned}
\succeq^{M_D}_{a_m} &= C_{a_m,1} \succeq^{M_D}_{a_m} \ldots \succeq^{M_D}_{a_m} C_{a_m,k} \succeq^{M_D}_{a_m} \ldots \succeq^{M_D}_{a_m} C_{a_m,2^{n-1}} \\
\succeq^{M_D}_{s} &= C_{a_m,1} \cup \{s\} \succ^{M_D}_{s} \ldots \succ^{M_D}_{s} C_{a_m,k-1} \cup \{s\} \succ^{M_D}_{s} C_{a_m,k+1} \cup \{s\} \\
&\quad \succ^{M_D}_{s} \ldots \succ^{M_D}_{s} C_{a_m,2^{n-1}} \cup \{s\} \succ^{M_D}_{s} C_{a_m,k} \cup \{s\}
\end{aligned}$$

Dans cette manipulation, l'agent Sybil prétend vouloir être dans toutes les coalitions contenant a_m autre que $C_{a_m,k}$ tandis que a_m refuse donc toutes les coalitions contenant s car il présente son véritable profil de préférence.

Exemple 4.2.1 - Reprenons l'exemple 4.1.1.

$$\begin{aligned}
N &= \{a_1,a_2,a_3,a_m\} \\
\succeq_{a_1} &= \{a_1,a_2\} \succ_{a_1} \{a_1,a_3,a_m\} \succ_{a_1} \{a_1,a_3\} \sim_{a_1} \{a_1,a_2,a_m\} \succ_{a_1} \{a_1,a_2,a_3,a_m\} \sim_{a_1} \{a_1\} \\
\succeq_{a_2} &= \{a_1,a_2\} \sim_{a_2} \{a_2,a_3,a_m\} \succ_{a_2} \{a_1,a_2,a_3\} \sim_{a_2} \{a_2,a_m\} \succ_{a_2} \{a_1,a_2,a_m\} \succ_{a_2} \{a_2,a_3\} \\
&\quad \succ_{a_2} \{a_1,a_2,a_3,a_m\} \sim_{a_2} \{a_2\} \\
\succeq_{a_3} &= \{a_1,a_3\} \sim_{a_3} \{a_2,a_3,a_m\} \succ_{a_3} \{a_3,a_m\} \succ_{a_3} \{a_1,a_2,a_3\} \succ_{a_3} \{a_2,a_3\} \\
&\quad \succ_{a_3} \{a_1,a_2,a_3,a_m\} \sim_{a_3} \{a_3\} \\
\succeq_{a_m} &= \{a_1,a_m\} \succ_{a_m} \{a_2,a_m\} \succ_{a_m} \{a_3,a_m\} \succ_{a_m} \{a_m\}
\end{aligned}$$

Les solutions stables du jeu sont $NS_{HG} = \{\{\{a_1,a_2\},\{a_3,a_m\}\},\{\{a_1,a_3\},\{a_2,a_m\}\}\}$. Supposons que l'agent malhonnête fixe $C_{a_m,k} = \{a_2,a_m\}$, la manipulation destructive a_m est alors :

$$\begin{aligned}
\succeq^{M_D}_{a_m} &= \{a_1,a_m\} \succ^{M_D}_{a_m} \{a_2,a_m\} \succ^{M_D}_{a_m} \{a_3,a_m\} \succ^{M_D}_{a_m} \{a_m\} \\
\succeq^{M_D}_{s} &= \{a_1,a_m,s\} \succ^{M_D}_{s} \{a_3,a_m,s\} \succ^{M_D}_{s} \{s\}
\end{aligned}$$

La mise en œuvre de M_D sur HG donne le jeu $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$ où :

$$N^{M_D} = \{a_1,a_2,a_3,a_m,s\}$$

$$\begin{aligned}
\succeq_1^{M_D} = {} & \{a_1,a_2\} \sim_{a_1}^{M_D} \{a_1,a_2,s\} \succ_{a_1}^{M_D} \{a_1,a_3,a_m\} \sim_{a_1}^{M_D} \{a_1,a_3,a_m,s\} \\
& \succ_{a_1}^{M_D} \{a_1,a_3\} \sim_{a_1}^{M_D} \{a_1,a_3,s\} \sim_{a_1}^{M_D} \{a_1,a_2,a_m\} \sim_{a_1}^{M_D} \{a_1,a_2,a_m,s\} \\
& \succ_{a_1}^{M_D} \{a_1,a_2,a_3,a_m\} \sim_{a_1}^{M_D} \{a_1,a_2,a_3,a_m,s\} \sim_{a_1}^{M_D} \{a_1\} \sim_{a_1}^{M_D} \{a_1,s\} \\
\succeq_2^{M_D} = {} & \{a_1,a_2\} \sim_{a_2}^{M_D} \{a_1,a_2,s\} \sim_{a_2}^{M_D} \{a_2,a_3,a_m\} \sim_{a_2}^{M_D} \{a_2,a_3,a_m,s\} \\
& \succ_{a_2}^{M_D} \{a_1,a_2,a_3\} \sim_{a_2}^{M_D} \{a_1,a_2,a_3,s\} \sim_{a_2}^{M_D} \{a_2,a_m\} \sim_{a_2}^{M_D} \{a_2,a_m,s\} \\
& \succ_{a_2}^{M_D} \{a_1,a_2,a_m\} \sim_{a_2}^{M_D} \{a_1,a_2,a_m,s\} \succ_{a_2}^{M_D} \{a_2,a_3\} \sim_{a_2}^{M_D} \{a_2,a_3,s\} \\
& \succ_{a_2}^{M_D} \{a_1,a_2,a_3,a_m\} \sim_{a_2}^{M_D} \{a_1,a_2,a_3,a_m,s\} \sim_{a_2}^{M_D} \{a_2\} \sim_{a_2}^{M_D} \{a_2,s\} \\
\succeq_3^{M_D} = {} & \{a_1,a_3\} \sim_{a_3}^{M_D} \{a_1,a_3,s\} \sim_{a_3}^{M_D} \{a_2,a_3,a_m\} \sim_{a_3}^{M_D} \{a_2,a_3,a_m,s\} \\
& \succ_{a_3}^{M_D} \{a_3,a_m\} \sim_{a_3}^{M_D} \{a_3,a_m,s\} \succ_{a_3}^{M_D} \{a_1,a_2,a_3\} \sim_{a_3}^{M_D} \{a_1,a_2,a_3,s\} \\
& \succ_{a_3}^{M_D} \{a_2,a_3\} \sim_{a_3}^{M_D} \{a_2,a_3,s\} \succ_{a_3}^{M_D} \{a_1,a_2,a_3,a_m\} \sim_{a_3}^{M_D} \{a_1,a_2,a_3,a_m,s\} \\
& \sim_{a_3}^{M_D} \{a_3\} \sim_{a_3}^{M_D} \{a_3,s\}
\end{aligned}$$

$$\begin{aligned}
\succeq_{a_m}^{M_D} &= \{a_1,a_m\} \succ_{a_m}^{M_D} \{a_2,a_m\} \succ_{a_m}^{M_D} \{a_3,a_m\} \succ_{a_m}^{M_D} \{a_m\} \\
\succeq_s^{M_D} &= \{a_1,a_m,s\} \succ_s^{M_D} \{a_3,a_m,s\} \succ_s^{M_D} \{s\}
\end{aligned}$$

La figure 4.3 représente l'ensemble des structures de coalitions stables pour les jeux HG et HG^{M_D}.

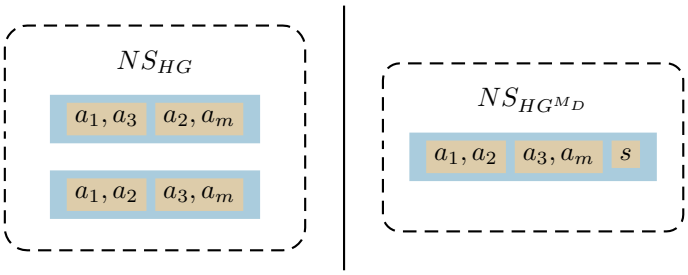

FIGURE 4.3 – Structures de coalitions stables au sens de Nash de HG et de HG^{M_D}

Dans cet exemple, la manipulation réduit le nombre de structures de coalitions stables afin que seules les structures de coalitions appartenant au concept d'acceptation de profondeur k aient une probabilité de sélection non nulle. Cela est vrai pour tout jeu hédonique. Pour prouver cela, étudions tout d'abord les conditions nécessaires à la stabilité d'une structure de coalitions dans $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$.

Propriété 4.2.1 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation destructive par $a_m \in N$ sur HG. Une structure de coalitions $\Pi' \in \mathcal{P}_{N^{M_D}}$ n'est pas stable au sens de Nash si une des deux conditions suivantes n'est pas vérifiée : $C_{a_m}^{\Pi'} = C_{a_m,k}$ et $C_s^{\Pi'} = \{s\}$.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation destructive par $a_m \in N$ sur HG. Supposons une structure de coalitions $\Pi' \in NS_{HG^{M_D}}$.
Supposons dans un premier temps que $C_{a_m}^{\Pi'} = C_s^{\Pi'}$. Par définition du profil de préférence de a_m, nous avons $\{a_m\} \succ_{a_m}^{M_D} C \cup \{s\}, \forall C \in \mathcal{C}_{a_m}^N$. Par conséquent, si $C_{a_m}^{\Pi'} = C_s^{\Pi'}$ alors $\{a_m\} \succ_{a_m}^{M_D} C_{a_m}^{\Pi'}$. Ceci contredit $\Pi' \in NS_{HG^{M_D}}$ et donc $C_{a_m}^{\Pi'} \neq C_s^{\Pi'}$.
Supposons maintenant que $C_{a_m}^{\Pi'} \neq C_{a_m,k}$. Par définition du profil de préférence de s, nous avons $C_{a_m}^{\Pi'} \cup \{s\} \succ_s^{M_D} C_s^{\Pi'}$ car $C_{a_m}^{\Pi'} \neq C_s^{\Pi'}$. Ceci contredit $\Pi' \in NS_{HG^{M_D}}$ et donc $C_{a_m}^{\Pi'} = C_{a_m,k}$.
Supposons enfin que $C_s^{\Pi'} \neq \{s\}$. Par définition du profil de préférence de s, nous avons $\{s\} \succ_s^{M_D} C_s^{\Pi'}$ car $C_{a_m}^{\Pi'} \neq C_s^{\Pi'}$. Ceci contredit $\Pi' \in NS_{HG^{M_D}}$ et donc $C_s^{\Pi'} = \{s\}$. □

Cette propriété nous permet d'affirmer que si le jeu $NS_{HG^{M_D}}$ n'est pas vide, alors toute structure de coalitions stable contient nécessairement la coalition $C_{a_m,k}$ et les probabilités de satisfaction de l'agent malhonnête sont caractérisées par le corollaire suivant :

Corollaire 4.2.1 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation destructive par $a_m \in N$ sur HG. Si $NS_{HG^{M_D}} \neq \emptyset$, nous avons alors :

1. $\forall i \in [1,k[: \mathrm{P}^*(a_m,i|HG^{M_D}) = 0$;
2. $\forall i \in [k,2^{n-1}] : \mathrm{P}^*(a_m,i|HG^{M_D}) = 1$.

Regardons maintenant quelles sont les structures de coalitions stables après mise en œuvre de la manipulation destructive. Pour cela, montrons dans un premier temps que si une structure de coalitions n'est pas stable dans HG, la structure de coalitions construite en y ajoutant un agent Sybil n'est pas stable dans $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$.

Propriété 4.2.2 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation destructive par $a_m \in N$ sur HG. Soit $\Pi \notin NS_{HG}$ une structure de coalitions non stable dans HG. $\forall C_0 \in \Pi \cup \{\emptyset\}$, la structure de coalitions $\Pi' = f(\Pi,s,C_0)$ n'est pas stable.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et une structure de coalitions $\Pi \notin NS_{HG}$. Par la propriété 4.2.1, $\forall C_0 \in \Pi$, nous avons $\Pi' = f(\Pi,s,C_0) \notin NS_{HG^{M_D}}$. En effet, par construction de Π', nous avons $C_s^{\Pi'} \neq \{s\}$ et donc Π' ne peux pas être stable.
Considérons maintenant la structure de coalitions $\Pi' = f(\Pi,s,\emptyset)$ et supposons que $\Pi' \in NS_{HG^{M_D}}$. Par définition de la stabilité (définition 2.1.9), nous avons $\forall a_i \in N^{M_D}$, $\nexists C \in \Pi' \cup \{\emptyset\}, C \cup \{a_i\} \succ_{a_i}^{M_D} C_{a_i}^{\Pi'}$.
Comme $C_s^{\Pi'} = \{s\}$, nous avons $\forall a_i \in N, C_{a_i}^{\Pi'} = C_{a_i}^{\Pi}$. Par définition du profil de préférence de a_m, si $\nexists C \in \Pi' \cup \{\emptyset\}$ telle que $C \cup \{a_m\} \succ_{a_m}^{M_D} C_{a_m}^{\Pi'}$ alors $\nexists C \in \Pi \cup \{\emptyset\}, C \cup \{a_m\} \succ_{a_m} C_{a_m}^{\Pi}$.
De même par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute (hypothèses 3.2.3 et 3.2.4), si $\forall a_i \in N \setminus \{a_m\}$, $\nexists C \in \Pi' \cup \{\emptyset\}, C \cup \{a_i\} \succ_{a_i}^{M_D} C_{a_i}^{\Pi'}$ alors $\forall a_i \in N \setminus \{a_m\}$, $\nexists C \in \Pi \cup \{\emptyset\}, C \cup \{a_i\} \succ_{a_i} C_{a_i}^{\Pi}$. Ainsi, par définition de la stabilité si $\Pi' \in NS_{HG^{M_D}}$ alors $\Pi \in NS$, ce qui est en contradiction avec $\Pi \notin NS_{HG}$.
Ainsi, pour toute structure de coalitions $\Pi \notin NS_{HG}$, quelle que soit $C_0 \in \Pi \cup \{\emptyset\}$, la structure de coalitions $\Pi' = f(\Pi,s,C_0)$ n'est pas stable dans le jeu HG^{M_D}. □

Étudions maintenant l'effet de la manipulation destructive sur les structures de coalitions stables.

Propriété 4.2.3 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation destructive par $a_m \in N$ sur HG. Pour toute structure de coalitions $\Pi \in NS_{HG}$ telle que $C_{a_m,k} \in \Pi$ (respectivement $C_{a_m,k} \notin \Pi$), la structure de coalitions $\Pi' = f(\Pi, s, \emptyset)$ est stable (respectivement non stable) pour le HG^{M_D}.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et une structure de coalitions $\Pi \in NS_{HG}$. Soit $\Pi' = f(\Pi, s, \emptyset)$ la structure de coalitions obtenue après ajout de l'agent s dans sa coalition singleton.
Montrons dans un premier que si $C_{a_m,k} \notin \Pi$ alors Π' n'est pas stable. Par construction de Π', nous avons $C^{\Pi'}_{a_m} = C^{\Pi}_{a_m}$. Comme $C_{a_m,k} \notin \Pi$, nous avons $C^{\Pi'}_{a_m} \neq C_{a_m,k}$. Ainsi, par la propriété 4.2.1, la structure de coalitions $\Pi' = f(\Pi, s, \emptyset)$ n'est pas stable.
Montrons dans un second que si $C_{a_m,k} \in \Pi$ alors Π' est stable. Par construction de Π', nous avons $\forall a_i \in N, C^{\Pi}_{a_i} = C^{\Pi'}_{a_i}$. Par la définition 2.1.9, nous avons $\forall a_i \in N,\ \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_i\} \succ_{a_i} C^{\Pi}_{a_i}$. Par définition des préférences de a_m et de s dans la manipulation destructive (définition 4.2.1), nous avons donc :

$$\nexists C \in \Pi' \cup \{\emptyset\} : C \cup \{a_m\} \succ^{M_D}_{a_m} C^{\Pi'}_{a_m}$$
$$C \cup \{s\} \succ^{M_D}_{s} C^{\Pi'}_{s}$$

Par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute (hypothèse 3.2.3 et 3.2.4), $\forall a_i \in N \setminus \{a_m\}$, nous avons aussi :

$$\nexists C \in \Pi' \cup \{\emptyset\}, C \cup \{a_i\} \succ^{M_D}_{a_i} C^{\Pi'}_{a_i}$$

Ainsi, la structure de coalitions Π' est stable pour le jeu HG^{M_D}. □

De cette propriété, nous pouvons déduire qu'il existe des structures de coalitions stables dans HG^{M_D} sous les conditions suivantes :

Corollaire 4.2.2 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation destructive par $a_m \in N$ sur HG. Il existe au moins une structure de coalitions stable pour le jeu HG^{M_D} si et seulement si :

$$\exists \Pi \in NS_{HG} \text{ telles que } C_{a_m,k} \in \Pi$$

4.2.2 Conditions de rationalité

Le corollaire 4.2.1 stipule que la manipulation destructive garantit à un agent malhonnête que, pour un k fixé, toute solution du jeu stable au sens de Nash contient la coalition $C_{a_m,k}$. Cependant, comme nous allons le montrer dans cette section, la manipulation destructive n'est pas nécessairement rationnelle. Étudions dans un premier temps le cas où le jeu hédonique HG ne possède aucune structure de coalitions stable.

Propriété 4.2.4 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête. Si $NS_{HG} = \emptyset$ alors la manipulation destructive M_D n'est pas k-rationnelle, $\forall k \in [1, 2^{n-1}]$.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ avec $N = \{a_1, \ldots, a_{n-1}, a_m\}$ tel que $NS_{HG} = \emptyset$ et que la solution soit donc l'ensemble des coalitions singletons. Par la propriété 4.2.2, nous avons $NS_{HG^{M_D}} = \emptyset$ et la solution est aussi l'ensemble des coalitions singleton. Ainsi, par l'hypothèse de rationalité individuelle minimale (hypothèse 3.1.1), nous avons $\mathbb{P}(HG) = \{ \{a_1\}, \ldots, \{a_{n-1}\}, \{a_m\} \}$ et $\mathbb{P}(HG^{M_D}) = \{ \{a_1\}, \ldots, \{a_{n-1}\}, \{a_m\}, \{s\} \}$. Par conséquent, par définition de la probabilité de satisfaction (définition 3.1.6), nous avons :

$$\forall i \in [1,k[, \mathrm{P}^*(a_m,i|HG) = \mathrm{P}^*(a_m,i|HG^{M_D}) = 0$$
$$\forall i \in [k,2^{n-1}], \mathrm{P}^*(a_m,i|HG) = \mathrm{P}^*(a_m,i|HG^{M_D}) = 1$$

C'est pourquoi la manipulation destructive M_D n'est pas k-rationnelle selon la définition 3.2.5 car $\mathrm{P}^*(a_m,k|HG) = \mathrm{P}^*(a_m,k|HG^{M_D})$. □

Étudions maintenant le cas où le jeu hédonique n'a qu'une unique structure de coalitions stable.

Propriété 4.2.5 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête. Si $|NS_{HG}| = 1$ alors la manipulation destructive M_D n'est pas k-rationnelle, $\forall k \in [1,2^{n-1}]$.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ avec $N = \{a_1, \ldots, a_{n-1}, a_m\}$ tel que $NS_{HG} = \{\Pi_0\}$. Par définition du profil de préférence de l'agent s dans la manipulation destructive (définition 4.2.1) et la propriété 4.2.3, nous avons $NS_{HG^{M_D}} = \{\Pi_0 \cup \{\{s\}\}\}$. Par conséquent, par le corollaire 4.2.1, nous avons :

$$\forall i \in [1,k[, \mathrm{P}^*(a_m,i|HG) = \mathrm{P}^*(a_m,i|HG^{M_D}) = 0$$
$$\forall i \in [k,2^{n-1}], \mathrm{P}^*(a_m,i|HG) = \mathrm{P}^*(a_m,i|HG^{M_D}) = 1$$

C'est pourquoi la manipulation destructive M_D n'est pas k-rationnelle selon la définition 3.2.5 car $\mathrm{P}^*(a_m,k|HG) = \mathrm{P}^*(a_m,k|HG^{M_D})$. □

Ces deux propriétés nous permettent de montrer que mettre en œuvre une manipulation destructive sur un jeu HG n'est pas rationnelle si $|NS_{HG}| \leq 1$. Plus précisément, la manipulation destructive est k-rationnelle si et seulement si les conditions suivantes sont satisfaites :

Propriété 4.2.6 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. La manipulation destructive M_D est k-rationnelle si et seulement si :

1. $\forall i \in [1,k[, \nexists \Pi \in NS_{HG} : C_{a_m,i} \in \Pi$;
2. $\exists \Pi \in NS_{HG} : C_{a_m,k} = C^{\Pi}_{a_m}$;
3. $\exists \Pi \in NS_{HG} : C_{a_m,k} \succ_{a_m} C^{\Pi}_{a_m}$.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ tel que $|NS_{HG}| > 1$, un $k \in [1,2^{n-1}]$ et un agent malhonnête $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$.
Montrons dans un premier temps que s'il existe un $i \in [1,k[$ tel que $\exists \Pi \in NS_{HG} : C_{a_m,i} \in \Pi$, la manipulation destructive n'est pas k-rationnelle. Pour cela, supposons qu'il existe un tel i. Par définition de la probabilité de sélection, nous avons $\mathrm{P}^*(a_m,i|HG) > 1$. Par le corollaire 4.2.1,

si $NS_{HG^{M_D}} \neq \emptyset$ alors $\mathrm{P}^*(a_m,i|HG^{M_D}) = 0$. Par l'hypothèse de rationalité individuelle minimale (hypothèse 3.1.1), nous avons également $\mathrm{P}^*(a_m,i|HG^{M_D}) = 0$. C'est pourquoi dans les deux cas, la manipulation destructive M_D n'est pas k-rationnelle selon la définition 3.2.5 car $\mathrm{P}^*(a_m,i|HG^{M_D}) < \mathrm{P}^*(a_m,i|HG)$.
Montrons dans un second temps que, s'il n'existe pas de structure de coalitions $\Pi \in NS_{HG}$: $C_{a_m,k} = C^{\Pi}_{a_m}$, la manipulation destructive n'est pas k-rationnelle. Par le corollaire 4.2.2, s'il n'existe pas de structure de coalitions *partition* $\in NS_{HG} : C_{a_m,k} \sim_{a_m} C^{\Pi}_{a_m}$, alors $NS_{HGM_D} = \emptyset$. Ainsi, par l'hypothèse de rationalité individuelle minimale (hypothèse 3.1.1), nous avons $\mathbb{P}(HG^{M_D}) = \{ \{a_1\},\ldots,\{a_{n-1}\},\{a_m\},\{s\} \}$. Par conséquent, par définition de la probabilité de satisfaction (définition 3.1.6), nous avons :

$$\mathrm{P}^*(a_m,k|HG) = \mathrm{P}^*(a_m,k|HG^{M_D}) = 0$$

C'est pourquoi la manipulation destructive M_D n'est pas k-rationnelle selon la définition 3.2.5 car $\mathrm{P}^*(a_m,k|HG) = \mathrm{P}^*(a_m,k|HG^{M_D})$.
Montrons enfin que, s'il n'existe pas de structure de coalitions $\Pi \in NS_{HG} : C_{a_m,k} \succ_{a_m} C^{\Pi}_{a_m}$, la manipulation destructive n'est pas k-rationnelle. Pour cela, supposons qu'une telle structure de coalitions n'existe pas. Selon les deux conditions prouvées précédemment, nous avons : $\forall \Pi \in NS_{HG}, C_{a_m,k} = C^{\Pi}_{a_m}$. Par conséquent et par le corollaire 4.2.1, nous avons :

$$\mathrm{P}^*(a_m,k|HG) = \mathrm{P}^*(a_m,k|HG^{M_D}) = 1$$

C'est pourquoi la manipulation destructive M_D n'est pas k-rationnelle selon la définition 3.2.5 car $\mathrm{P}^*(a_m,k|HG) = \mathrm{P}^*(a_m,k|HG^{M_D})$.
Ainsi, si l'une des trois conditions n'est pas satisfaite, la manipulation destructive n'est pas k-rationnelle. □

Remarquons cependant que si la première condition n'est pas satisfaite, cela ne signifie pas pour autant que la manipulation destructive n'est pas rationnelle dans le cas général. En effet, si la première condition n'est pas satisfaite, cela signifie qu'il existe un $i < k$ (pour un k fixé) tel qu'il existe une manipulation destructive plus intéressante pour l'agent malhonnête. Ainsi, il n'est pas rationnel de vouloir intégrer la coalition $C_{a_m,k}$ alors que l'agent malhonnête pourrait intégrer $C_{a_m,i}$.

Enfin, nous pouvons déduire un corollaire à la propriété 4.2.6 qui caractérise les conditions nécessaires à la rationalité de la manipulation destructive.

Corollaire 4.2.3 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique, $a_m \in N$ un agent malhonnête. Pour tout $k \in [1,2^{n-1}]$, la manipulation destructive est k-rationnelle si et seulement si $\exists \Pi_1,\Pi_2 \in NS_{HG}$ et $\nexists \Pi_3 \in NS_{HG}$ telles que :

1. $C_{a_m,k} = C^{\Pi_1}_{a_m}$;
2. $C^{\Pi_1}_{a_m} \succ_{a_m} C^{\Pi_2}_{a_m}$;
3. $C^{\Pi_3}_{a_m} \succ_{a_m} C^{\Pi_1}_{a_m}$.

4.3 Robustesse de la stabilité au sens de Nash

Dans la section précédente, nous avons montré que, sous certaines conditions, il est rationnel pour un agent malhonnête d'effectuer une manipulation constructive ou destructive. Cependant,

nous montrons dans cette section qu'utiliser un protocole de sélection définissant la solution du jeu en la choisissant aléatoirement uniformément parmi l'ensemble des structures de coalitions stables au sens de Nash permet d'être *robuste* aux manipulations.

4.3.1 Complexité des manipulations

La robustesse d'un jeu hédonique aux manipulations est définie par l'incapacité d'un agent malhonnête à mettre en œuvre une manipulation M rationnelle.

Définition 4.3.1 - *Robustesse d'un jeu hédonique* : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique. Le jeu HG est *robuste aux manipulations* si pour tout agent $a_i \in N$, pour tout $k \in [1,2^{n-1}]$, il n'existe pas de manipulation M k-rationnelle.

Étudions dans un premier temps les conditions nécessaires à *toute forme de manipulation*. Intuitivement, nous montrons qu'il n'existe pas de manipulation dont les conditions nécessaires à la rationalité ne comportent pas celles de la manipulation constructive ou destructive.

Propriété 4.3.1 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête. Toute manipulation M mise en œuvre par a_m est k-rationnelle si et seulement si soit la manipulation constructive M_C, soit la manipulation destructive M_D est k-rationnelle.

La preuve de cette propriété repose sur des contradictions entre l'existence d'une manipulation M k-rationnelle et la non-rationalité des manipulations constructive et destructive. Pour des raisons de lisibilité, la preuve est donnée en annexe B. Intuitivement, la preuve repose sur les trois points suivants :

1. si la manipulation M est k-rationnelle alors $\mathrm{P}^*(a_m,k|HG) < 1$;
2. si $\mathrm{P}^*(a_m,k|HG) > 0$ alors la manipulation destructive est également rationnelle ;
3. si $\mathrm{P}^*(a_m,k|HG) = 0$ alors la manipulation constructive est également rationnelle.

Cette propriété signifie donc que les conditions de la rationalité de la manipulation constructive et de la manipulation destructive sont les conditions minimales nécessaires à la rationalité de toute manipulation sur un jeu hédonique. Cependant, bien qu'il existe des manipulations rationnelles qui peuvent être simples à construire dans le cas particulier des manipulations constructives et destructives, calculer si une manipulation est rationnelle est problème difficile.

Considérons le problème de décision suivant : *soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique, existe-t-il une manipulation M et un $k \in [1,2^{n-1}]$ tels que la mise en œuvre de M sur HG est k-rationnelle ?* Montrons qu'il est difficile pour un agent malhonnête de résoudre ce problème de décision. La complexité de ce problème repose sur le fait que décider si une manipulation est rationnelle nécessite de calculer l'ensemble des structures de coalitions stables au sens de Nash qui est un problème prouvé comme NP-complet [Ballester, 2004].

Propriété 4.3.2 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$. Décider de l'existence d'une manipulation M rationnelle sur HG est un problème NP-complet.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ avec $a_m \in N$ un agent malhonnête. Fixons $C_{a_m,1} \succeq_{a_m} \ldots, \succeq_{a_m} C_{a_m,2^{n-1}}$ le profil de préférence de l'agent a_m.
Par la propriété 4.3.1, il est rationnel pour un agent malhonnête a_m de manipuler HG si la manipulation constructive (définition 4.1.1) ou la manipulation destructive (définition 4.2.1) sur

HG est rationnelle. Par conséquent, décider de l'existence d'une manipulation rationnelle sur HG revient à montrer qu'il existe un $k \in [1,2^{n-1}]$ tel que soit la manipulation constructive, soit la manipulation destructive est k-rationnelle.
Par la propriété 4.1.5, a_m doit calculer la cardinalité de NS_{HG} pour décider si la manipulation constructive est k-rationnelle. Par le corollaire 4.2.3, a_m doit calculer NS_{HG} pour décider si la manipulation destructive est k-rationnelle. Ainsi, dans les deux cas, pour décider de l'existence d'un $k \in [1,2^{n-1}]$ tel que la manipulation constructive ou la manipulation destructive est k-rationnel, l'agent malhonnête doit préalablement résoudre un problème NP-complet. □

Ainsi, si certains jeux hédoniques ne sont pas robustes aux manipulations (définition 4.3.1), il reste difficile pour un agent malhonnête de décider la mise en œuvre de cette manipulation. Rappelons par ailleurs que les hypothèses de notre modèle (hypothèse 3.2.1, 3.2.3 et 3.2.4) sont très favorables aux agents malhonnêtes : il est intuitivement bien plus difficile de décider de la mise en œuvre d'une manipulation sans elles.

4.3.2 Les jeux manipulables sont rares

Si calculer la rationalité d'une manipulation sur un jeu hédonique est un problème NP-complet, cela ne signifie pas qu'en moyenne ce problème soit difficile. [Conitzer et Sandholm, 2006] ont montré l'existence d'une procédure permettant de manipuler une règle de vote lorsqu'elle satisfait l'axiome de monotonie faible et que le vote d'agents malhonnêtes en collusions peut faire gagner un candidat parmi deux. Ils montrent empiriquement que les jeux satisfaisant ces conditions sont fréquents. Ainsi, si certaines règles de votes sont dites robustes aux manipulations, car il est NP-difficile de décider d'une manipulation efficace, en pratique ce problème de décision est souvent facile.

C'est pourquoi, pour montrer que la stabilité au sens de Nash permet de garantir une robustesse aux manipulations, nous montrons ici que les jeux hédoniques satisfaisant les conditions nécessaires à la mise en œuvre d'une manipulation rationnelle sont rares. Pour cela, nous estimons empiriquement la probabilité d'existence des jeux hédoniques rationnellement manipulables par au moins un agent du système, soit par la manipulation constructive (définition 4.1.1), soit par la manipulation destructive (définition 4.2.1).

Pour cela, nous générons k jeux hédoniques HG où les profils de préférence des n agents sont définis aléatoirement uniformément. Pour chacun de ces jeux hédoniques, nous considérons tour à tour chaque agent $a_i \in N$ et calculons s'il est rationnel pour cet agent d'effectuer soit une manipulation constructive, soit une manipulation destructive. Notons que dans certains cas, tels que le jeu présenté dans les exemples 4.1.1 et 4.2.1, ces deux manipulations sont rationnelles.

Afin que la confiance en nos résultats soit suffisante, nous fixons pour nos simulations $k =$ 10 000 et considérons le pourcentage de jeux manipulables par au moins un agent. Dans nos simulations, nous faisons varier n entre 3 et 10 agents. Notons que nous ne considérons pas le cas où $n = 2$ car soit les deux agents désirent coopérer et forment la grande coalition, soit l'un des deux agents ne désire pas coopérer et ils forment les coalitions singletons.

La figure 4.4 donne le pourcentage de jeux hédoniques où la manipulation constructive ou destructive est k-rationnelle pour au moins un agent $a_i \in N$.

Il est intéressant de constater que, bien qu'en théorie la manipulation destructive est la plus efficace lorsqu'elle est k-rationnelle puisque $\mathrm{P}^*(a_m,k|HG^{M_D}) = 1$ (propriété 4.2.1), les conditions nécessaires à sa rationalité sont rarement satisfaites en pratique. Par exemple, seuls 1,71 % des jeux à 5 agents sont manipulables par une manipulation destructive. Ceci est dû aux faits que plus n est important, moins il existe de structures coalitions stables (puisque pour une structure

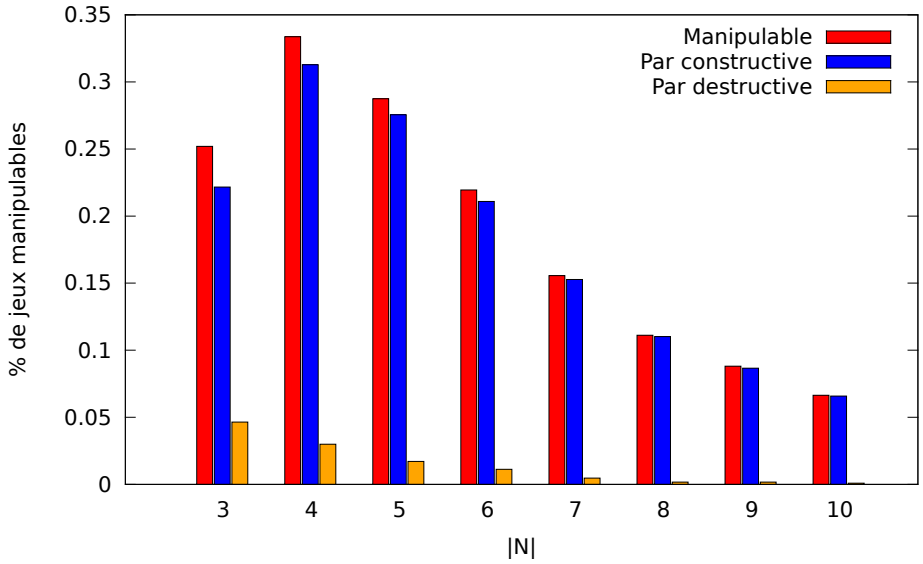

FIGURE 4.4 – Taux de jeux hédoniques manipulables en fonction du nombre d'agents

de coalitions donnée, il est fréquent qu'au moins un agent désire changer de coalition).

La manipulation constructive est plus souvent rationnelle. Par exemple, elle est rationnelle dans environ 11 % des jeux hédoniques à 8 agents. Cependant, comme pour la manipulation destructive, plus n est important, moins il existe de jeux hédoniques manipulables. Ceci s'explique par le fait que plus il y a d'agents participant aux jeux, moins il existe de structures de coalitions où un agent désirant manipuler le jeu est l'unique responsable de la non-stabilité.

Remarquons la présence d'un cas particulier. En effet, le pourcentage de jeux manipulables augmente en passant de jeux à 3 agents à des jeux à 4 agents. Cette augmentation est due au fait qu'il n'existe que 5 structures de coalitions possibles dans les jeux à 3 agents et qu'il est fréquent que la solution soit individuellement optimale pour chaque agent, et qu'aucun n'ait besoin de mettre en œuvre une manipulation.

Par ces simulations, nous montrons que le nombre de jeux hédoniques manipulables par au moins un agent est relativement faible. Ainsi, utiliser un protocole de sélection satisfaisant la stabilité au sens de Nash permet de fortement accroître la robustesse en pratique des jeux hédoniques, tout en garantissant que la solution du jeu sera rationnelle pour l'ensemble des agents participants.

4.4 Conclusion

Dans ce chapitre, nous avons étudié la robustesse des jeux hédoniques aux manipulations lorsque le protocole de sélection satisfait la *stabilité au sens de Nash*. Nous avons considéré deux manipulations : la *manipulation constructive* et la *manipulation destructive*.

La première consiste à augmenter la probabilité de satisfaction de l'agent malhonnête en augmentant artificiellement le nombre de structures de coalitions stables au sens de Nash. Pour ce faire, l'agent malhonnête définit son profil de préférence en prétendant être indifférent à toutes les coalitions possibles et introduit un seul agent Sybil présentant ses véritables préférences. Nous avons montré que cette manipulation permet de rendre stables des structures de coalitions

où l'agent malhonnête est *l'unique responsable de la non-stabilité*. La rationalité de cette manipulation dépend alors principalement des coalitions présentes dans les structures de coalitions dont l'agent malhonnête est l'unique responsable de la non-stabilité. La seconde manipulation est dite destructive, car elle permet de réduire le nombre de structures de coalitions stables qui ne sont pas satisfaisantes pour l'agent malhonnête. Pour effectuer cette manipulation, l'agent malhonnête présente son véritable profil de préférence et introduit un seul agent Sybil prétendant vouloir être dans toutes les coalitions contenant l'agent malhonnête. Nous avons montré que cette manipulation est rationnelle s'il existe au minimum deux structures de coalitions stables telles que l'une est préférée à l'autre par l'agent malhonnête.

Pour montrer la robustesse des jeux hédoniques utilisant la stabilité au sens de Nash, nous avons dans un premier temps montré que les conditions de rationalité de ces deux manipulations sont les conditions minimales à la rationalité de toute manipulation. En effet, nous avons montré que si un agent malhonnête met en œuvre une manipulation rationnelle sur un jeu hédonique alors la manipulation constructive ou la manipulation destructive sont aussi rationnelles.

Nous avons ensuite montré que décider si une de ces deux manipulations est rationnelle revient à résoudre un problème NP-complet. Enfin nous avons estimé empiriquement la proportion de jeux hédoniques où il existe au moins un agent pour qui la manipulation constructive ou la manipulation destructive est rationnelle. Nous avons montré que les conditions nécessaires à la rationalité de la manipulation destructive sont très rarement réunies. Bien que la manipulation constructive soit rationnelle dans un plus grand nombre de cas, ceux-ci restent relativement peu nombreux. Par ailleurs, pour ces deux manipulations, plus le nombre d'agents participant au jeu est important, plus il est rare que les conditions nécessaires à la rationalité de l'une des manipulations soient satisfaites.

Nous pouvons ainsi conclure que sélectionner aléatoirement uniformément la solution d'un jeu parmi l'ensemble des structures de coalitions stables au sens de Nash permet de lutter efficacement contre les manipulations. Cependant, pour montrer cette robustesse, nous avons émis certaines hypothèses favorables aux agents malhonnêtes. De plus, l'ensemble des structures stables au sens de Nash est parfois vide, ce qui limite son utilisation dans des applications pratiques. Ainsi, dans le chapitre suivant, nous poursuivons cette étude en remettant en cause l'hypothèse de bénéfice du doute et en considérant d'autres concepts de solution.

Chapitre 5

Forces et faiblesses des autres jeux

Sommaire

Résumé.
Si dans le chapitre 4 nous avons montré que la stabilité au sens de Nash est robuste aux manipulations lorsque le jeu satisfait certaines hypothèses, nous étudions dans ce chapitre la robustesse des jeux hédoniques sous d'autres conditions. Dans un premier temps, nous étudions la rationalité des manipulations *constructives* et *destructives* sans l'hypothèse du *bénéfice du doute* et considérons des hypothèses de *sous-additivité* et *sur-additivité*. Dans un second temps, nous étudions deux autres concepts de solutions canoniques qui garantissent que la solution est individuellement rationnelle : la *stabilité individuelle* et la *stabilité au sens du cœur*. La stabilité individuelle désigne l'ensemble des structures de coalitions où un agent peut refuser qu'un autre agent rejoigne sa coalition. La stabilité au sens du cœur correspond, quant à elle, à l'ensemble des structures de coalitions où il n'existe pas de sous-ensembles d'agents désirant collectivement quitter leurs coalitions respectives pour en former une autre ensemble. Pour ces deux concepts de solution, nous montrons que les jeux hédoniques dont la solution est sélectionnée aléatoirement uniformément parmi l'un de ces deux concepts de solution sont *fortement sensibles* aux manipulations.

Dans ce chapitre, les principales notations utilisées sont celles définies dans les chapitres 3 et 4 et sont résumées en l'annexe A.1.

5.1 Remise en cause du bénéfice du doute

Les manipulations *constructives* et *destructives* (définitions 4.1.1 et 4.2.1) ne sont rationnelles que sur un jeu hédonique respectant certaines conditions, et ces conditions reposent sur les hypothèses émises au chapitre 3. Dans cette section, nous étudions la robustesse des jeux hédoniques lorsque l'hypothèse de *bénéfice du doute*, intuitivement très favorable aux agents malhonnêtes, est remise en cause. Toutefois, pour qu'un agent malhonnête puisse décider de la rationalité de toute manipulation utilisant un agent Sybil, il lui faut dans tous les cas émettre des hypothèses, analogues au bénéfice du doute, permettant d'estimer ce que seront les profils de préférence des autres agents pendant la mise en œuvre de la manipulation.

5.1.1 Hypothèse de sous-additivité

Considérons dans un temps une hypothèse de sous-additivité, c'est-à-dire que les agents honnêtes vont préférer la coalition C à la coalition $C \cup \{a_j\}$ lorsque l'agent a_j qu'ils ne connaissent pas rejoint le jeu. Cela représente une forme de méfiance.

Hypothèse 5.1.1 - *Préférences sous-additives* : Un agent $a_i \in N$ a un *profil de préférence sous-additif* vis-à-vis d'un ensemble d'agents U qu'il ne connaît pas si, dans le jeu $HG' = \langle N \cup U, \succeq' \rangle$, $\succeq'_{a_i}$ est défini tel que :

$$\forall C_1, C_2 \subseteq \mathcal{C}^N_{a_i} \text{ telles que } C_1 \succ_{a_i} C_2, \forall C_3 \subseteq U : C_1 \succ'_{a_i} C_1 \cup C_3 \succ'_{a_i} C_2$$

Ceci nous permet de définir le profil de préférence des agents honnêtes.

Exemple 5.1.1 - Considérons le jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ où :

$$N = \{a_1, a_2, a_3, a_m\}$$
$$\succeq_{a_1} = \{a_1, a_2\} \succ_{a_1} \{a_1, a_3, a_m\} \succ_{a_1} \{a_1, a_3\} \succ_{a_1} \{a_1, a_2, a_m\} \succ_{a_1} \{a_1, a_2, a_3, a_m\} \succ_{a_1} \{a_1\}$$

Considérons une manipulation M mise en œuvre par a_m utilisant un agent Sybil s. Soit $HG^M = \langle N^M, \succeq^M, \mathbb{P} \rangle$ le jeu résultant de la manipulation M sur HG. Par les hypothèses d'indépendance des alternatives non-pertinentes et de sous-additivité, nous avons :

$$\begin{aligned} \succeq^M_1 = \{a_1, a_2\} &\succ^M_{a_1} \{a_1, a_2, s\} \succ^M_{a_1} \{a_1, a_3, a_m\} \succ^M_{a_1} \{a_1, a_3, a_m, s\} \\ &\succ^M_{a_1} \{a_1, a_3\} \succ^M_{a_1} \{a_1, a_3, s\} \succ^M_{a_1} \{a_1, a_2, a_m\} \succ^M_{a_1} \{a_1, a_2, a_m, s\} \\ &\succ^M_{a_1} \{a_1, a_2, a_3, a_m\} \succ^M_{a_1} \{a_1, a_2, a_3, a_m, s\} \succ^M_{a_1} \{a_1\} \succ^M_{a_1} \{a_1, s\} \end{aligned}$$

Étudions dans un premier temps l'influence de cette hypothèse sur la manipulation constructive. De manière générale, elle restreint les conditions nécessaires à la stabilité d'une structure de coalitions.

Propriété 5.1.1 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Pour toute $\Pi \in \mathcal{P}_N$, si $\exists C_0 \in \Pi$ telle que $|C_0| = 1$ alors la structure de coalitions $\Pi' = f(\Pi, s, C_0)$ n'est pas stable au sens de Nash dans HG^{M_C}.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$, une structure de coalitions $\Pi \in \mathcal{P}_N$ et une coalition $C_0 \in \Pi$ telle que $C_0 = \{a_i\}$. Soit la structure $\Pi' = f(\Pi, s, C_0)$.
Supposons dans un premier temps que $a_i = a_m$. Par définition du profil de préférence de a_m dans la manipulation constructive, $\{a_m\} \succ_{a_m}^{M_C} \{a_m, s\}$. Par construction de Π', nous avons donc $\{a_m, s\} \in \Pi'$ et, par conséquent, la structure de coalitions Π' n'est pas stable au sens de Nash dans HG^{M_C}.
Supposons maintenant que $a_i \in N \setminus \{a_m\}$. Par les hypothèses d'indépendance des alternatives non-pertinentes et de sous-additivité, nous avons $\{a_i\} \succ_{a_i}^{M_C} \{a_i, s\}$. Or, par construction de Π', $C_{a_i}^{\Pi'} = \{a_i, s\}$. Par conséquent, il existe une coalition $C \in \Pi' \cup \{\emptyset\}$ telle que $C \cup \{a_i\} \succ_{a_i}^{M_C} C_{a_i}^{\Pi'}$. La structure de coalitions Π' n'est donc pas stable au sens de Nash dans lHG^{M_C}.
Ainsi, pour toute structure de coalitions Π et toute coalition $C_0 \in \Pi$ telle que $|C_0| = 1$, la structure de coalitions $\Pi' = f(\Pi, s, C)$ n'est pas stable au sens de Nash dans HG^{M_C}. □

Intuitivement, sous l'hypothèse de sous-additivité, si l'agent Sybil désire rejoindre un agent honnête isolé, ce dernier préfère former sa coalition singleton, rendant la structure de coalitions non stable au sens de Nash. Ainsi, l'hypothèse de sous-additivité rend caduques les propriétés 4.1.1 et 4.1.3. Ces deux propriétés doivent alors se réécrire :

Propriété 5.1.2 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Soit $\Pi \in NS_{HG}$ une structure de coalitions stable au sens de Nash de HG. Sous l'hypothèse de sous-additivité, la structure de coalitions $\Pi' = f(\Pi, s, C_0)$ de HG^{M_C} est stable au sens de Nash si et seulement si les trois conditions suivantes sont satisfaites :

1. $C_0 \neq C_{a_m}^{\Pi}$;
2. $\forall C \in (\Pi \setminus \{C_{a_m}^{\Pi}\}) \cup \{\emptyset\} : C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}$;
3. $|C_0| \neq 1$.

Propriété 5.1.3 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Soit Π une structure de coalitions dont l'agent malhonnête a_m est l'unique responsable de la non-stabilité. Sous l'hypothèse de sous-additivité, la structure de coalitions $\Pi' = f(\Pi, s, C_0)$ est stable pour le jeu HG^{M_C} si et seulement si :

1. $\forall C_1 \in \Pi \cup \{\emptyset\} : C_0 \cup \{a_m\} \succeq_{a_m} C_1 \cup \{a_m\}$;
2. $|C_0| \neq 1$.

Comme il s'agit d'une réécriture des propriétés 4.1.1 et 4.1.3, leurs démonstrations reposent sur les mêmes principes et nous ne les présentons pas ici par soucis de lisibilité. Enfin, remarquons que la propriété 5.1.1 ne remet en cause la véracité de la propriété 4.1.2.

Propriété 5.1.4 : Sous l'hypothèse de sous-additivité, la propriété 4.1.2 reste vraie.

Intuitivement, cette propriété est toujours vraie, car sous l'hypothèse de sous-additivité, un agent honnête préférant changer de coalition dans une structure de coalitions préfère toujours en changer indépendamment de la coalition C_0 que rejoint l'agent Sybil.

Ainsi, l'hypothèse de sous-additivité n'a pas d'influence directe sur les conditions de la k-rationalité de la manipulation constructive, mais influe sur les valeurs de $card_{M_C}(|HG)$.

Corollaire 5.1.1 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Pour toute structure de coalitions $\Pi \in NS_{HG}$, nous avons :

$$card_{M_C}(\Pi|HG) = |\{C_0 \in \Pi | \forall C \in (\Pi \setminus \{C^{\Pi}_{a_m}\}) \cup \{\emptyset\}, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\} \wedge |C_0| \neq 1\}|$$

Pour toute structure de coalitions $\Pi \notin NS_{HG} \cup UR^{HG}_{a_m}$, nous avons :

$$card_{M_C}(\Pi|HG) = 0$$

Pour toute structure de coalitions $\Pi \in UR^{HG}_{a_m}$, nous avons :

$$card_{M_C}(\Pi|HG) = |\{C_0 \in \Pi | \forall C \in \Pi \cup \{\emptyset\}, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\} \wedge |C_0| \neq 1\}|$$

De manière intéressante, la sous-additivité permet de caractériser trivialement certaines situations où la manipulation constructive n'est pas k-rationnelle.

Propriété 5.1.5 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Sous l'hypothèse de sous-additivité, si $|C_{a_m,k}| = 2$ alors la manipulation constructive n'est pas k-rationnelle.

Étudions dans un second temps l'influence de l'hypothèse de sous-additivité sur la manipulation destructive.

Propriété 5.1.6 : Les conditions nécessaires à la k-rationalité de la manipulation destructive sur un jeu HG sont les mêmes sous l'hypothèse de sous-additivité que sous l'hypothèse de bénéfice du doute.

Pour rappel, la propriété 4.2.6 est déduite des propriétés 4.2.1, 4.2.2 et 4.2.3.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Fixons $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation destructive par $a_m \in N$ sur HG.
Montrons dans un premier temps que la propriété 4.2.1 est vraie : une structure de coalitions $\Pi' \in \mathcal{P}_{N^{M_D}}$ n'est pas stable au sens de Nash si $C^{\Pi'}_{a_m} = C_{a_m,k}$ ou $C^{\Pi'}_{s} = \{s\}$ ne sont pas vérifiées. Considérons une structure de coalitions $\Pi' \in \mathcal{P}_{N^{M_D}}$. Supposons que $C^{\Pi'}_{a_m} \neq C_{a_m,k}$. Par définition des profils de préférence de a_m et de s, soit $C^{\Pi'}_{a_m} = C^{\Pi'}_{s}$ et donc $\{a_m\} \succ^{M_D}_{a_m} C^{\Pi'}_{a_m}$, soit $C^{\Pi'}_{a_m} \neq C^{\Pi'}_{s}$ et donc $C^{\Pi'}_{a_m} \cup \{s\} \succ^{M_D}_{s} C^{\Pi'}_{s}$. Dans les deux cas, la structure de coalitions Π' n'est pas stable. Supposons maintenant que $C^{\Pi'}_{a_m} = C_{a_m,k}$ et que $C^{\Pi'}_{s} \neq \{s\}$. Par définition du profil de préférence de s, nous avons $C^{\Pi'}_{a_m} \cup \{s\} \succ^{M_D}_{s} C^{\Pi'}_{s}$ et donc Π' n'est pas stable au sens de Nash. Ainsi, la

propriété 4.2.1 est vraie sous hypothèse de sous-additivité.
Montrons dans un second temps que la propriété 4.2.2 est vraie : $\forall \Pi \notin NS_{HG}$, $\forall C_0 \in \Pi \cup \{\emptyset\}$, la structure de coalitions $\Pi' = f(\Pi,s,C_0)$ n'est pas stable. Considérons une structure de coalitions $\Pi \notin NS_{HG}$. Par la propriété 4.2.1, $\forall C_0 \in \Pi$, nous avons $\Pi' = f(\Pi,s,C_0) \notin NS_{HG^{M_D}}$. En effet, par construction de Π', nous avons $C_s^{\Pi'} \neq \{s\}$ et donc Π' n'est pas stable. Considérons maintenant la structure de coalitions $\Pi' = f(\Pi,s,\emptyset)$ et supposons que $\Pi' \in NS_{HG^{M_D}}$. Par définition de la stabilité (définition 2.1.9), nous avons $\forall a_i \in N^{M_D}$, $\nexists C \in \Pi' \cup \{\emptyset\}, C \cup \{a_i\} \succ_{a_i}^{M_D} C_{a_i}^{\Pi'}$. Comme $C_s^{\Pi'} = \{s\}$, nous avons $\forall a_i \in N, C_{a_i}^{\Pi'} = C_{a_i}^{\Pi}$. Par définition du profil de préférence de a_m, si $\nexists C \in \Pi' \cup \{\emptyset\}$ telle que $C \cup \{a_m\} \succ_{a_m}^{M_D} C_{a_m}^{\Pi'}$ alors $\nexists C \in \Pi \cup \{\emptyset\}, C \cup \{a_m\} \succ_{a_m} C_{a_m}^{\Pi}$. De même, par les hypothèses de sous-additivité et de bénéfice du doute si $\forall a_i \in N \setminus \{a_m\}$, $\nexists C \in \Pi' \cup \{\emptyset\}, C \cup \{a_i\} \succ_{a_i}^{M_D} C_{a_i}^{\Pi'}$ alors $\forall a_i \in N \setminus \{a_m\}$, $\nexists C \in \Pi \cup \{\emptyset\}, C \cup \{a_i\} \succ_{a_i} C_{a_i}^{\Pi}$. Ainsi, si $\Pi' \in NS_{HG^{M_D}}$ alors $\Pi \in NS$, ce qui est en contradiction avec $\Pi \notin NS_{HG}$. La propriété 4.2.2 est donc vraie sous hypothèse de sous-additivité.
Montrons enfin que la propriété 4.2.3 est vraie : $\forall \Pi \in NS_{HG}$ telle que $C_{a_m,k} \in \Pi$ (respectivement $C_{a_m,k} \notin \Pi$), la structure de coalitions $\Pi' = f(\Pi,s,\emptyset)$ est stable (respectivement non stable) pour le HG^{M_D}. Considérons une structure de coalitions $\Pi \in NS_{HG}$. Par construction de Π', $\forall a_i \in N$, $\forall a_i \in N, C_{a_i}^{\Pi'} = C_{a_i}^{\Pi}$. Supposons que $C_{a_m,k} \notin \Pi$. Par la propriété 4.2.1, comme $C_{a_m}^{\Pi'} \neq C_{a_m,k}$, la structure de coalitions Π' n'est pas stable. Supposons enfin que $C_{a_m,k} \in \Pi$. Par définition du profil de préférence de s, comme $C_{a_m}^{\Pi'} = C_{a_m,k}$, nous avons $\nexists C \in \Pi' \cup \{\emptyset\} : C \cup \{s\} \succ_s M_D\{s\}$. Comme $\Pi \in NS_{HG}$. Par définition des profils de préférence de a_m, nous avons également $\nexists C \in \Pi' \cup \{\emptyset\} : C \cup \{a_m\} \succ_{a_m}^{M_D} C_{a_m}^{\Pi'}$. Par l'hypothèse de sous-additivité, pour tout agent $a_i \in N \setminus \{a_m\}$, nous avons $C_{a_i}^{\Pi'} \succ_{a_m}^{M_D} \{a_i,s\}$. Comme $\Pi \in NS_{HG}$, nous avons donc $\forall a_i \in N \setminus \{a_m\}$, $\nexists C \in \Pi' \cup \{\emptyset\} : C \cup \{a_i\} \succ_{a_i}^{M_D} C_{a_i}^{\Pi'}$. Ainsi, si $C_{a_m}^{\Pi'} = C_{a_m,k}$ alors la structure de coalitions Π' est stable. Par conséquent, la propriété 4.2.3 est vraie sous hypothèse de sous-additivité. □

Ainsi, la relaxation de l'hypothèse de bénéfice du doute par l'hypothèse de sous-additivité renforce la robustesse des jeux hédoniques face à la manipulation constructive mais pas face à la manipulation destructive.

5.1.2 Hypothèse de sur-additivité

Considérons maintenant une autre relaxation intuitive du bénéfice du doute : une hypothèse de sur-additivité, correspondant au fait que les agents honnêtes vont préférer la coalition $C \cup \{a_j\}$ à la coalition C lorsqu'un agent a_j qu'ils ne connaissent pas rejoint le jeu. Cette hypothèse représente une préférence à la présence d'agents inconnus dans les coalitions.

Hypothèse 5.1.2 - *Préférences sur-additives* : Un agent $a_i \in N$ a un *profil de préférence sur-additif* vis-à-vis d'un ensemble d'agents U qu'il ne connaît pas si, dans le jeu $HG^{'} = \langle N \cup U, \succeq^{'} \rangle$, $\succeq_{a_i}^{'}$ est défini tel que :

$$\forall C_1,C_2 \subseteq \mathcal{C}_{a_i}^N \text{ telles que } C_1 \succ_{a_i} C_2, \forall C_3 \subseteq U : C_1 \cup C_3 \succ_{a_i}^{'} C_1 \succ_{a_i}^{'} C_2$$

Intuitivement cette hypothèse semble être avantageuse pour les agents malhonnêtes utilisant au moins un agent Sybil puisque la présence de ce dernier dans une coalition est préférable à son absence. Cependant, cette intuition est fausse. En effet, l'hypothèse de sur-additivité n'est pas à l'avantage d'un agent malhonnête puisqu'elle réduit le nombre de structures stables lors de la mise en œuvre de la manipulation.

Propriété 5.1.7 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique, $a_m \in N$ un agent malhonnête et $HG^M = \langle N^M, \succeq^M, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation M utilisant un ensemble $S = \{s_1, \ldots, s_l\}$ (non vide) d'agents Sybil. Soit $\Pi' \in \mathcal{P}_{N^M}$ une structure de coalitions. Si $\exists a_i \in N \setminus \{a_m\}$ tel que $C^{\Pi'}_{a_i} = \{a_i\}$ et que $\exists s_j \in S$ tel que $C^{\Pi'}_{s_j} = \{s_j\}$ alors, sous l'hypothèse de sur-additivité, la structure de coalitions Π' n'est pas stable au sens de Nash.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$, $a_m \in N$ un agent malhonnête et $HG^M = \langle N^M, \succeq^M, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation M utilisant un ensemble $S = \{s_1, \ldots, s_l\}$ (non vide) d'agents Sybil. Fixons une structure de coalitions $\Pi' \in \mathcal{P}_{N^M}$ telle qu'il existe $a_i \in N \setminus \{a_m\} : C^{\Pi'}_{a_i} = \{a_i\}$ et qu'il existe $s_j \in S : C^{\Pi'}_{s_j} = \{s_i\}$. Par l'hypothèse de sur-additivité, nous avons $\{a_i, s_j\} \succ^M_{a_m} \{a_i\}$. Par conséquent, la structure de coalitions Π' n'est pas stable au sens de Nash. □

Cette propriété indique qu'une structure de coalitions contenant un agent honnête et un agent Sybil dans leur coalition singleton respective ne peut pas être stable au sens de Nash, car l'agent honnête préfère toujours rejoindre l'agent Sybil. Ainsi, l'hypothèse de sur-additivité rend caduque la propriété 4.1.1. Cette dernière doit donc être réécrite :

Propriété 5.1.8 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Soit $\Pi \in NS_{HG}$ une structure de coalitions stable au sens de Nash de HG. Sous l'hypothèse de sur-additivité, $\forall C_0 \in \Pi \cup \{\emptyset\}$, la structure de coalitions $\Pi' = f(\Pi, s, C_0)$ de HG^{M_C} est stable si et seulement si les trois conditions suivantes sont satisfaites :

1. $C_0 \neq C^{\Pi}_{a_m}$;
2. $\forall C \in (\Pi \setminus \{C^{\Pi}_{a_m}\}) \cup \{\emptyset\} : C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}$;
3. $C_0 \neq \emptyset \vee \nexists a_i \in N \setminus \{a_m\} : C^{\Pi}_{a_i} = \{a_i\}$.

Remarquons que la propriété 5.1.7 ne remet en cause la véracité des propriétés 4.1.2 et 4.1.3.

Propriété 5.1.9 : Sous l'hypothèse de sur-additivité, les propriétés 4.1.2 et 4.1.3 restent vraies.

Intuitivement, la propriété 4.1.2 est toujours vraie, car, sous hypothèse de sur-additivité, un agent honnête préférant changer de coalition dans une structure de coalitions préfère toujours en changer indépendamment de la coalition C_0 que rejoint l'agent Sybil. Quant à la propriété 4.1.3, elle est incluse implicitement dans la propriété 5.1.7. Ainsi, l'hypothèse de sur-addivité n'a pas d'influence directe sur les conditions de la k-rationalité de la manipulation constructive, mais influe sur les valeurs de $card_{M_C}(\Pi | HG)$.

Corollaire 5.1.2 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation constructive par $a_m \in N$ sur HG. Sous hypothèse de sur-additivité, nous avons :

1. pour toute structure $\Pi \in NS_{HG}$:

$$card_{M_C}(\Pi | HG) = |\{C_0 \in \Pi | \forall C \in \Pi \cup \{\emptyset\}, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\} \wedge (\, C_0 \neq \emptyset \;\vee\; \nexists a_i \in N \setminus \{a_m\} : C^{\Pi}_{a_i} = \{a_i\}\,)\}|$$

2. pour toute structure $\Pi \notin NS_{HG} \cup UR_{a_m}^{HG}$:

$$card_{M_C}(\Pi|HG) = 0$$

3. pour toute structure $\Pi \in UR_{a_m}^{HG}$:

$$card_{M_C}(\Pi|HG) = |\{C_0 \in \Pi | \forall C \in (\Pi \setminus \{C_{a_m}^{\Pi}\}) \cup \{\emptyset\}, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}\}|$$

Étudions dans un second temps l'influence de l'hypothèse de sur-additivité sur la manipulation destructive. La caractérisation de cette manipulation repose sur les propriétés 4.2.1, 4.2.2 et 4.2.3. Trivialement, les deux premières propriétés sont toujours vraies sous l'hypothèse de sur-additivité. Ce n'est pas le cas pour la propriété 4.2.3 qui doit être reformulée.

Propriété 5.1.10 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation destructive par $a_m \in N$ sur HG. Sous hypothèse de sur-addivité, pour toute structure de coalitions $\Pi \in NS_{HG}$, la structure $\Pi' = f(\Pi, s, \emptyset)$ est stable dans HG^{M_D} si et seulement si $C_{a_m,k} \in \Pi$ et $\nexists a_i \in N \setminus \{a_m\}$ tel que $C_{a_i}^{\Pi} = \{a_i\}$.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Fixons $HG^{M_D} = \langle N^{M_D}, \succeq^{M_D}, \mathbb{P} \rangle$ le jeu résultant de la mise en œuvre de la manipulation destructive par $a_m \in N$ sur HG. Fixons les structures de coalitions $\Pi \in NS_{HG}$ et $\Pi' = f(\Pi, s, \emptyset)$. Comme par construction de Π', nous avons $\forall a_i \in N, C_{a_i}^{\Pi} = C_{a_i}^{\Pi'}$ si $C_{a_m,k} \notin \Pi$. Donc, nous avons $C_{a_m,k} \notin \Pi'$. Par la propriété 4.2.1, la structure de coalitions Π' n'est donc pas stable.
Suppons maintenant que $, C_{a_m,k} \in \Pi$ mais qu'il existe un agent $a_i \in N \setminus \{a_m\}$ tel que $C_{a_i}^{\Pi} = \{a_i\}$. Par l'hypothèse de sur-additivité, nous avons $\{a_i, s\} \succ_{a_i}^{M_D} \{a_i\}$. Comme par construction nous avons $\{s\} \in \Pi'$, la structure de coalitions Π' n'est pas stable au sens de Nash. □

Ainsi, sous hypothèse de sur-additivité, les conditions nécessaires à la rationalité de la manipulation destructive sont :

Propriété 5.1.11 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Sous l'hypothèse de sur-additivité, la manipulation destructive M_D est k-rationnelle si et seulement si :

1. $\forall i \in [1, k[, \nexists \Pi \in NS_{HG} : C_{a_m,i} \in \Pi$;
2. $\exists \Pi \in NS_{HG}$ telle que :
   (a) $C_{a_m,k} = C_{a_m}^{\Pi}$;
   (b) $\nexists a_i \in N \setminus \{a_m\}$ tel que $C_{a_i}^{\Pi} = \{a_i\}$;
3. $\exists \Pi \in NS_{HG} : C_{a_m,k} \succ_{a_m} C_{a_m}^{\Pi}$.

Contre toute attente, l'hypothèse de sur-additivité n'est pas plus avantageuse pour l'agent malhonnête que l'hypothèse de bénéfice du doute. Au contraire, cette hypothèse rend plus restrictives les conditions nécessaires à la rationalité des manipulations constructive et destructive.

5.2 Robustesse des autres concepts de solution

Dans le chapitre 4, nous avons étudié la robustesse des jeux hédoniques lorsque le protocole de sélection satisfait la stabilité au sens de Nash. Cependant, il existe d'autres concepts de solution satisfaisant la rationalité individuelle (définition 2.1.13). Dans cette section, nous nous intéressons à deux de ces concepts : la *stabilité individuelle* (définition 2.1.10) et la *stabilité au sens du cœur* (définition 2.1.12). Comme dans le chapitre précédent, nous supposons que les hypothèses 3.2.3 (indépendance des alternatives non-pertinentes), 3.2.4 (bénéfice du doute), 3.2.2 (unicité de l'agent malhonnête), 3.2.1 (connaissance du jeu initiale) et 3.1.1 (satisfaction minimale de la rationalité individuelle) sont respectées.

5.2.1 La manipulation destructive n'est plus rationnelle

Nous l'avons montré dans la section 4.2, la manipulation destructive permet de réduire le nombre de structures de coalitions stables au sens de Nash afin de garantir que la solution d'un jeu soit satisfaisante pour un agent malhonnête. Cependant, cette même manipulation n'est pas rationnelle pour d'autres concepts de solution tels que la stabilité du cœur et la stabilité individuelle.

Dans le cadre de la stabilité individuelle, le fait que la manipulation destructive soit non rationnelle découle directement de la définition de ce concept de solution. En effet, pour qu'une structure de coalitions ne soit pas individuellement stable, il suffit qu'un agent préfère changer de coalition et qu'il soit accepté par les agents qu'il désire rejoindre. Or, l'efficacité de la manipulation destructive repose sur le fait que l'agent Sybil s désire rejoindre la coalition de a_m et que ce dernier l'y refuse, rendant une structure de coalitions stable au sens de Nash dans HG non stable dans HG^{M_D}. Ainsi, la stabilité individuelle ne permet plus à l'agent Sybil s de rendre instables des structures de coalitions.

Propriété 5.2.1 : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ tel que $\mathbb{P}$ retourne la solution du jeu HG aléatoirement uniformément parmi IS_{HG}, l'ensemble des structures de coalitions individuellement stables dans HG. La manipulation destructive M_D n'est pas k-rationnelle pour tout $k \in [1,2^{n-1}]$.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et un agent malhonnête $a_m \in N$ ayant le profil de préférence suivant : $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$.
Montrons dans un premier temps que pour toute structure de coalitions $\Pi \notin IS_{HG}$ et toute $C_0 \in \Pi \cup \{\emptyset\}$, la structure de coalitions $\Pi' = f(\Pi, s, C_0)$ n'est pas individuellement stable dans HG^{M_D}. Rappelons que par construction de Π', $\forall a_i \in N, C_{a_i}^{\Pi} = C_{a_i}^{\Pi'} \setminus \{s\}$. Par définition de la stabilité individuelle (définition 2.1.10), pour toute structure de coalitions $\Pi \notin IS_{HG}$, il existe $a_i \in N$ et $C \in \Pi \cup \{\emptyset\}$ tels que $C \cup \{a_i\} \succ_{a_i} C_{a_i}^{\Pi}$ et que $\forall a_j \in C, C \cup \{a_i\} \succeq_{a_j} C$.

1. Supposons $C_0 = \emptyset$. Par construction du profil de préférence de a_m dans la manipulation destructive et les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute (hypothèses 3.2.3 et 3.2.4), $\exists C \in \Pi' : C \cup \{a_i\} \succ_{a_i}^{M_D} C_{a_i}^{\Pi'} \wedge \forall a_j \in C, C \cup \{a_i\} \succeq_{a_j}^{M_D} C$. Par conséquent, la structure de coalitions Π' n'est pas individuellement stable.
2. Supposons $C_0 = C_{a_m}^{\Pi}$. Par définition du profil de préférence de a_m dans la manipulation destructive, nous avons $\{a_m\} \succ_{a_m}^{M_D} C_{a_m}^{\Pi'}$ et donc la structure de coalitions Π' n'est pas individuellement stable.

3. Supposons enfin $C_0 \in \Pi \setminus \{C^{\Pi}_{a_m}\}$. Par définition du profil de préférence de s dans la manipulation destructive, nous avons $\{s\} \succ^{M_D}_s C^{\Pi'}_s$ et donc la structure de coalitions Π' n'est pas individuellement stable.

Montrons maintenant que, pour toute structure de coalitions $\Pi \in IS_{HG}$ et toute coalition $C_0 \in \Pi$, la structure de coalitions $\Pi' = f(\Pi,s,C_0)$ n'est pas individuellement stable dans HG^{M_D}. Nous avons ici deux cas possibles : $C_0 = C^{\Pi}_{a_m}$ et $C_0 \in \Pi \setminus \{C^{\Pi}_{a_m}\}$. Comme précédemment, si $C_0 = C^{\Pi}_{a_m}$, l'agent a_i préfère sa coalition singleton, et si $C_0 \in \Pi \setminus \{C^{\Pi}_{a_m}\}$, l'agent Sybil préfère sa coalition singleton. Dans les deux cas, la structure de coalitions Π' n'est pas individuellement stable.
Montrons enfin que, pour toute structure de coalitions $\Pi \in IS_{HG}$, la structure de coalitions $\Pi' = f(\Pi,s,\emptyset)$ est individuellement stable. Comme Π est individuellement stable, $\forall a_i \in N$, $\nexists C \in \Pi \cup \{\emptyset\}$ telle que : $C \cup \{a_i\} \succ_{a_i} C^{\Pi}_{a_i}$ et $\forall a_j \in C, C \cup \{a_i\} \succeq_{a_j} C$. Par construction du profil de préférence de a_m dans la manipulation destructive, par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute (hypothèses 3.2.3 et 3.2.4) et comme $\forall C_1 \in \Pi$, nous avons également $C_1 \in \Pi'$, $\forall a_i \in N$, $\nexists C \in \Pi' \cup \{\emptyset\}$ telle que : $C \cup \{a_i\} \succ^{M_D}_{a_i} C^{\Pi'}_{a_i}$ et $\forall a_j \in C, C \cup \{a_i\} \succeq^{M_D}_{a_j} C$.
Si cette dernière propriété est également vraie pour s alors Π' est individuellement stable. Par construction de $\succeq^{M_D}_s$, si $C^{\Pi'}_{a_m} = C_{a_m,k}$ alors $\nexists C \in \Pi' : C \cup \{s\} \succ^{M_D}_s \{s\}$. Par conséquent Π' est bien individuellement stable. Si $C^{\Pi'}_{a_m} \neq C_{a_m,k}$, nous avons $C^{\Pi'}_{a_m} \cup \{s\} \succ^{M_D}_s \{s\}$. Cependant, la partition Π' est individuellement stable, car par définition des préférences de a_m, nous avons $C^{\Pi'}_{a_m} \succ^{M_D}_{a_m} C^{\Pi'}_{a_m} \cup \{s\}$.
Nous pouvons enfin déduire de ces trois points que $|IS_{HG}| = |IS_{HG^{M_D}}|$ avec $\forall \Pi \in IS_{HG}$, $f(\Pi,s,\emptyset) \in IS_{HG^{M_D}}$. Par conséquent, $\forall k \in [1,2^{n-1}]$, $\mathrm{P}^*(a_m,k|HG) = \mathrm{P}^*(a_m,k|HG^{M_D})$ et la manipulation destructive ne peut donc pas être k-rationnelle. □

Dans le même principe, la manipulation destructive n'est pas rationnelle lorsque le protocole de sélection satisfait la stabilité au sens du cœur. Intuitivement, l'introduction de la coalition $\{s\}$ dans une structure stable au sens du cœur ne change pas le fait qu'aucun sous-ensemble d'agent ne souhaite collectivement former une nouvelle coalition. La manipulation destructive ne permet donc pas de rendre instable une structure de coalitions appartenant à CS_{HG}.

Propriété 5.2.2 : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$ tel que $\mathbb{P}$ retourne la solution du jeu HG aléatoirement uniformément parmi CS_{HG}, l'ensemble des structures de coalitions stables au sens du cœur. La manipulation destructive M_D n'est pas k-rationnelle pour tout $k \in [1,2^{n-1}]$.

La démonstration de cette propriété est similaire à celle de la propriété 5.2.1.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$ et un agent malhonnête $a_m \in N$ ayant le profil de préférence suivant : $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Fixons les structures de coalitions $\Pi \in \mathcal{P}_N$, une coalition $C_0 \in \Pi \cup \{\emptyset\}$ et $\Pi' = f(\Pi,s,C_0)$.
Supposons dans un premier temps que $C_0 \neq \emptyset$. Par définition des préférences de a_m et de s dans la manipulation destructive, si $C_0 \neq C^{\Pi}_{a_m}$ alors $\{s\} \succ^{M_D}_s C_0 \cup \{s\}$, et si $C_0 = C^{\Pi}_{a_m}$ alors $\{a_m\} \succ^{M_D}_{a_m} C_0 \cup \{s\}$. Dans les deux cas, Π' n'est pas stable au sens du cœur.
Par conséquent, pour que Π' soit stable au sens du cœur, il faut que $C_0 = \emptyset$. Dans ce cas, nous avons $\forall a_i \in N, C^{\Pi}_{a_i} = C^{\Pi'}_{a_i}$. Par l'hypothèse d'indépendance des alternatives non-pertinentes (hypothèses 3.2.3), s'il existe $N_2 \subseteq N$ tel que $\forall a_i \in N_2, N_2 \succ_{a_i} C^{\Pi}_{a_i}$ alors $\forall a_i \in N_2, N_2 \succ^{M_D}_{a_i} C^{\Pi'}_{a_i}$ et donc $\Pi' \notin CS_{HG^{M_D}}$. Inversement, s'il n'existe pas un tel groupe N_2 alors la structure de coalitions $\Pi' = f(\Pi,s,\emptyset)$ est stable au sens du cœur.

Comme pour la stabilité individuelle, $\forall k \in [1,2^{n-1}]$, $\mathrm{P}^*(a_m,k|HG) = \mathrm{P}^*(a_m,k|HG^{M_D})$ et la manipulation destructive n'est donc pas k-rationnelle. □

Ainsi, les concepts de stabilité individuelle et de stabilité au sens du cœur ne permettent pas la rationalité de la manipulation destructive.

5.2.2 La manipulation constructive reste rationnelle

Si la manipulation destructive n'est pas rationnelle dans le cadre de la stabilité individuelle et de la stabilité au sens du cœur, ces deux concepts de solution sont tout de même sensibles à la manipulation constructive. Dans le cadre de la stabilité individuelle, le fait que la manipulation constructive soit rationnelle découle des hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute. Intuitivement, pour toute structure de coalitions individuellement stable contenant une coalition que veut rejoindre a_m, l'agent Sybil s peut rejoindre cette coalition. Il en est de même pour les structures de coalitions où l'agent malhonnête est l'unique responsable de la non-stabilité.

Propriété 5.2.3 : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$ tel que $\mathbb{P}$ retourne la solution du jeu HG aléatoirement uniformément parmi IS_{HG}, l'ensemble des structures de coalitions individuellement stables dans HG. Soit $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Soit $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P}\rangle$ le jeu résultant de la manipulation constructive mis en œuvre par a_m. Soit $\Pi \in IS_{HG}$ une structure de coalitions individuellement stable dans le jeu HG. Pour toute coalition $C_0 \in \Pi \cup \{\emptyset\}$, la structure de coalitions $\Pi' = f(\Pi,s,C_0)$ est individuellement stable dans le jeu $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P}\rangle$ si et seulement si les trois conditions suivantes sont satisfaites :

1. $C_0 \neq C^{\Pi}_{a_m}$;
2. $\nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$;
3. $\nexists a_i \in N \setminus \{a_m\}$ tel que $C^{\Pi}_{a_m} \cup \{a_i\} \succ_{a_i} C^{\Pi}_{a_i}$ et $\forall a_j \in C^{\Pi}_{a_m} \setminus \{a_m\}, C^{\Pi}_{a_m} \cup \{a_i\} \succeq_{a_j} C^{\Pi}_{a_m}$.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$, un agent malhonnête $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$, une structure de coalitions $\Pi \in CS_{HG}$ et une coalition $C_0 \in \Pi \cup \{\emptyset\}$. Soit la structure de coalitions Π' telle que $\Pi' = f(\Pi,s,C_0)$.
Montrons dans un premier temps que si $C_0 = C^{\Pi}_{a_m}$, la structure de coalitions Π' n'est pas individuellement stable dans HG^{M_C}. Supposons que $C_0 = C^{\Pi}_{a_m}$. Par définition du profil de préférence de s, nous avons $\{s\} \succ^{M_C}_s C^{\Pi}_{a_m} \cup \{s\}$. Ainsi, si $C_0 = C^{\Pi}_{a_m}$ alors la structure Π' n'est pas individuellement stable dans HG^{M_C}.
Montrons ensuite que si $\exists C \in \Pi \cup \{\emptyset\} : C \cup \{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$, la structure Π' n'est pas individuellement stable dans le jeu HG^{M_C}. Supposons qu'il existe une telle coalition $C \in \Pi \cup \{\emptyset\}$. Par définition de profil de préférence de s, nous avons $C \cup \{s\} \succ^{M_C}_s C_0 \cup \{s\}$. Par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute, nous avons $\forall a_j \in C, C \sim^{M_C}_{a_j} C \cup \{s\}$. Par conséquent, la structure Π' n'est pas individuellement stable.
Montrons maintenant que la troisième condition est nécessaire pour que la structure de coalitions Π' soit individuellement stable. Spposons qu'il existe $a_i \in N \setminus \{a_m\}$ tel que $C^{\Pi}_{a_m} \cup \{a_i\} \succ_{a_i} C^{\Pi}_{a_i}$ et que $\forall a_j \in C^{\Pi}_{a_m} \setminus \{a_m\}, C^{\Pi}_{a_m} \cup \{a_i\} \succeq_{a_j} C^{\Pi}_{a_m}$. Comme $\Pi \in IS_{HG}$, nous avons $C^{\Pi}_{a_m} \succeq_{a_m} C^{\Pi}_{a_m} \cup \{a_i\}$. Par l'hypothèse d'indépendance des alternatives non-pertinentes, nous avons alors $a_i \in N \setminus \{a_m\}$ tel que $C^{\Pi'}_{a_m} \cup \{a_i\} \succ^{M_C}_{a_i} C^{\Pi'}_{a_i}$ et que $\forall a_j \in C^{\Pi'}_{a_m} \setminus \{a_m\}, C^{\Pi'}_{a_m} \cup \{a_i\} \succeq^{M_C}_{a_j} C^{\Pi'}_{a_m}$. Par définition du

profil de préférence de a_m, nous avons $C^{\Pi'}_{a_m} \sim^{M_C}_{a_m} C^{\Pi'}_{a_m} \cup \{a_i\}$. Par conséquent, la structure Π' n'est pas individuellement stable.
Montrons enfin que si les conditions (1), (2) et (3) sont satisfaites, la structure de coalitions Π' est individuellement stable dans HG^{M_C}. Supposons que la structure de coalitions Π vérifie $\nexists a_i \in N \setminus \{a_m\}$ tel que $C^{\Pi}_{a_m} \cup \{a_i\} \succ_{a_i} C^{\Pi}_{a_i}$ et $\forall a_j \in C^{\Pi}_{a_m} \setminus \{a_m\}$,$C^{\Pi}_{a_m} \cup \{a_i\} \succeq_{a_j} C^{\Pi}_{a_m}$ et fixons la structure de coalitions C_0 telle que $C_0 \neq C^{\Pi}_{a_m}$ et que $\nexists C \in \Pi \cup \{\emptyset\}$ telle que $C \cup \{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$. Comme $C_0 \neq C^{\Pi}_{a_m}$, par construction de Π', nous avons $C^{\Pi}_{a_m} = C^{\Pi'}_{a_m}$. Par définition du profil de préférence de a_m, nous avons $\nexists C \in \Pi' \cup \{\emptyset\} : C \cup \{a_m\} \succ^{M_C}_{a_m} C^{\Pi'}_{a_m}$. Comme $\nexists C \in \Pi \cup \{\emptyset\}$ telle que $C \cup \{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$, par définition du profil de préférence de s, nous avons également $\nexists C \in \Pi' \cup \{\emptyset\}$ telle que $C \cup \{a_m\} \succ^{M_C}_{s} C^{\Pi'}_{a_m}$. Par conséquent, ni a_m, ni s ne désirent changer de coalition dans Π'.
Comme $\Pi \in IS_{HG}$, nous avons $\forall a_i \in N$, $\nexists C \in \Pi \cup \{\emptyset\}$ telle que $C \cup \{a_i\} \succ_{a_i} C^{\Pi}_{a_i}$ et $\forall a_j \in C$,$C \cup \{a_i\} \succeq_{a_j} C^{\Pi}_{a_j}$. Par les hypothèses d'indépendance des alternatives non-pertinentes, de bénéfice du doute et par définition du profil de préférence de a_m, nous avons donc $\forall a_i \in N$, $\nexists C \in \Pi' \cup \{\emptyset\}$ telle que $C \cup \{a_i\} \succ^{M_C}_{a_i} C^{\Pi'}_{a_i}$ et $\forall a_j \in C$,$C \cup \{a_i\} \succeq^{M_C}_{a_j} C$. Par conséquent, tout agent honnête désirant changer de coalition dans Π' sera rejeté.
Ainsi, si les conditions (1), (2) et (3) sont satisfaites, la structure de coalitions Π' est individuellement stable dans HG^{M_C}. □

Nous pouvons alors en déduire le nombre de structures de coalitions individuellement stables dans $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P}\rangle$ construites à partir d'une structure $\Pi \in IS_{HG}$. Si la troisième condition qui porte sur les préférences des agents, et donc sur les structures stables, n'est pas satisfaite, il n'existe pas de structure stable. Sinon, il en existe autant que de coalitions $C_0 \in \Pi \cup \{\emptyset\}$ vérifiant les deux premières conditions.

Corollaire 5.2.1 : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$ et $a_m \in N$ agent malhonnête. Pour toute structure de coalitions $\Pi \in IS_{HG}$, si $\exists a_i \in N \setminus \{a_m\}$ tel que $C^{\Pi}_{a_m} \cup \{a_i\} \succ_{a_i} C^{\Pi}_{a_i}$ et $\forall a_j \in C^{\Pi}_{a_m} \setminus \{a_m\}$,$C^{\Pi}_{a_m} \cup \{a_i\} \succeq_{a_j} C^{\Pi}_{a_m}$ alors $card_{M_C}(\Pi|HG) = 0$. Sinon $card_{M_C}(\Pi|HG) = |\{C_0 \in \Pi \cup \{\emptyset\} \setminus C^{\Pi}_{a_m} |\ \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_m\} \succ_{a_m} C_0 \cup \{a_m\}\}|$.

Étudions dans un second temps l'influence de la manipulation constructive sur les structures de coalitions qui ne sont pas individuellement stables. Intuitivement, l'ajout de l'agent s dans l'une des coalitions d'une structure non stable la rend stable uniquement si l'agent malhonnête est l'unique responsable de la non-stabilité de Π.

Propriété 5.2.4 : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$ tel que $\mathbb{P}$ retourne la solution du jeu HG aléatoirement uniformément parmi IS_{HG}, l'ensemble des structures de coalitions individuellement stables dans HG. Soit $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \dots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \dots \succeq_{a_m} C_{a_m,2^{n-1}}$. Soit $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P}\rangle$ le jeu résultant de la manipulation constructive mis en œuvre par a_m. Soit $\Pi \notin IS_{HG}$ une structure de coalitions individuellement stable. Pour toute coalition $C_0 \in \Pi \cup \{\emptyset\}$, la structure de coalitions $\Pi' = f(\Pi,s,C_0)$ est individuellement stable dans le jeu $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P}\rangle$ si et seulement si les trois conditions suivantes sont satisfaites :

1. $C_0 \neq C^{\Pi}_{a_m}$;
2. $\nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$;
3. $\Pi \in UR^{HG}_{a_m}$.

La démonstration des deux premières conditions est identique à celle de la démonstration 5.2.2. Pour des raisons de lisibilité, nous ne présentons ici que la démonstration de la troisième condition.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$, un agent malhonnête $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$, une structure de coalitions $\Pi \notin CS_{HG}$ et une coalition $C_0 \in \Pi \cup \{\emptyset\}$. Soit la structure de coalitions Π' telle que $\Pi' = f(\Pi, s, C_0)$. Supposons que les conditions (1) et (2) sont satisfaites et que $\Pi' \in IS_{HG^{M_C}}$.
Par définition de la stabilité individuelle, nous avons $\forall a_i \in N \setminus \{a_m\}$, $\nexists C \in \Pi'$ telle que $C \cup \{a_i\} \succ_{a_i}^{M_C} C_{a_i}^{\Pi'}$ et $\forall a_j \in C, C \cup \{a_i\} \succeq_{a_i}^{M_C} C$. Par construction de Π', pour toute coalition $C \in \Pi'$, soit $C \in \Pi$, soit $C \setminus \{s\} \in \Pi$. Par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute, nous avons $\forall a_i \in N \setminus \{a_m\}$, $\nexists C \in \Pi$ telle que $C \cup \{a_i\} \succ_{a_i} C_{a_i}^{\Pi}$ et $\forall a_j \in C, C \cup \{a_i\} \succeq_{a_i} C$. Or, $\Pi \notin IS_{HG}$. Comme il n'existe pas d'agent $a_i \in N \setminus \{a_m\}$ souhaitant changer de coalition pour un autre où il est accepté, nous avons $\Pi \in UR_{a_m}^{HG}$.
Montrons enfin que si les conditions (1), (2) et (3) sont satisfaites, la structure de coalitions Π' est individuellement stable dans HG^{M_C}. Supposons $\Pi \in UR_{a_m}^{HG}$ et fixons $C_0 \neq C_{a_m}^{\Pi}$ où $\nexists C \in \Pi \cup \{\emptyset\}$ telle que $C \cup \{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$. Comme $C_0 \neq C_{a_m}^{\Pi}$, par construction de Π', nous avons $C_{a_m}^{\Pi} = C_{a_m}^{\Pi'}$. Par définition du profil de préférence de a_m, nous avons $\nexists C \in \Pi' \cup \{\emptyset\}$: $C \cup \{a_m\} \succ_{a_m}^{M_C} C_{a_m}^{\Pi'}$. Comme $\nexists C \in \Pi \cup \{\emptyset\}$ telle que $C \cup \{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$, par définition du profil de préférence de s, nous avons également $\nexists C \in \Pi' \cup \{\emptyset\}$ telle que $C \cup \{a_m\} \succ_s^{M_C} C_{a_m}^{\Pi'}$. Par conséquent, ni a_m, ni s ne désirent changer de coalition dans Π'.
Comme $\Pi \in UR_{a_m}^{HG}$, nous avons $\forall a_i \in N \setminus \{a_m\}$, $\nexists C \in \Pi \cup \{\emptyset\}$ tel que $C \cup \{a_i\} \succ_{a_i} C_{a_i}^{\Pi}$ et $\forall a_j \in C, C \cup \{a_i\} \succeq_{a_j} C_{a_j}^{\Pi}$. Par les hypothèses d'indépendance des alternatives non-pertinentes, de bénéfice du doute et par définition du profil de préférence de a_m, nous avons donc $\forall a_i \in N$, / $\exists C \in \Pi' \cup \{\emptyset\}$ telle que $C \cup \{a_i\} \succ_{a_i}^{M_C} C_{a_i}^{\Pi'}$ et $\forall a_j \in C, C \cup \{a_i\} \succeq_{a_j}^{M_C} C$. Par conséquent, tout agent honnête désirant changer de coalition dans Π' sera rejeté.
Par conséquent, si les conditions (1), (2) et (3) sont satisfaites alors Π' est individuellement stable. □

Nous pouvons aussi en déduire le nombre de structures de coalitions individuellement stables dans $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ construites à partir d'une structure de coalitions $\Pi \notin IS_{HG}$.

Corollaire 5.2.2 : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et $a_m \in N$ agent malhonnête. Pour toute structure de coalitions $\Pi \notin IS_{HG}$, si $\Pi \notin UR_{a_m}^{HG}$ alors $card_{M_C}(\Pi | HG) = 0$. Sinon $card_{M_C}(\Pi | HG) = |\{C_0 \in \Pi \cup \{\emptyset\} \setminus C_{a_m}^{\Pi} | \ \nexists C \in \Pi \cup \{\emptyset\} : C \cup \{a_m\} \succ_{a_m} C_0 \cup \{a_m\}\}|$.

À partir des corollaires 5.2.1 et 5.2.2, nous pouvons déduire les conditions à la rationalité de la manipulation constructive sur les jeux utilisant la stabilité individuelle comme concept de solution.

Propriété 5.2.5 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique tel que $\mathbb{P}$ retourne la solution du jeu HG aléatoirement uniformément parmi IS_{HG}, l'ensemble des structures de coalitions individuellement stables dans HG et $a_m \in N$ un agent malhonnête ayant le profil de préférence $\succeq_{a_m} = C_{a_m,1} \succeq_{a_m} C_{a_m,2} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Soit $(1)_i$ le nombre de structures de coalitions $\Pi \in IS_{HG}$ telles que :

$$C_{a_m,i} \sim_{a_m} C_{a_m}^{\Pi} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$$
$$\text{et} \quad \nexists C \in \Pi : C \cup \{a_m\} \succ_{a_m} C_{a_m,i}$$

Soit $(2)_i$ le nombre de structures de coalitions $\Pi \in UR_{a_m}^{HG}$ telles que :

$$C_{a_m,i} \sim_{a_m} C_{a_m}^{\Pi} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$$

Si $IS_{HG} = \emptyset$, la manipulation constructive mise en œuvre par a_m est k-rationnelle si et seulement si $\forall i \in [1,k[$, $(2)_i = 0$ et $(2)_k > 0$. Si $IS_{HG} \neq \emptyset$, la manipulation constructive mise en œuvre par a_m est k-rationnelle si et seulement si :

$$\forall i \in [1,k[, \frac{|\{\Pi \in IS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,i}\}|}{|IS_{HG}|} = \frac{(1)_i + (2)_i}{card_{M_C}(\mathcal{P}_{HG}|HG)}$$
$$\text{et } \frac{|\{\Pi \in IS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,k}\}|}{|IS_{HG}|} < \frac{(1)_k + (2)_k}{card_{M_C}(\mathcal{P}_{HG}|HG)}$$

Intuitivement, $(1)_i$ (respectivement $(2)_i$) représente le nombre de structures de coalitions individuellement stables dans HG^{M_C} construites à partir d'une structure de coalitions $\Pi \in IS_{HG}$ (respectivement $\Pi \in UR_{a_m}^{HG}$) telles que soit a_m, soit l'agent Sybil forme la coalition $C_{a_m,i}$.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ tel que $\mathbb{P}$ retourne la solution du jeu HG aléatoirement uniformément parmi IS_{HG} et un agent malhonnête $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Fixons HG^{M_C} le jeu résultant de la mise en œuvre de la manipulation constructive par a_m sur HG. Rappelons que par la définition 3.2.5, la manipulation M_C est k-rationnelle si, $\forall i \in [1,k[$, $\mathrm{P}^*(a_m,i|HG) = \mathrm{P}^*(a_m,i|HG^M)$ et $\mathrm{P}^*(a_m,k|HG) < \mathrm{P}^*(a_m,k|HG^M)$.
Montrons dans un premier temps les conditions nécessaires lorsque $IS_{HG} = \emptyset$. Comme $IS_{HG} = \emptyset$, pour tout k tel que $C_{a_m,k} \succ_{a_m} \{a_m\}$, nous avons $\mathrm{P}^*(a_m,k|HG) = 0$. Par conséquent, la manipulation M_C est k-rationnelle si :

1. pour tout $i \in [1,k[$, il n'existe pas de structure de coalitions $\Pi' \in IS_{HG^{M_C}}$ telle que $C_{a_m}^{\Pi'} \sim_{a_m} C_{a_m,i}$ ou que $(C_s^{\Pi'} \setminus \{s\}) \cup \{a_m\} \sim_{a_m} C_{a_m,i}$;
2. il existe une structure de coalitions $\Pi' \in IS_{HG^{M_C}}$ telle que $C_{a_m}^{\Pi'} \sim_{a_m} C_{a_m,k}$ ou que $(C_s^{\Pi'} \setminus \{s\}) \cup \{a_m\} \sim_{a_m} C_{a_m,k}$.

Par la propriété 5.2.4, s'il existe une structure de coalitions $\Pi \in UR_{a_m}^{HG}$ telle que $C_{a_m,i} \sim_{a_m} C_{a_m}^{\Pi} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$ alors il existe au moins une structure de coalitions $\Pi' \in IS_{HG^{M_C}}$ telle que $C_{a_m}^{\Pi'} \sim_{a_m} C_{a_m,i}$ ou que $(C_s^{\Pi'} \setminus \{s\}) \cup \{a_m\} \sim_{a_m} C_{a_m,i}$. Si nous notons par $(2)_i$ le nombre de ces structures Π alors les conditions nécessaires à la k-rationalité de la manipulation constructive lorsque $IS_{HG} = \emptyset$ sont $\forall i \in [1,k[$, $(2)_i = 0$ et $(2)_k > 0$.
Montrons maintenant les conditions nécessaires à la k-rationalité de la manipulation constructive lorsque $IS_{HG} \neq \emptyset$. Par définition, la manipulation constructive est k-rationnelle si :

$$\forall i \in [1,k[, \frac{|\{\Pi \in IS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,i}\}|}{|IS_{HG}|} = \frac{|\{\Pi \in IS_{HG^{M_C}} | \mathrm{Cond}_1(i) \vee \mathrm{Cond}_2(i)\}|}{card_{M_C}(\mathcal{P}_{HG}|HG)}$$
$$\text{et } \frac{|\{\Pi \in IS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,k}\}|}{|IS_{HG}|} < \frac{|\{\Pi \in IS_{HG^{M_C}} | \mathrm{Cond}_1(k) \vee \mathrm{Cond}_2(k)\}|}{|IS_{HG^{M_C}}|}$$

où $\mathrm{Cond}_1(i)$ et $\mathrm{Cond}_2(i)$ désignent respectivement les conditions $C_{a_m}^{\Pi'} \sim_{a_m} C_{a_m,i}$ et $(C_s^{\Pi'} \setminus \{s\}) \cup \{a_m\} \sim_{a_m} C_{a_m,i}$. Fixons une structure de coalitions $\Pi' \in IS_{HG^{M_C}}$. Soit $\Pi \in \mathcal{P}_N$ la structure

de coalitions telle que $\Pi = f^{-1}(\Pi')$ et $C_0 = C_s^{\Pi'}$. Par la propriété 5.2.3, si $\Pi \in IS_{HG}$ alors $\text{Cond}_1(i) \vee \text{Cond}_2(i)$ est satisfaite si :

$$C_{a_m,i} \sim_{a_m} C_{a_m}^{\Pi} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$$
$$\text{et} \quad \nexists C \in \Pi : C \cup \{a_m\} \succ_{a_m} C_{a_m,i}$$

Notons par $(1)_i$ le nombre de structures de coalitions $\Pi' \in IS_{HG^{M_C}}$ telles que $f^{-1}(\Pi') \in IS_{HG}$ satisfait $\text{Cond}_1(i) \vee \text{Cond}_2(i)$. Comme nous l'avons montré précédemment par la propriété 5.2.7, $(2)_i$ représente le nombre de structures de coalitions $\Pi' \in IS_{HG^{M_C}}$ tel que $f^{-1}(\Pi') \notin IS_{HG}$ satisfaisant $\text{Cond}_1(i) \vee \text{Cond}_2(i)$. Ainsi, nous avons $|\{\Pi \in IS_{HG^{M_C}} | \text{Cond}_1(i) \vee \text{Cond}_2(i)\}| = (1)_i + (2)_i$.
Par conséquent, comme $|IS_{HG^{M_C}}| = card_{M_C}(\mathcal{P}_{HG}|HG)$, si $IS_{HG} \neq \emptyset$ alors la manipulation constructive est k-rationnelle si et seulement si :

$$\forall i \in [1,k[, \frac{|\{\Pi \in IS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,i}\}|}{|IS_{HG}|} = \frac{(1)_i + (2)_i}{card_{M_C}(\mathcal{P}_{HG}|HG)}$$
$$\text{et } \frac{|\{\Pi \in IS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,k}\}|}{|IS_{HG}|} < \frac{(1)_k + (2)_k}{card_{M_C}(\mathcal{P}_{HG}|HG)}$$

□

Étudions maintenant le cas de la stabilité au sens du cœur. La rationalité de la manipulation constructive pour ce concept de solution découle également du fait que l'agent Sybil s est accepté dans toute coalition par les agents honnêtes. Ainsi, l'agent Sybil peut rejoindre la coalition qu'il désire.

Propriété 5.2.6 : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ tel que $\mathbb{P}$ retourne la solution du jeu HG aléatoirement uniformément parmi CS_{HG}, l'ensemble des structures de coalitions stables au sens du cœur dans HG. Soit $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Soit $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la manipulation constructive mis en œuvre par a_m. Soit $\Pi \in CS_{HG}$ une structure de coalitions stable au sens du cœur dans HG. Pour toute coalition $C_0 \in \Pi \cup \{\emptyset\}$, la structure de coalitions $\Pi' = f(\Pi,s,C_0)$ est stable au sens du cœur dans $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ si et seulement si :

$$C_0 \cup \{a_m\} \succeq \{a_m\} \text{ et } C_0 \neq C_{a_m}^{\Pi}$$

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$, un agent malhonnête $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$, une structure de coalitions $\Pi \in CS_{HG}$ et une coalition $C_0 \in \Pi \cup \{\emptyset\}$. Soit la structure de coalitions Π' telle que $\Pi' = f(\Pi,s,C_0)$.
Montrons dans un premier temps que si $\{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$ ou que $C_0 = C_{a_m}^{\Pi}$, la structure de coalitions Π' n'est pas stable au sens du cœur dans le jeu HG^{M_C}. Supposons que $\{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$. Par définition du profil de préférence de s, nous avons $\{s\} \succ_s^{M_C} C_0 \cup \{s\}$. Or, par construction de Π', $C_0 \cup \{s\} \in \Pi'$. Par conséquent, si $\{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$, la structure de coalitions Π n'est pas stable au sens du cœur dans le jeu $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$. Supposons maintenant que $C_0 = C_{a_m}^{\Pi}$. Par définition du profil de préférence de s, nous avons $\{s\} \succ_s^{M_C} C_{a_m}^{\Pi} \cup \{s\}$. Ainsi, si $C_0 = C_{a_m}^{\Pi}$, la structure de coalitions Π' n'est pas stable au sens du cœur

dans le jeu HG^{M_C}.
Montrons maintenant que si $C_0 \cup \{a_m\} \succeq \{a_m\}$ et que $C_0 \neq C^{\Pi}_{a_m}$, la structure de coalitions Π' est stable au sens du cœur. Par construction du profil de préférence de a_m, comme $C_0 \neq C^{\Pi}_{a_m}$, nous avons $\nexists N_2 \subseteq N \cup \{s\}$ tel que $a_m \in N_2$ et que $N_2 \succ^{M_C}_{a_m} C^{\Pi'}_{a_m}$. Comme $\Pi \in CS_{HG}$, par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute, nous avons $\nexists N_2 \subseteq N \cup \{s\} : \forall a_i \in N_2 \setminus \{a_m,s\}, N_2 \succ^{M_C}_{a_i} C^{\Pi}_{a_i}$. Par conséquent, la structure de coalitions Π' est stable au sens du cœur dans le jeu HG^{M_C}. □

Nous pouvons en déduire le nombre de structures de coalitions stables au sens du cœur dans $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P}\rangle$ construites à partir d'une structure de coalitions $\Pi \in CS_{HG}$. Intuitivement, il en existe autant que de coalitions dans Π vérifiant les conditions de la propriété 5.2.3.

Corollaire 5.2.3 : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$ et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P}\rangle$ le jeu résultant de la manipulation constructive mise en œuvre par un agent malhonnête $a_m \in N$. Pour toute structure de coalitions $\Pi \in CS_{HG}$, nous avons :

$$card_{M_C}(\Pi|HG) = |\{C_0 \in \Pi \cup \{\emptyset\} \setminus C^{\Pi}_{a_m} | C_0 \cup \{a_m\} \succeq \{a_m\}\}|$$

Étudions dans un second temps l'influence de la manipulation constructive sur les structures de coalitions non stables au sens du cœur.

Propriété 5.2.7 : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$ tel que $\mathbb{P}$ retourne la solution du jeu HG aléatoirement uniformément parmi CS_{HG}, l'ensemble des structures de coalitions stables au sens du cœur dans HG. Soit $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Soit $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P}\rangle$ le jeu résultant de la manipulation constructive mis en œuvre par a_m. Soit $\Pi \notin CS_{HG}$ une structure de coalitions non stable au sens du cœur dans HG. Pour toute coalition $C_0 \in \Pi \cup \{\emptyset\}$, la structure de coalitions $\Pi' = f(\Pi, s, C_0)$ est stable au sens du cœur dans HG^{M_C} si et seulement si :

1. $C_0 \cup \{a_m\} \succeq \{a_m\}$ et $C_0 \neq C^{\Pi}_{a_m}$;
2. $\forall N_2 \subseteq N : \forall a_i \in N_2, N_2 \succ_{a_i} C^{\Pi}_{a_i}$, $a_m \in N_2$.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P}\rangle$, un agent malhonnête $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$, une structure de coalitions $\Pi \notin CS_{HG}$ et une coalition $C_0 \in \Pi \cup \{\emptyset\}$. Soit la structure de coalitions Π' telle que $\Pi' = f(\Pi, s, C_0)$.
Montrons dans un premier temps que si $\{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$ ou que $C_0 = C^{\Pi}_{a_m}$, la structure de coalitions Π' n'est pas stable au sens du cœur dans le jeu. Pour cela, supposons que $\{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$. Par définition du profil de préférence de s, nous avons $\{s\} \succ^{M_C}_{s} C_0 \cup \{s\}$. Or, par construction de Π', $C_0 \cup \{s\} \in \Pi'$. Par conséquent, si $\{a_m\} \succ_{a_m} C_0 \cup \{a_m\}$ alors la structure de coalitions Π n'est pas stable au sens du cœur dans $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P}\rangle$. Supposons maintenant que $C_0 = C^{\Pi}_{a_m}$. Par définition du profil de préférence de s, nous avons $\{s\} \succ^{M_C}_{s} C^{\Pi}_{a_m} \cup \{s\}$. Ainsi, si $C_0 = C^{\Pi}_{a_m}$ alors la structure de coalitions Π' n'est pas stable au sens du cœur dans HG^{M_C}.
Montrons dans un second temps que si il existe $N_2 \subseteq N : \forall a_i \in N_2, N_2 \succ_{a_i} C^{\Pi}_{a_i}$ tel que $a_m \notin N_2$, la structure de coalitions Π' n'est pas stable au sens du cœur dans $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P}\rangle$. Supposons qu'il existe un tel sous-ensemble d'agents N_2. Par les hypothèses d'indépendance des

alternatives non-pertinentes et de bénéfice du doute, nous avons : $\forall a_i \in N_2, N_2 \succ_{a_i}^{M_C} C_{a_i}^{\Pi} \succ_{a_i}^{M_C} C_{a_i}^{\Pi} \cup \{s\}$. Par construction de Π', nous avons $\forall a_i \in N$, soit $C_{a_i}^{\Pi'} = C_{a_i}^{\Pi}$, soit $C_{a_i}^{\Pi'} = C_{a_i}^{\Pi} \cup \{s\}$. Dans les deux cas, il existe $N_2 \subseteq N \cup \{s\}$: ,$N_2 \succ_{a_i}^{M_C} C_{a_i}^{\Pi'}$. Par conséquent, la structure de coalitions Π' n'est pas stable au sens du cœur dans le jeu HG^{M_C}.
Supposons enfin que $\forall N_2 \subseteq N : \forall a_i \in N_2, N_2 \succ_{a_i} C_{a_i}^{\Pi}$, $a_m \in N_2$. Par définition du profil de préférence de a_m, nous avons $N_2 \sim_{a_m}^{M_C} C_{a_i}^{\Pi}$. Ainsi, dans le jeu HG^{M_C}, $\exists a_i \in N_2 : C_{a_i}^{\Pi} \sim_{a_m}^{M_C} N_2$. Par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute, nous avons donc $\nexists N_2 \subseteq N \cup \{s\} : \forall a_i \in N_2, N_2 \succ_{a_i} C_{a_i}^{\Pi}$. Par conséquent, la structure de coalitions Π' est stable au sens du cœur dans HG^{M_C}. □

Nous pouvons déduire le nombre de structures de coalitions stables au sens du cœur dans $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ construites à partir d'une structure de coalitions $\Pi \notin CS_{HG}$.

Corollaire 5.2.4 : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et $HG^{M_C} = \langle N^{M_C}, \succeq^{M_C}, \mathbb{P} \rangle$ le jeu résultant de la manipulation constructive mise en œuvre par un agent malhonnête $a_m \in N$. Pour toute structure de coalitions $\Pi \notin CS_{HG}$, si $\exists N_2 \subseteq N : \forall a_i \in N_2, N_2 \succ_{a_i} C_{a_i}^{\Pi}$ et que $a_m \notin N_2$ alors $card_{M_C}(\Pi | HG) = 0$. Sinon :

$$card_{M_C}(\Pi | HG) = |\{C_0 \in \Pi \cup \{\emptyset\} \setminus C_{a_m}^{\Pi} | C_0 \cup \{a_m\} \succeq \{a_m\}\}|$$

À partir des corollaires 5.2.3 et 5.2.4, nous pouvons déduire les conditions à la rationalité de la manipulation constructive sur les jeux utilisant la stabilité du cœur comme concept de solution.

Propriété 5.2.8 : Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique telle que $\mathbb{P}$ retourne la solution du jeu HG aléatoirement uniformément parmi CS_{HG}, l'ensemble des structures de coalitions stables au sens du cœur dans HG et $a_m \in N$ un agent malhonnête ayant le profil de préférence $\succeq_{a_m} = C_{a_m,1} \succeq_{a_m} C_{a_m,2} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Soit $(1)_i$ le nombre de structures de coalitions $\Pi \in CS_{HG}$ telles que :

$$C_{a_m,i} \sim_{a_m} C_{a_m}^{\Pi} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$$

Soit $(2)_i$ le nombre de structures de coalitions $\Pi \notin CS_{HG}$ telles que :

$$C_{a_m,i} \sim_{a_m} C_{a_m}^{\Pi} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$$
$$\text{et } \forall N_2 \subseteq N : \forall a_i \in N_2, N_2 \succ_{a_i} C_{a_i}^{\Pi}, a_m \in N_2$$

Si $CS_{HG} = \emptyset$, la manipulation constructive mise en œuvre par a_m est k-rationnelle si et seulement si $\forall i \in [1,k[$, $(2)_i = 0$ et $(2)_k > 0$. Si $CS_{HG} \neq \emptyset$, la manipulation constructive mise en œuvre par a_m est k-rationnelle si et seulement si :

$$\forall i \in [1,k[, \frac{|\{\Pi \in CS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,i}\}|}{|CS_{HG}|} = \frac{(1)_i + (2)_i}{card_{M_C}(\mathcal{P}_{HG} | HG)}$$
$$\text{et } \frac{|\{\Pi \in CS_{HG} | C_{a_m}^{\Pi} \sim_{a_m} C_{a_m,k}\}|}{|CS_{HG}|} < \frac{(1)_k + (2)_k}{card_{M_C}(\mathcal{P}_{HG} | HG)}$$

Pour des raisons de lisibilité, la démonstration de cette propriété peut être trouvée en annexe B et repose sur le même principe que la démonstration de la propriété 5.2.5. Intuitivement,

$(1)_i$ (respectivement $(2)_i$) représente le nombre de structures de coalitions stables au sens du cœur dans HG^{M_C} construites à partir d'une structure de coalitions $\Pi \in CS_{HG}$ (respectivement $\Pi \notin CS_{HG}$) telles que soit a_m, soit l'agent Sybil forme la coalition $C_{a_m,i}$.

5.2.3 Destruction du cœur

Si la manipulation destructive n'est pas k-rationnelle, existe-t-il d'autres formes de manipulation destructive rationnelles, c'est-à-dire réduisant l'ensemble des structures de coalitions stables, pour la stabilité au sens du cœur ? C'est le cas lorsque l'agent malhonnête fournit un faux profil de préférence tout en introduisant deux agents Sybils s_1 et s_2.

Définition 5.2.1 - *Destruction du cœur* : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ avec un agent $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$ et $k \in [1,2^{n-1}]$ tel que :

1. $\forall i \in [1,k[,\ \nexists \Pi \in CS_{HG} : C_{a_m,i} \in \Pi$;
2. $\exists \Pi \in CS_{HG} : C_{a_m,k} \in \Pi$.

La *destruction du cœur* est la manipulation définie par :

$$
\begin{array}{rcl}
M_K & = & \langle \{a_m,s_1,s_2\}, \{\succeq_{a_m}^{M_K}, \succeq_{s_1}^{M_K}, \succeq_{s_2}^{M_K}\}, \succeq_{a_m} \rangle \\
\text{avec} \quad \succeq_{a_m}^{M_K} & = & C_{a_m,k} \succ_{a_m}^{M_K} \{a_m,s_1\} \succ_{a_m}^{M_K} \{a_m,s_2\} \succ_{a_m}^{M_K} \{a_m\} \\
\succeq_{s_1}^{M_K} & = & \{s_1,s_2\} \succ_{s_1}^{M_K} \{a_m,s_1\} \succ_{s_1}^{M_K} \{s_1\} \\
\succeq_{s_2}^{M_K} & = & \{a_m,s_2\} \succ_{s_2}^{M_K} \{s_1,s_2\} \succ_{s_2}^{M_K} \{s_2\}
\end{array}
$$

Cette manipulation consiste à créer des profils de préférence circulaires entre a_m, s_1, et s_2 afin que toutes les structures de coalitions $\Pi' \in \mathcal{P}_{N \cup \{s_1,s_2\}}$ ne contenant pas la coalition $C_{a_m,k}$ ne puissent pas être stables au sens du cœur.

Exemple 5.2.1 - Considérons le jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ suivant :

$$
\begin{array}{rcl}
N & = & \{a_1,a_2,a_m\} \\
\succeq_{a_1} & = & \{a_1,a_2,a_m\} \succ_{a_1} \{a_1,a_2\} \sim_{a_1} \{a_1,a_m\} \succ_{a_1} \{a_1\} \\
\succeq_{a_2} & = & \{a_1,a_2,a_m\} \succ_{a_2} \{a_1,a_2\} \succ_{a_2} \{a_2,a_m\} \succ_{a_2} \{a_2\} \\
\succeq_{a_m} & = & \{a_1,a_m\} \succ_{a_m} \{a_2,a_m\} \succ_{a_m} \{a_m\}
\end{array}
$$

Comme illustré sur la figure 5.1, nous avons $CS_{HG} = \{\{\{a_1,a_m\},\{a_2\}\},\{\{a_1,a_2\},\{a_m\}\}\}$. Ainsi, l'agent a_m peut fixer $C_{a_m,k} = \{a_1,a_m\}$ et construire la manipulation M_K suivante :

$$
\begin{array}{rcl}
M_K & = & \langle \{a_m,s_1,s_2\}, \{\succeq_{a_m}^{M_K}, \succeq_{s_1}^{M_K}, \succeq_{s_2}^{M_K}\}, \succeq_{a_m} \rangle \\
\text{avec} \quad \succeq_{a_m}^{M_K} & = & \{a_1,a_m\} \succ_{a_m}^{M_K} \{a_m,s_1\} \succ_{a_m}^{M_K} \{a_m,s_2\} \succ_{a_m}^{M_K} \{a_m\} \\
\succeq_{s_1}^{M_K} & = & \{s_1,s_2\} \succ_{s_1}^{M_K} \{a_m,s_1\} \succ_{s_1}^{M_K} \{s_1\} \\
\succeq_{s_2}^{M_K} & = & \{a_m,s_2\} \succ_{s_2}^{M_K} \{s_1,s_2\} \succ_{s_2}^{M_K} \{s_2\}
\end{array}
$$

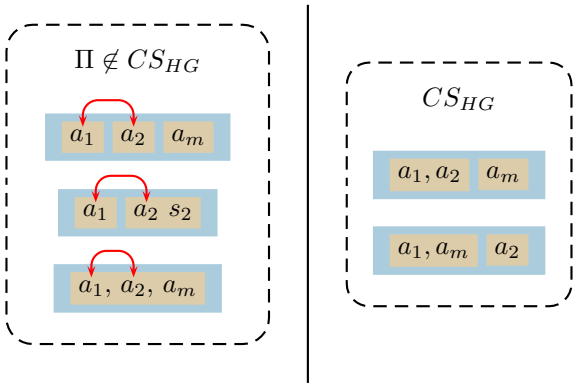

FIGURE 5.1 – Stablité au sens du cœur des structures de coalitions de HG

La manipulation M_K produit le jeu HG^{M_K} où $CS_{HG^{M_K}} = \{ \{ \{a_1,a_m\},\{a_2\},\{s_1,s_2\} \} \}$.

Dans cet exemple, la destruction du cœur est 1-rationnelle. Cependant, sa rationalité n'est pas garantie pour l'ensemble des jeux hédoniques. En effet trivialement, comme pour la manipulation destructive, la manipulation M_K n'est pas rationnelle lorsque l'ensemble des structures de coalitions stables au sens du cœur est vide. Caractérisons maintenant les conditions nécessaires à la k-rationalité de M_K lorsque $CS_{HG} \neq \emptyset$.

Propriété 5.2.9 : Soit un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succ_{a_m} \ldots \succ_{a_m} C_{a_m,k} \succ_{a_m} \ldots \succ_{a_m} C_{a_m,2^{n-1}}$. Soit $\mathbb{P}$ un protocole de sélection tirant la solution d'un jeu parmi l'ensemble des structures de coalitions stables au sens du cœur. La manipulation M_K est k-rationnelle si et seulement si :

1. $\exists \Pi \in CS_{HG} : C^{\Pi}_{a_m} = C_{a_m,k}$;
2. $\exists i \in]k,2^{n-1}], \exists \Pi \in CS_{HG} : C^{\Pi}_{a_m} = C_{a_m,i}$.

Cette propriété repose sur le fait que les profils de préférence de a_m, de s_1 et de s_2 rendent toute structure de coalitions ne contenant pas $C_{a_m,k}$ non stable au sens du cœur.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et un agent malhonnête $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$.
Par définition du profils de préférence de a_m dans la destruction du cœur, il n'existe pas de structure de coalitions $\Pi \in CS_{HG}$ telle que $\exists i \in [1,k[: C_{a_m,i} \in \Pi$ et donc $\forall i \in [1,k[\mathrm{P}^*(a_m,i|HG) = 0$. Par définition de la k-rationalité d'une manipulation, la destruction du cœur est k-rationnelle si $\forall i \in [1,k[\mathrm{P}^*(a_m,i|HG^{M_K}) = 0$. Par ailleurs, s'il n'existe pas de $i \in]k,2^{n-1}]$ tel que $\exists \Pi \in CS_{HG} : C^{\Pi}_{a_m} = C_{a_m,i}$, nous avons $\mathrm{P}^*(a_m,k|HG^{M_K}) = 1$ et donc la destruction du cœur ne peut pas être k-rationnelle.
Fixons une structure de coalitions $\Pi' \in \mathcal{P}_{N \cup \{s_1,s_2\}}$. Supposons dans un premier temps que $C^{\Pi'}_{a_m} \neq C_{a_m,k}$. Par définition des profils de préférence de a_m, s_1 et de s_2, si $\exists a_i \in N \setminus \{a_m\}$ tel que $C^{\Pi'}_{a_i} = C^{\Pi'}_{s_1}$ (respectivement si $C^{\Pi'}_{a_i} = C^{\Pi'}_{s_2}$) alors la structure de coalitions Π' ne peut pas être stable au sens du cœur puisque $\{s_1\} \succ^{M_K}_{s_1} C^{\Pi'}_{s_1}$ (respectivement $\{s_2\} \succ^{M_K}_{s_2} C^{\Pi'}_{s_2}$). De même, si $C^{\Pi'}_{a_m} = \{a_m,s_1,s_2\}$ alors Π' n'est pas stable. Enfin, si $\exists a_i \in N \setminus \{a_m\}$ tel que $C^{\Pi'}_{a_i} = C^{\Pi'}_{a_m}$ alors Π' n'est toujours pas stable.
Ainsi, il ne reste que quatre coalitions possibles pour les agents a_m, s_2 et s_1 : $\{a_m\},\{s_1\},\{s_2\}$,

$\{a_m,s_1\},\{s_2\}$, $\{a_m,s_2\},\{s_1\}$ et $\{a_m\},\{s_1,s_2\}$. Dans ces quatre cas, indépendamment des coalitions des agents de $N \setminus \{a_m\}$, la structure de coalitions Π' ne peut pas être stable au sens du cœur, car au moins 2 des agents (parmi a_m, s_2 et s_1) désirent changer de coalition pour en former une nouvelle ensemble. Ainsi, si $C^{\Pi'}_{a_m} \neq C_{a_m,k}$ alors la structure de coalitions Π' ne peut pas stable au sens du cœur. La figure 5.2 représente pour ces 4 cas quels agents rendent la structure non stable.

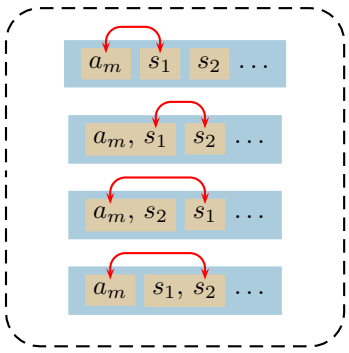

FIGURE 5.2 – Instabilité du cœur due aux agents malhonnêtes

Considérons maintenant que $C^{\Pi'}_{a_m} = C_{a_m,k}$. Par définition des profils de préférence de s_1 et de s_2, pour toute structure de coalitions $\Pi' \in \mathcal{P}_{N\cup\{s_1,s_2\}}$, si $C^{\Pi'}_{s_1} \neq \{s_1,s_2\}$ alors la structure de coalitions Π' ne peut pas être stable au sens de du cœur. Nous avons deux cas :

1. Considérons une structure de coalitions $\Pi \in CS_{HG}$ telle que $C_{a_m,k} \in \Pi$ et la structure $\Pi' = \Pi \cup \{s_1,s_2\}$. Par définition des profils de préférence de s_1 et de s_2, les agents a_m, s_1 et s_2 ne désirent pas changer de coalition dans Π'. Comme la structure de coalitions $\Pi \in CS_{HG}$, par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute (hypothèse 3.2.3 et 3.2.4), aucun autre sous-ensemble de $N\setminus\{a_m\}$ ne souhaite aussi changer de coalition dans Π'. Donc, Π' est stable au sens du cœur dans le jeu HG^{M_K}.

2. Considérons enfin une structure de coalitions $\Pi \notin CS_{HG}$ telle que $C_{a_m,k} \in \Pi$ et la structure $\Pi' = \Pi \cup \{s_1,s_2\}$. Comme $\Pi \notin CS_{HG}$, $C_{a_m,k} \in \Pi$, nous avons $\exists N_2 \subseteq N \setminus \{a_m\}$ tel que $\forall a_i \in N_2, N_2 \succ_{a_i} C^{\Pi}_{a_i}$. Comme par construction, comme $\{s_1,s_2\} \in \Pi'$, $C^{\Pi}_{a_i} = C^{\Pi'}_{a_i}$. Ainsi, par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute, nous avons $\forall a_i \in N_2, N_2 \succ^{M_K}_{a_i} C^{\Pi'}_{a_i}$ et donc la structure de coalitions Π' n'est pas stable au sens du cœur dans le jeu HG^{M_K}.

Ainsi, il existe une structure de coalitions $\Pi' \in CS_{HG^{M_K}}$ si et seulement s'il existe $\Pi \in CS_{HG}$ telle que $C_{a_m,k} \in \Pi$. Comme toute structure de coalitions $\Pi' \in CS_{HG^{M_K}}$ contient la coalition $C_{a_m,k}$, nous avons :

- $\forall i \in [1,k[\mathrm{P}^*(a_m,i|HG^{M_K}) = \mathrm{P}^*(a_m,i|HG) = 0$;
- $\mathrm{P}^*(a_m,k|HG^{M_K}) = 1 > \mathrm{P}^*(a_m,k|HG)$.

Par conséquent, la manipulation M_K n'est k-rationnelle que si et seulement si les conditions (1) et (2) sont satisfaites. □

Remarquons que les conditions nécessaires à la rationalité de la destruction du cœur sont les mêmes que celles de la manipulation destructive au concept de solution près (voir les conditions 2 et 3 de la propriété 4.2.6).

5.2.4 Fréquences des jeux manipulables

Dans la section 4.3.2, nous avons montré empiriquement que les conditions nécessaires à la rationalité d'une manipulation sont rarement satisfaites lorsque le concept de solution satisfait la stabilité au sens de Nash. Nous montrons ici que la proportion de jeux hédoniques manipulables est beaucoup plus importante pour les concepts de solution de stabilité individuelle et de stabilité au cœur. Pour cela, nous proposons de suivre le même protocole expérimental. Pour un ensemble d'agents variant entre 3 et 10, nous générons 10 000 jeux hédoniques dont les préférences des agents sont définies aléatoirement uniformément. Nous calculons pour chacun de ces jeux s'il existe au moins un agent pour qui une manipulation est k-rationnelle.

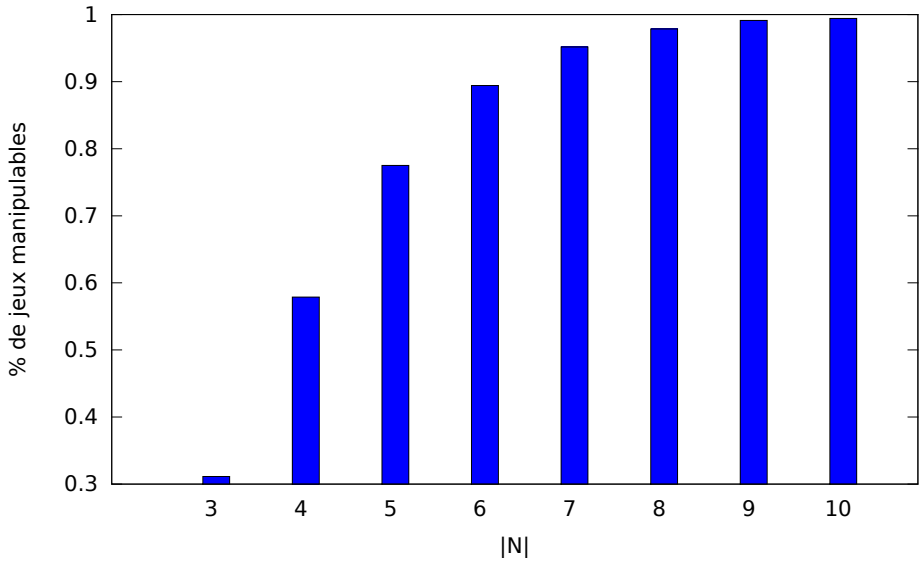

FIGURE 5.3 – Jeux manipulables en fonction du nombre d'agents pour la stabilité individuelle

Étudions d'abord le cas de la stabilité individuelle où, parmi les manipulations que nous avons considérées, seule la manipulation constructive peut être k-rationnelle. La figure 5.3 montre la proportion de jeux hédoniques manipulables dans ces conditions. Le premier constat que nous pouvons faire est que, contrairement à la stabilité au sens de Nash, la stabilité individuelle est majoritairement sensibles à la manipulation constructive et d'autant plus fréquemment qu'il y a d'agents dans le système. Un tiers des jeux où $n = 3$ sont manipulables et plus de 90 % des jeux le sont aussi pour $n \geq 6$. Cela s'explique par le fait que, dans une structure de coalitions individuellement stable, si un agent préfère rejoindre une autre coalition, C mais qu'il est rejeté, il lui suffit d'introduire un agent Sybil qui sera alors accepté dans C.

Regardons maintenant le cas de la stabilité au sens du cœur. Nous étudions ici la proportion de jeux manipulable par la manipulation constructive ainsi que par la destruction du cœur. La figure 5.4 montre la proportion jeux hédoniques manipulables dans ces conditions. Comme pour la stabilité individuelle, nous constatons que la stabilité au sens du cœur n'est pas un concept de solution robuste aux manipulations. Comparativement à la stabilité individuelle, la manipulation constructive est moins fréquemment rationnelle en règle générale, sauf dans le cas particulier où $n = 3$, car il y a plus souvent des jeux où l'agent malhonnête est l'unique responsable de la non-stabilité. La destruction du cœur, quant à elle, est k-rationnelle dans environs 25 % des jeux où $n = 3$ et cela augmente légèrement lorsque le nombre d'agents augmente. Pour $n = 10$, un tiers

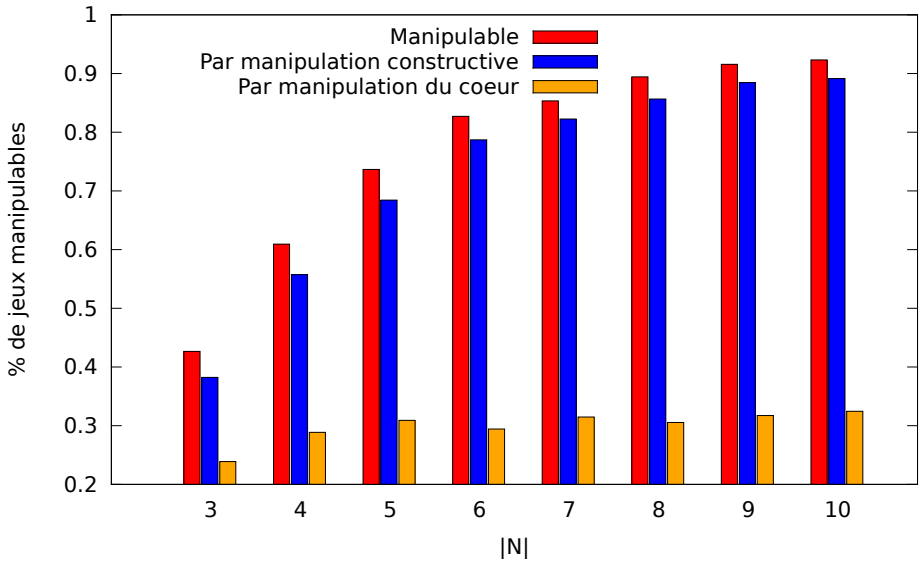

FIGURE 5.4 – Jeux manipulables selon le nombre d'agents pour la stabilité au sens du cœur

des jeux sont manipulable. Il est intéressant de constater qu'avec des conditions structurellement proches de la manipulation destructive pour la stabilité au sens de Nash, la destruction du cœur reste globalement toujours autant efficace, et cela contrairement à la manipulation destructive.

Ainsi, nous pouvons dire que la stabilité individuelle et la stabilité au sens du cœur ne sont pas robustes aux manipulations.

5.3 Conclusion

Dans ce chapitre, nous avons étudié la robustesse des jeux hédoniques face aux manipulations sous d'autres hypothèses que celles décrites dans la section 4. En premier lieu, nous avons remis en cause l'hypothèse, favorable aux agents malhonnêtes, du bénéfice du doute. À la place, nous avons considéré d'une part une *hypothèse de sous-additivité* qui rend plus restrictives les conditions nécessaires à la rationalité de la manipulation constructive, et d'autre part une *hypothèse de sur-additivité* qui, de manière intéressante, n'est pas avantageuse pour les agents malhonnêtes car elle rend plus restrictive les conditions à la rationalité des deux types de manipulations.

En second lieu, nous avons étudié d'autres concepts de solutions que la stabilité au sens de Nash. La *stabilité individuelle* et la *stabilité au sens du cœur* rendent non rationnelle la manipulation destructive. Bien que la manipulation destructive ne soit plus rationnelle sur ces deux concepts, la manipulation constructive le reste sous certaines conditions. De plus, nous avons exhibé une manipulation spécifique au cœur, la *destruction du cœur* dont les conditions de rationalité sont similaires à celles de la manipulation destructive sur la stabilité au sens de Nash.

Nous avons montré empiriquement que les deux concepts de solutions sont très sensibles aux manipulations. Contrairement à la stabilité au sens de Nash, plus le nombre d'agents est important, plus les jeux hédoniques sont fréquemment manipulables. De plus, nous n'avons pas prouvé que la manipulation constructive et la destruction du cœur sont des manipulations aux conditions minimalement restrictives. Ainsi, cela renforce nos résultats, car, s'il existe des

manipulations moins restrictives, les jeux hédoniques seront d'autant plus manipulables.

Cependant, nous avons considéré la stabilité individuelle et la stabilité au sens du cœur sous l'hypothèse du bénéfice du doute. Relâcher cette hypothèse permettrait d'accroître la robustesse de ces concepts. Intuitivement, la manipulation constructive n'est jamais rationnelle sur la stabilité individuelle lorsque les préférences sont sous-additives car les agents honnêtes refuse alors que l'agent Sybil rejoigne leur coalition.

Troisième partie

Manipulations d'un modèle dynamique - les systèmes de réputation

Chapitre 6

Un modèle d'interaction

Sommaire

Résumé.
Dans ce chapitre, nous présentons un modèle générique d'interaction entre agents sous forme d'échange de services afin d'étudier la robustesse des systèmes de réputation aux manipulations. Nous présentons comment les agents peuvent utiliser le résultat de leurs interactions passées afin de calculer une valeur de réputation correspondant à une estimation collective du comportement des autres agents. Cette estimation collective du comportement des agents a pour objectif d'aider les agents du système qui cherchent à maximiser leurs gains personnels. Nous présentons dans une seconde section des agents malveillants dont l'objectif est de diminuer le gain des autres agents du système. Nous présentons ensuite un modèle des principales manipulations que peut utiliser une collusion d'agents malveillants pour atteindre cet objectif.

Afin de simplifier la lecture de ce chapitre, le tableau 6.1 résume les principales notations utilisées. Nous donnons en annexe A.2 les notations portant sur les systèmes de réputation utilisées dans ce manuscrit.

Échange de services	
$N = \{a_1, \ldots, a_n\}$	Ensemble des agents du système
$S = \{s_1, \ldots, s_m\}$	Ensemble des services fournis par les agents
$N_x \subseteq N$	Ensembles des agents fournissant le service s_x
$S_k \subseteq S$	Ensemble des services fournis par l'agent a_k
$\varepsilon_{k,x}$	Expertise de l'agent $a_k \in N_x$ pour le service $s_x \in S_k$
Évaluation de l'agent	
v_i	Fonction d'évaluation de l'agent $a_i \in N$
$v^t_{i,k,x}$	Évaluation de a_i de la qualité du service s_x fourni par l'agent a_k à l'instant t
g^t_i	Gains totaux observés par l'agent $a_i \in N$ à l'instant t
$c_{i,k,x}$	Confiance de l'agent $a_i N$ envers l'agent $a_k \in N$ pour fournir le service $s_x \in S$
f_i	Fonction de réputation l'agent $a_i \in N$
Observations et témoignages	
$O_{i,k,x}$	Ensemble des observations de l'agent $a_i \in N$ pour les services $s_x \in S$ fournis par l'agent $a_k \in N_x$
$F_{i,j,k,x}$	Ensemble des témoignages que l'agent $a_j \in N$ a fournis à l'agent $a_i \in N$ vis-à-vis des services $s_x \in S$ rendus par l'agent $a_k \in N_x$
$\mathcal{F}_i$	Ensemble des témoignages et des observations de l'agent $a_i \in N$

Tableau 6.1 – Principales notations du modèle d'échange de service

6.1 Système d'échange de services

6.1.1 Un modèle générique d'interaction

Dans la suite de ce manuscrit, nous nous intéressons à un système où les agents sont amenés à interagir entre eux au cours du temps. Pour modéliser ces interactions entre les agents, nous considérons un système générique d'échanges de services entre agents. Il s'agit d'un système multi-agents hétérogène ouvert où les agents disposent de ressources propres qu'ils peuvent temporairement mettre à disposition des autres agents du système. Dans un tel système, les agents sont à la fois consommateurs et fournisseurs de services. Les réseaux pair-à-pair de partage de fichiers tels que Gnutella en sont un exemple [Adar et Huberman, 2000].

Afin de rester génériques, nous considérons les services comme un ensemble d'entités abstraites $S = \{s_1, \ldots, s_m\}$, chaque agent $a_k \in N$ pouvant en fournir un sous-ensemble S_k. Dans la suite, nous distinguons un agent qui reçoit un service en le désignant comme l'agent a_i d'un agent qui fournit un service en le désignant comme l'agent a_k.

Définition 6.1.1 - *Système d'échange de services* : Un *système d'échange de service* est

un couple $\langle N,S\rangle$ où N est l'ensemble des agents du système et S l'ensemble des services qui peuvent être fournis.

Nous considérons que chaque agent $a_k \in N$ peut fournir un sous-ensemble $S_k \subseteq S$ (potentiellement vide) de services et que chaque service $s_x \in S$ peut être fourni par un sous-ensemble $N_x \subseteq N$ (non vide) d'agents du système. Nous considérons aussi qu'un agent $a_i \in N$ interagit avec l'agent $a_k \in N_x$ lorsque l'agent a_i demande à l'agent a_k de lui fournir le service $s_x \in S$.

Exemple 6.1.1 - Pour illustrer nos propos tout au long de ce chapitre, nous considérons un scénario où les agents a_1,a_2,a_3,a_4,a_5 peuvent fournir différentes fonctionnalités présentées dans le tableau 6.2.

	a_1	a_2	a_3	a_4	a_5
s_1	√	√	×	×	×
s_2	×	√	×	×	√
s_3	√	×	×	√	√

Tableau 6.2 – Ensembles des services fournis par les différents agents du système

Si plusieurs agents peuvent proposer le même service $s_x \in S$, la capacité à le fournir avec qualité peut varier d'un fournisseur à un autre. Par ailleurs, pour un même agent, la qualité du service s_x peut être influencée par divers facteurs tels que le niveau d'utilisation de ces ressources propres lorsque le service lui est demandé. La qualité d'un service peut alors être considérée d'un point de vue utilitariste et peut correspondre au gain apporté à l'agent a_i lorsqu'il le reçoit. Pour capturer les différents paramètres influençant la qualité d'un service et considérer une variance de qualité, nous considérons que la qualité d'un service s_x fourni par l'agent $a_k \in N_x$ soit définie par une fonction de probabilité $\theta_{k,x}$ (définie sur un domaine $\mathcal{D}_x$ commun à l'ensemble des agents). Cette fonction de probabilité étant propre à chaque agent, nous considérons un facteur d'expertise désignant l'espérance mathématique de la qualité de ce service.

Définition 6.1.2 - *Expertise* : L'*expertise* de l'agent $a_k \in N$ pour le service $s_x \in S_k$, notée $\varepsilon_{k,x}$, est la valeur espérée de la qualité lorsque a_k fournit ce service à un autre agent.

Exemple 6.1.2 - Reprenons l'exemple 6.1.1 et considérons les facteurs d'expertise donnés dans le tableau 6.3. Par simplification, nous supposons ici que la qualité des différents services est définie par une valeur booléenne. Ainsi, la valeur d'expertise $\varepsilon_{2,2} = 0.8$ signifie que dans 4 cas sur 5 l'agent a_2 effectue correctement le service s_2 qui lui est demandé. Par abus de notation, nous désignons par $\emptyset$ l'expertise d'un agent a_i ne pouvant par fournir le service s_x.

	a_1	a_2	a_3	a_4	a_5
s_1	0.4	0.3	$\emptyset$	$\emptyset$	$\emptyset$
s_2	$\emptyset$	0.8	$\emptyset$	$\emptyset$	0.75
s_3	0.9	$\emptyset$	$\emptyset$	0.95	0.6

Tableau 6.3 – Expertises des agents selon les différents services proposés

Pour un même service, plusieurs agents peuvent évaluer sa qualité selon divers critères. Par exemple, dans le cadre d'un système de partage de fichiers, deux agents a_{i1} et a_{i2} peuvent évaluer la qualité d'un même fichier sur différents critères tels que le délai de réception ou son niveau de compression. Pour modéliser ces différences d'évaluations de la qualité, nous considérons que chaque agent a_i dispose d'une fonction d'évaluation lui permettant de mesurer la qualité du service selon ses propres critères.

Définition 6.1.3 - *Fonction d'évaluation* : À l'instant t, la qualité du service s_x fourni par l'agent a_k à l'agent a_i est mesurée par la *fonction d'évaluation* de a_i : $v_i(a_k,s_x,t) \in \mathcal{D}_x$.

Par simplicité, nous notons par $v^t_{i,k,x}$ la qualité du service s_x fourni par l'agent a_k à l'agent a_i à l'instant t.

Exemple 6.1.3 - Considérons les agents a_1, a_2 et le service s_2. Supposons que a_1 a demandé à 6 reprises le service s_2 à l'agent a_2. Le tableau 6.4 représente les évaluations de a_1 où ce dernier a considéré qu'aux temps t_3 et t_5 le service s_2 n'as pas été correctement fourni par a_2.

t	t_1	t_2	t_3	t_4	t_5	t_6
$v^t_{1,2,2}$	1	1	0	1	0	1

Tableau 6.4 – Évaluation de a_1 des services s_2 fournis par a_2

Les interactions entre un $a_i \in N$ et un agent $a_k \in N_x$ se résument en 4 étapes présentées dans la figure 6.1 :

1. a_i demande le service s_x dont il a besoin à a_k ;
2. a_k réalise s_x avec une qualité d'espérance $\varepsilon_{k,x}$;
3. a_k fournit s_x à a_i ;
4. a_i évalue la qualité de s_x via sa fonction d'évaluation v_i.

6.1.2 Un modèle générique de fonction de réputation

Comme nous l'avons présenté en section 2.2, les systèmes de réputation permettent aux agents d'estimer collectivement le comportement des autres agents lors de futures interactions. Dans notre modèle d'interaction, nous considérons l'évaluation de la qualité d'un service comme le gain que l'agent a_i estime avoir reçu en demandant ce service. Ainsi, le comportement des agents est représenté par cette estimation pour les interactions futures.

Comme nous considérons qu'un même service peut être fourni par plusieurs agents du système, nous intégrons au modèle d'interaction entre les agents un système de réputation aidant les agents à décider à quels agents $a_k \in N_x$ demander de réaliser le service s_x. Nous ne nous intéressons pas dans ce manuscrit à la définition d'une nouvelle fonction permettant aux agents de calculer les valeurs de réputation mais considérons que les agents disposent de l'une des nombreuses fonctions de réputation issue de la littérature [6]. Cette abstraction de la fonction de réputation nous permet de rester générique dans la description de notre modèle.

Comme l'ont fait remarquer [Resnick *et al.*, 2000], les résultats des interactions passées doivent être pris en compte dans le calcul de la valeur de confiance. Pour modéliser cela, nous considérons que les agents conservent les évaluations des interactions.

6. Se référer à la section 2.2 pour une description des diverses fonctions de réputation.

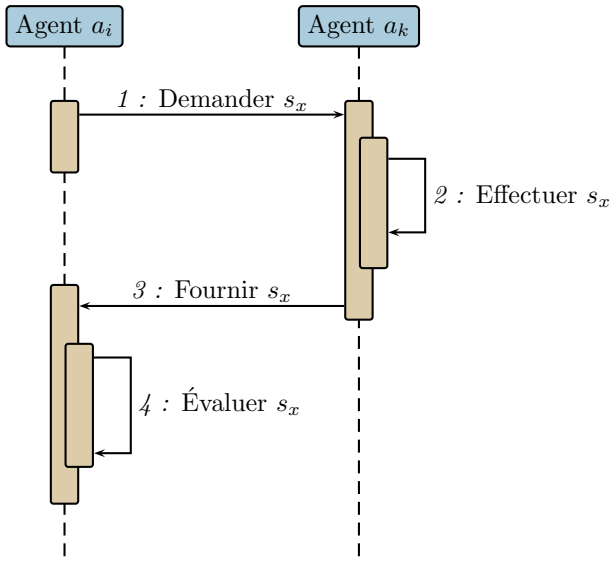

FIGURE 6.1 – Diagramme de séquence d'une interaction entre a_i et a_k

Définition 6.1.4 - *Observations* : Les *observations* de l'agent $a_i \in N$ vis-à-vis de l'agent $a_k \in N$ pour le service $s_x \in S_k$ désignent les évaluations de toutes les m interactions passées entre a_i et a_k pour le service s_x :

$$O_{i,k,x} = \{v_{i,k,x}^{t_1}, \ldots, v_{i,k,x}^{t_m}\}$$

Exemple 6.1.4 - Dans l'exemple 6.1.3, l'ensemble des observations de a_1 pour le service s_2 fourni par a_2 est $O_{1,2,2} = \{\ 1,1,0,1,0,1\ \}$.

À partir de ses observations, l'agent a_i peut calculer deux valeurs : le gain qu'il a obtenu à participer au système et la confiance qu'il a envers l'agent a_k pour fournir le service s_x.

Définition 6.1.5 - *Gains* : Le *gain* de l'agent $a_i \in N$ à l'instant t est défini par :

$$g_i^t = \sum_{v_{i,k,x}^{t'} \in O_i} v_{i,k,x}^{t'} \text{ où } O_i = \bigcup_{s_x \in S, a_k \in N_x} O_{i,k,x}$$

Définition 6.1.6 - *Confiance* : La *confiance* de l'agent $a_i \in N$ vis-à-vis de l'agent $a_k \in N$ pour le service $s_x \in S_k$ est l'estimation $c_{i,k,x}$ de l'expertise de a_k pour le s_x fondé sur $O_{i,k,x}$.

Comme nous considérons la fonction de réputation comme abstraite, nous ne donnons pas ici de domaine de définition de la confiance. Celle-ci peut être définie comme dans EigenTrust [Kamvar *et al.*, 2003] par le nombre d'interactions satisfaisantes moins le nombre d'interactions non satisfaisantes, par un tuple $\langle r,s \rangle$ comme dans BetaReputation [Jøsang et Ismail, 2002] ou même par de la logique floue comme dans Repage [Sabater *et al.*, 2006].

Exemple 6.1.5 - Supposons que l'agent a_1 définisse la confiance comme la moyenne des qualités observées. La confiance de a_1 envers a_2 pour le service s_2 est alors :

$$\begin{aligned} c_{1,2,2} &= (1+1+0+1+0+1)/(6) \\ &= 2/3 \end{aligned}$$

Pour calculer la réputation de l'agent a_k, les agents du système partagent leurs observations. Comme pour la confiance, la représentation des témoignages dépend du modèle de réputation considéré. Par exemple, nous pouvons considérer que les agents partagent directement le résultat de leurs interactions avec les autres agents du système ou que les agents partagent leurs valeurs de confiance normalisées comme dans EigenTrust.

Définition 6.1.7 - *Témoignage* : Un *témoignage* de l'agent $a_j \in N$ à $a_i \in N$ désigne un ensemble d'informations $F_{i,j,k,x}$ que a_j partage avec a_i vis-à-vis de $a_k \in N$ pour le service $s_x \in S_k$.

Dans la suite, nous désignons par a_j un agent qui fournit un témoignage. Comme à partir des observations d'un agent il est possible de construire les valeurs de confiance, nous considérons ici que les agents s'échangent un ensemble de valeurs qu'ils déclarent être leurs observations. Nous ne nous intéressons pas dans ce manuscrit au protocole utilisé par les agents pour l'échange des témoignages et considérons qu'un agent a_i peut à tout instant demander à un agent a_j de lui fournir un témoignage.

Exemple 6.1.6 - Supposons que a_1 partage avec a_3 ses observations portant sur la capacité de a_2 à effectuer s_2. a_3 reçoit alors le témoignage $F_{3,1,2,2} = \{\ 1,1,0,1,0,1\ \}$.

La réputation d'un agent étant une estimation de son comportement fondé sur les observations et les témoignages reçus, nous désignons par $\mathcal{F}_i$ la matrice $N \times N \times S$ telle que nous avons $\forall a_k \in N, \forall s_x \in S$:

$$\begin{aligned} \mathcal{F}_i[i][k][x] &= O_{i,k,x} \\ \forall a_j \in N \setminus \{a_i, a_k\} : \mathcal{F}_i[j][k][x] &= F_{i,j,k,x} \end{aligned}$$

À partir de $\mathcal{F}_i$, l'agent a_i peut calculer une estimation collective de l'expertise de a_k pour le service s_x.

Définition 6.1.8 - *Réputation* : La *réputation* de l'agent $a_k \in N$ pour le service $s_x \in S_k$ du point de vue de l'agent a_i est une estimation collective de $\varepsilon_{k,x}$ calculée par une *fonction de réputation* $f_i : N \times S \times 2^{\mathcal{F}} \rightarrow \mathcal{R}_i$ où par abus de notation $2^{\mathcal{F}}$ désigne l'ensemble des matrices de $\mathcal{F}_i$ possibles et $\mathcal{R}_i$ le domaine de définition des valeurs de réputation.

Cette définition abstraite permet de considérer de multiples fonctions de réputation spécifiques. Celles-ci peuvent autant être globales que personnalisées, dépendantes ou indépendantes du contexte. Nous considérons cependant que le système est décentralisé et qu'il n'existe pas d'autorité centrale collectant les témoignages chargée de calculer la réputation des agents.

Exemple 6.1.7 - Supposons que l'agent a_1 calcule la réputation de a_2 en effectuant la moyenne des valeurs de confiance. Supposons que pour le service s_2, a_1 a reçu les témoignages suivants vis-à-vis de a_2 :

$$F_{1,3,2,2} = \{\ 0{,}1{,}1\ \} \quad F_{1,4,2,2} = \{\ 1{,}1\ \} \quad F_{1,5,2,2} = \{\ 1{,}1{,}1{,}1\ \}$$

À partir de ces témoignages, a_1 peut calculer les valeurs de confiance a_3, a_4 et a_5 ainsi que la réputation de a_2 pour le service s_2 :

$$\begin{array}{rcc} c_{1,2,2} & = & 2/3 \\ c_{3,2,2} & = & 2/3 \\ c_{4,2,2} & = & 1 \\ c_{5,2,2} & = & 1 \\ \hline f_1(a_2,s_2,\mathcal{F}_i) & = & 5/6 \end{array}$$

Pour décider avec qui interagir, nous considérons que les agents disposent d'une *politique de sélection* π_i. Intuitivement, la politique de sélection est le processus par lequel l'agent a_i décide à quel agent $a_k \in N$ demander le service s_x.

La figure 6.2 résume ce modèle. Les arcs en pointillés correspondent à des échanges entre l'agent a_i et un autre agent du système. Lorsque l'agent $a_i \in N$ a besoin d'un service $s_x \in S$, il détermine à quel agent $a_k \in N_x$ le demander en suivant sa politique de sélection π_i. Il demande alors ce service s_x à l'agent a_k (flèche 1). Lorsqu'il reçoit le service (flèche 2), il en évalue la qualité, mets à jour ses observations puis recalcule les valeurs de réputation. Les flèches 3 et 4 désignent respectivement un témoignage reçu par un agent a_j et un témoignage que fourni a_i à un autre agent du système. Enfin, les flèches 5 et 6 correspondent à une demande de service reçu par a_i et ce même service une fois que a_i l'a effectué.

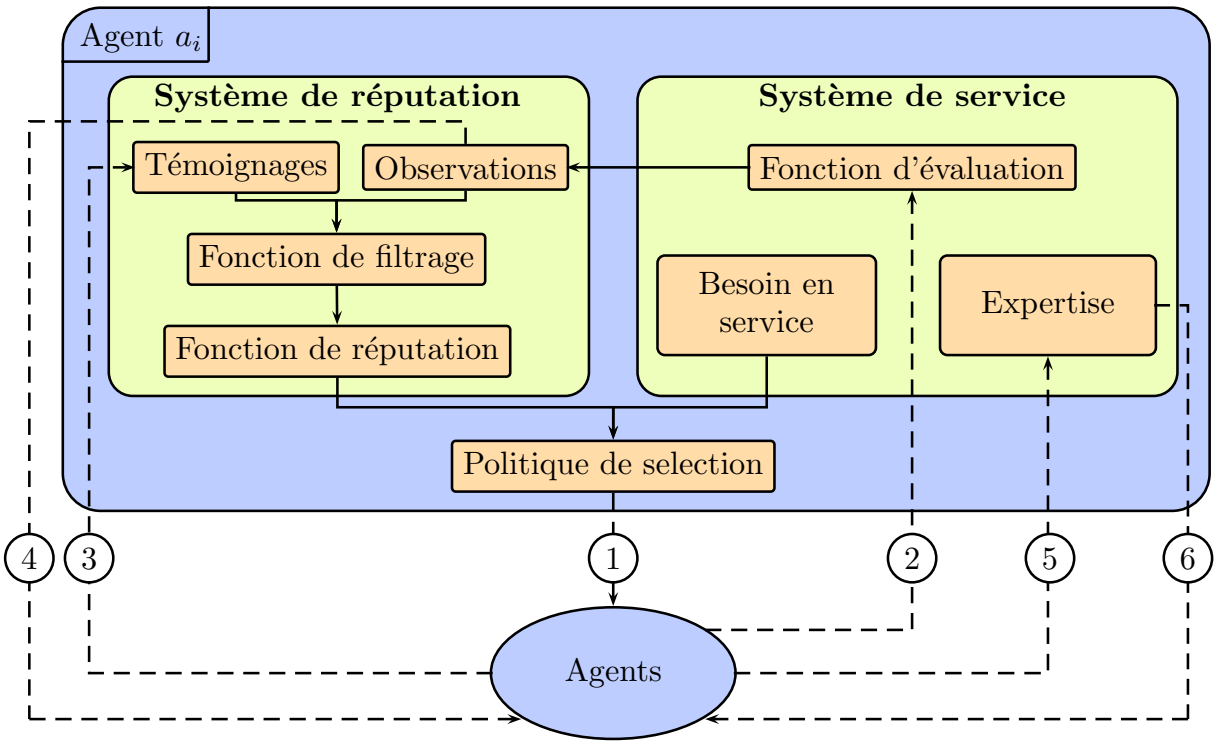

FIGURE 6.2 – Modélisation des interactions entre agents

6.2 Manipulation de la réputation

Dans cette section, nous présentons les différentes manipulations que nous considérons dans le contexte des systèmes de réputation.

6.2.1 Des agents malveillants en collusion

Comme une interaction entre un agent a_i et un agent a_k correspond au fait que a_i délégue à a_k la réalisation d'une tâche, nous représentons par l'expertise la capacité des différents agents du système à fournir les services demandés avec une bonne qualité. Comme nous considérons des agents rationels, nous faisons l'hypothèse que les agents cherchent à recevoir les services avec la meilleure qualité possible.

Définition 6.2.1 - *Agent honnête* : Dans un système d'échange de services $\langle N,S\rangle$, un *agent honnête* $a_i \in N$ est un agent dont l'objectif est de maximiser la qualité des services qu'il demande.

La définition de la malveillance (définition 1.2.4) nous permet de considérer l'objectif des agents malveillants en opposition à celui des agents honnêtes.

Définition 6.2.2 - *Agent malveillant* : Dans un système d'échange de services $\langle N,S\rangle$, un *agent malveillant* est un agent $a_k \in N$ dont l'objectif est de minimiser la qualité des services reçus par les autres agents du système. Pour atteindre son objectif, un agent malveillant a_k peut présenter deux comportements lorsqu'un agent a_i lui demande un service s_x :

- un comportement *honnête* si la qualité de s_x dépend de son expertise $\varepsilon_{k,x}$;
- un comportement *malveillant* si a_k fournit sciemment s_x avec une expertise $\overline{\varepsilon_{k,x}} < \varepsilon_{k,x}$.

Un exemple classique de comportement malveillant dans les réseaux pairs à pairs est celui de la diffusion de virus. De plus, si un agent malveillant peut être seul dans le système et chercher à manipuler l'ensemble des agents du système, il peut également être en collusion (définition 1.2.5) avec d'autres agents malveillants. Nous nous intéressons ici à une collusion d'agents malveillants $M \subset N$ cherchant collectivement à réduire le gain des autres agents du système.

Hypothèse 6.2.1 - : Les agents malveillants de M cherchent à réduire le gain de tout agent $a_i \in N \setminus M$ sans faire de discrimination.

Exemple 6.2.1 - Reprenons l'exemple 6.1.1. Supposons que les agents a_4 et a_5 sont malveillants et forment la collusion M. Lorsque l'agent a_4 fournit le service s_3 à l'agent a_2, il peut :

- interagir honnêtement en fonction de son facteur d'expertise $\varepsilon_{4,3} = 0.95$;
- être malveillant en fournissant volontairement s_3 avec une expertise $\overline{\varepsilon_{4,3}} = 0$.

Si l'utilisation du système de réputation permet aux agents honnêtes de détecter les comportements malveillants et de ne plus interagir avec les agents qui en sont responsables, ces derniers peuvent recourir à diverses manipulations afin de tromper le système de réputation.

6.2.2 Manipulations considérées

Les premières manipulations que nous considérons ici sont des manipulations explicites portant sur les informations partagées par les agents malveillants. Comme nous nous plaçons dans un

système multi-agents ouvert décentralisé, nous considérons que les agents malveillants peuvent introduire dans le système des agents Sybil. Contrairement à [Cheng et Friedman, 2005] qui considèrent l'utilisation des agents Sybil uniquement d'un point de vue statique, nous considérons les agents Sybil dans leur dynamique : ils peuvent interagir et rendre des services aux agents honnêtes.

Définition 6.2.3 - *Agent Sybil* : Un agent malveillant a_i effectue une attaque Sybil en introduisant un nouvel agent a_s dans le système d'échange de service.

Nous considérons aussi du blanchiment permettant de *réinitialiser* les observations des agents honnêtes.

Définition 6.2.4 - *Blanchiment* : Un agent malveillant $a_i \in M$ effectue du blanchiment en quittant le système et en le réintégrant sous une nouvelle identité $a_{i'}$ telle que pour $\forall a_j \in N \setminus M, \forall s_x \in S_{i'}$, $O_{j,i',x} = \emptyset$.

Exemple 6.2.2 - Dans l'exemple 6.1.1, l'agent malveillant a_4 peut intégrer un agent Sybil a_6 afin d'augmenter le nombre d'agents malveillants présents dans le système. L'agent a_5 peut effectuer du blanchiment en quittant le système est en le réintégrant sous une autre identité $a_{5'}$. Si l'agent a_3 a l'ensemble d'observations $O_{3,5,2} = \{1,0,0,0,0\}$, après blanchiment, a_3 considère $a_{5'}$ comme un nouvel agent avec $O_{3,5',2} = \emptyset$.

Si les agents malveillants peuvent mentir sur leurs identités, nous considérons également l'une des principales manipulations dans les systèmes de réputation : les faux témoignages. L'objectif de cette manipulation est de partager avec les autres agents du système des informations que l'on sait être fausses afin que la fonction de réputation calcule des valeurs à l'avantage d'au moins l'un des agents malveillants.

Définition 6.2.5 - *Faux témoignages* : Un agent malveillant $a_j \in M$ ayant eu les observations $O_{j,k,x}$ fournit un faux témoignage en partageant avec un agent $a_i \in N \setminus M$ le témoignage $F_{i,j,k,x}$ tel que $F_{i,j,k,x} \neq O_{j,k,x}$.

En opposition aux faux témoignages que peuvent fournir les agents malveillants, nous considérons que les agents honnêtes partagent toujours leurs véritables observations.

Hypothèse 6.2.2 - : Lorsqu'un agent $a_j \in N \setminus M$ fournit un témoignage à l'agent $a_i \in N$ vis-à-vis de $a_k \in N$ pour le service $s_x \in S_k$, celui-ci est toujours *honnête* : $F_{i,j,k,x} = O_{j,k,x}$.

Nous distinguons deux catégories de faux témoignages : la *promotion* visant à augmenter la valeur de réputation d'un agent malveillant $a_k \in M$ et la *diffamation* visant à réduire la réputation d'un agent honnête $a_k \in N \setminus M$.

Définition 6.2.6 - *Promotion* : L'agent malveillant $a_j \in M$ *promeut* l'agent malveillant $a_k \in M$ en fournissant à tout agent $a_i \in N \setminus M$ un faux témoignage $F_{i,j,k,x}$ tel que :

$$\sum_{v^t_{j,k,x} \in F_{i,j,k,x}} v^t_{j,k,x} > \sum_{v^t_{j,k,x} \in O_{j,k,x}} v^t_{j,k,x}$$

Définition 6.2.7 - *Diffamation* : L'agent malveillant $a_j \in M$ *diffame* l'agent honnête $a_k \in N \setminus M$ en fournissant à tout agent $a_i \in N \setminus M$ un faux témoignage $F_{i,j,k,x}$ tel que :

$$\sum_{v^t_{j,k,x} \in F_{i,j,k,x}} v^t_{j,k,x} < \sum_{v^t_{j,k,x} \in O_{j,k,x}} v^t_{j,k,x}$$

Exemple 6.2.3 - Le tableau 6.5 présente deux exemples de faux témoignages fournis par les agents malveillants. Dans le premier cas, l'agent a_4 diffame l'agent a_2 pour le service s_2. Dans le second cas, a_4 promeut l'agent a_5 vis-à-vis du service s_3.

a_k	s_x	$O_{4,k,x}$	$F_{1,4,k,x}$
a_2	s_2	{ 1,1,1,1,0 }	{0,0,0,1}
a_5	s_3	$\emptyset$	{ 1,1,1 }

Tableau 6.5 – Exemples de faux témoignages fournis par a_4 à a_1

L'agent malveillant a_6 effectue de la promotion vis-à-vis de a_4 pour le service s_3 en fournissant à l'agent a_1 le témoignage $F_{1,6,4,3} = \{1,1,1,1\}$ alors qu'il n'a pas interagi avec a_4. Supposons que a_4 a les observations $O_{4,1,3} = \{ 1,1,1 \}$. Il effectue une diffamation vis-à-vis de a_1 pour le service s_3 en partagent avec a_2 le témoignage $F_{2,4,1,3} = \{ 0,0,0 \}$.

La troisième catégorie de manipulation que nous considérons dans notre système d'échange de service est une manipulation comportementale : la traîtrise. Il s'agit pour l'agent malveillant de présenter lors de ses premières interactions un comportement honnête puis de changer de comportement après un certain temps.

Définition 6.2.8 - *Traîtrise* : Un agent malveillant $a_k \in N$ effectue une *traîtrise* s'il présente dans un premier temps un comportement honnête puis dans une seconde phase fournie volontairement ses services avec une mauvaise qualité.

Notons que nous ne définissons pas ici l'instant où l'agent malveillant passe du comportement honnête au comportement malveillant. Cet instant peut être prédéfini ou être calculé dynamiquement en fonction de la valeur de réputation de l'agent malveillant. De même, dans le cas d'une fonction de réputation dépendante du contexte, l'agent malveillant peut présenter un comportement honnête pour un service s_1 et un comportement malveillant pour un service s_2. Cette manipulation s'attaque à l'hypothèse généralement faite que le comportement des agents reste le même au cours du temps et que la qualité des services fournis ne varie que légèrement.

Exemple 6.2.4 - Considérons que l'agent malveillant a_4 présente entre $t = 0$ et $t = t^*$ un comportement honnête (en fournissant le service avec une qualité moyenne $\varepsilon_{4,3} = 0.95$) puis effectue une traîtrise à l'instant $t = t^*$. La figure 6.3 illustre la qualité moyenne des services s_3 fournis par a_4 aux autres agents du système en fonction de l'instant où il lui est demandé. La courbe bleue représente l'évolution de la réputation de a_4. La zone hachurée correspond à l'intervalle de temps durant lesquels l'agent malveillant présente un mauvais comportement, mais que sa valeur de réputation reste suffisamment élevée pour interagir.

La dernière manipulation que nous considérons est l'*attaque oscillante* qui est agrégation des différentes manipulations présentées ci-dessus.

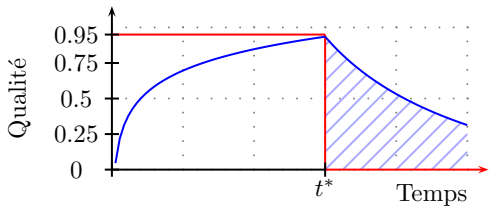

FIGURE 6.3 – Qualité moyenne du service s_3 fourni par a_4 au cours du temps

Définition 6.2.9 - *Attaque oscillante* : Une collusion M d'agents malveillants effectue une *attaque oscillante* en se partitionnant en deux sous-groupes M_1 et M_2 où les agents de chaque groupe jouent un rôle différent :

- les agents de M_1 ont un comportement honnête et promeuvent les agents de M_2 ;
- les agents de M_2 ont un comportement malveillant et diffament les agents de $N \setminus M$.

Lorsque la réputation d'un agent de M_2 est inférieure à la réputation de la majorité des agents honnêtes :

- cet agent de M_2 se blanchit et sa nouvelle identité rejoint M_1 ;
- le plus ancien agent de M_1 rejoint M_2.

Exemple 6.2.5 - Considérons un système d'échange de service où $M = \{a_4,a_5,a_6,a_7\}$. Les agents malveillants peuvent mettre en œuvre une attaque oscillante en se répartissant dans deux groupes : $M_1 = \{a_4,a_5\}$ et $M_2 = \{a_6,a_7\}$. La figure 6.4 illustre les oscillations du comportement des agents malveillants. En rouge est représenté le comportement des agents de M_1 et en bleu le comportement des agents de M_2.

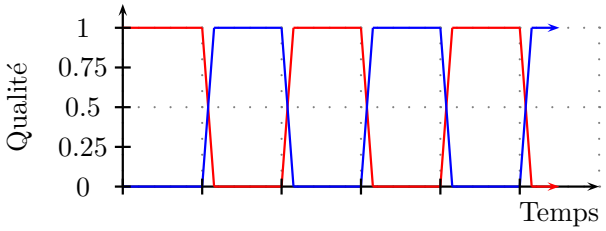

FIGURE 6.4 – Comportement des agents malveillants effectuant une attaque oscillante

Dans la suite de ce manuscrit, nous considérons des agents malveillants mettant en œuvre des attaques oscillantes. Cette manipulation étant une agrégation des principaux types de manipulations, il est plus difficile d'y être robuste, de la détecter ou de la prévenir.

6.3 Conclusion

Dans ce chapitre, nous avons présenté un modèle générique d'interaction entre agents dans un système ouvert et décentralisé sous forme d'échange de services. Une interaction est représentée par la réalisation d'un service s_x par un agent a_k pour un agent a_i. Dans notre modèle, la capacité des agents à fournir des services de bonne qualité est définie par un facteur d'*expertise*. La qualité d'un service représente le gain que perçoit a_i après interaction avec a_k. Pour estimer

collectivement l'expertise des différents agents, nous considérons un *système de réputation*. Les agents du système s'échangent par le biais de *témoignages* leurs différentes observations puis les agrègent via une fonction de réputation.

Ce modèle d'interaction permet à des agents honnêtes de maximiser la qualité des services qu'ils reçoivent. Cependant, nous considérons dans ce système une seconde catégorie d'agents : des agents malveillants qui par opposition aux agents honnêtes cherchent à minimiser les gains totaux des autres agents en fournissant volontairement de mauvais services. Nous avons présenté une modélisation des principales manipulations que peut réaliser une collusion d'agents malveillants afin d'atteindre leurs objectifs :

- le blanchiment, consistant à changer d'identité pour réinitialiser les connaissances des autres agents ;
- les faux témoignages (promotion et diffamations), consistant à partager avec les autres agents de fausses observations ;
- la traîtrise, consistant à changer de comportement au cours du temps.

Dans la suite de ce manuscrit, nous considérons des agents malveillants menant une *attaque oscillante* qui combinent toute ses manipulations consistant à alterner entre deux blanchiments bons et mauvais comportements tout en effectuant de la promotion et de la diffamation. Nous proposons deux approches pour lutter contre ces agents malveillants. La première consiste à choisir avec discernement la politique de sélection qui spécifie comment utiliser la réputation des agents dans le processus de décision. La seconde approche est destinée à limiter l'influence des faux témoignages en filtrant les témoignages en fonction d'une mesure de crédibilité.

Chapitre 7

Des politiques de bandits manchots

Sommaire

Résumé.
Dans le chapitre précédent, nous avons proposé un modèle générique d'interaction où les agents utilisent un système de réputation pour estimer la qualité de leurs futures interactions. Dans ce chapitre, nous nous intéressons à l'utilisation de la valeur de réputation dans le processus de décision des agents, et plus précisément à l'influence de cette utilisation sur la robustesse du système de réputation. L'approche que nous proposons d'étudier ici a fait l'objet de deux publications [Vallée *et al.*, 2014a, Vallée *et al.*, 2014b]. Dans cette approche, nous abordons la problématique de la confiance et de la réputation comme un problème d'apprentissage par renforcement et proposons d'utiliser le modèle des *bandits manchots*. Dans la première section de ce chapitre, nous présentons de manière canonique les bandits manchots et le problème de décision associé. Nous faisons ensuite une analogie entre notre modèle générique d'interaction et les modèles de bandits manchots. Dans la suite, nous présentons comment nous pouvons utiliser les politiques de sélection issue de la littérature des bandits manchots dans le contexte des systèmes de réputation. Pour finir, nous étudions empiriquement comment l'utilisation de ces différentes politiques de sélection influence la robustesse d'un tel système.

Afin de simplifier la lecture de ce chapitre, le tableau 7.1 résume les principales notations utilisées.

Bandits manchots	
$B = \{b_1, \ldots, b_k\}$	Ensemble des bras d'un bandit manchot
θ_i	Distribution de probabilité de la récompense associée aux bras b_i
$b_i^t \in B$	Bras sélectionné à l'instant t
π_i	Politique de sélection de l'agent a_i
r_i^t	Regret de l'agent a_i à l'instant t

Tableau 7.1 – Principales notations de l'analogie avec les bandits manchots

7.1 Système de réputation et bandits manchots

7.1.1 Bandits manchots, un problème de décision

L'un des problèmes classiques de prise de décision séquentielle utilisant le résultat des observations passées est celui des *bandits manchots* (ou Multi-Armed Bandits) [Robbins, 1952, Anantharam *et al.*, 1987, Katehakis et Veinott Jr, 1987, Auer *et al.*, 1995, Agrawal, 1995, Auer *et al.*, 2002, Vermorel et Mohri, 2005, Leskovec *et al.*, 2008, Auer et Ortner, 2010]. Il s'agit pour un joueur de casino (représentant l'agent) de décider parmi un ensemble de machines à sous laquelle utiliser afin de maximiser son gain.

Le modèle canonique d'un problème de bandits manchots est un ensemble de bras $B = \{b_1, \ldots, b_k\}$ où chaque bras représente une machine à sous [Robbins, 1952]. À chaque instant t, l'agent doit décider quel bras $b_i^t \in B$ tirer (utiliser la machine à sous correspondante) pour maximiser son gain. À chaque bras $b_i \in B$ est associée une distribution de probabilité θ_i de paramètres inconnus. Lorsqu'à l'instant t, l'agent a tire le bras b_i, il reçoit une récompense $g_i^t \in \mathbb{R}$ selon une *fonction de récompense* qui suit la distribution θ_i.

De nombreuses variantes de modèles de bandits manchots ont été étudiées dans la littérature et permettent de considérer :

- des fonctions de récompenses non stationnaires qui varient au cours du temps [Gittins, 1979, Ishikida et Varaiya, 1994, Koulouriotis et Xanthopoulos, 2008, Bubeck et Cesa-Bianchi, 2012] ;
- plusieurs agents tirant simultanément des bras et se partageant éventuellement les récompenses [Anantharam *et al.*, 1987, Liu et Zhao, 2010] ;
- des adversaires qui choisissent les récompenses de chaque bras à chaque pas de temps avant qu'ils ne soient tirés [Auer *et al.*, 1995] ;
- un agent qui ne peut tirer qu'un nombre limité de bras [Tran-Thanh *et al.*, 2010].

Pour décider quels bras un agent doit tirer, de nombreux travaux s'intéressent à la définition d'une *politique de sélection* permettant à l'agent de maximiser son gain [Whittle, 1980, Katehakis et Veinott Jr, 1987, Auer *et al.*, 1995, Agrawal, 1995, Burnetas et Katehakis, 1997, Auer *et al.*, 2002, Vermorel et Mohri, 2005, Leskovec *et al.*, 2008, Audibert *et al.*, 2009, Auer et Ortner, 2010]. De manière générique, nous définissons une politique de bandit manchot de la manière suivante :

Définition 7.1.1 - *Politiques de sélection d'un bandit manchot* : La *politique de sélection* de l'agent a est une fonction $\pi : B^{t-1} \times \mathcal{R}^{t-1} \to B$ qui définit quel bras de B l'agent a doit tirer à l'instant t en se fondant sur les récompenses observées lors des $t-1$ actions précédentes.

Dans le cas d'un agent omniscient qui connaîtrait les espérances de récompenses de chaque bras, la politique optimale consiste à tirer à chaque pas de temps le bras, dont la fonction qui maximise cette espérance. Cependant, comme l'agent ne connaît a priori pas les paramètres des distributions de récompense, il lui faut donc utiliser les résultats de ses observations passées pour estimer cette espérance. Pour faire cette estimation tout en maximisant son gain en parallèle, les politiques de sélection proposent des compromis entre l'*exploration* qui consistent à tirer les différents bras pour améliorer ses connaissances et l'*exploitation* où l'agent tire le meilleur bras selon ses connaissances courantes.

Parmi les nombreuses politiques, nous pouvons notamment considérer :

- les *politiques non contextuelles*, c'est-à-dire qui sélectionnent un bras en se fondant uniquement sur son espérance de récompense et en acceptant parfois de ne pas sélectionner le meilleur afin d'obtenir plus d'information. Par exemple, la famille des politiques ε-gloutonnes consiste à tirer le bras ayant la meilleure espérance estimée avec une probabilité $1-\varepsilon$ et à tirer un autre bras (choisi aléatoirement uniformément parmi tous les bras) avec une probabilité ε [Tran-Thanh *et al.*, 2010,Kuleshov et Precup, 2014]. Un exemple est la famille des politiques élitistes qui sélectionnent le bras à tirer parmi un sous-ensemble de bras ayant une bonne espérance de gains ou la politique Poker qui sélectionne le bras qui maximise la récompense obtenue plus son espérance de récompense [Bubeck et Cesa-Bianchi, 2012] ;
- les *politiques contextuelles*, c'est-à-dire qui sélectionnent explicitement un bras en fonction d'un compromis entre son espérance de récompense et une mesure de l'information que possède l'agent sur ce bras. Par exemple, la famille des politiques *Exp3*[7] [Auer *et al.*, 2002] sélectionne le bras $b_i \in B$ une probabilité :

$$p_i(t+1) = (1-\gamma)\frac{w_i(t)}{\sum_{j=1}^{k} w_j(t)} + \frac{\gamma}{k}$$

où $\gamma \in (0,1]$ est une constance et les poids sont $w_j(t+1) = w_j(t)exp(\gamma\frac{g_j^t}{p_j(t)k})$ pour le bras b_j ayant été sélectionné à l'instant t et $w_j(t+1) = w_j(t)$ pour tous les autres bras. La principale famille de politiques contextuelles étudiées dans la littérature est *UCB*[8] [Agrawal, 1995, Auer et Ortner, 2010] qui consiste à sélectionner le bras qui maximise l'espérance de récompense (notée μ_i) plus un facteur d'exploration dépendant de n_i, le nombre de fois que le bras b_i déjà été tiré parmi tous les bras :

$$b_i^{t+1} = \underset{b_j \in B}{\text{argmax}}(\mu_j + \sqrt{\frac{2\ln t}{n_j}})$$

La principale propriété de ces politiques est qu'elles permettent à l'agent a de minimiser son *regret*, c'est-à-dire la différence entre le gain que l'agent aurait obtenu s'il avait suivi une politique de sélection optimale et le gain qu'il a réellement obtenu.

7.1.2 Analogie entre notre modèle d'interaction et les bandits manchots

Dans les systèmes de réputation, le problème de décision de chaque agent est similaire à celui d'un agent dans un problème de bandit manchot. En effet, étant donné un ensemble d'observations et un ensemble d'agents pouvant fournir le service $s_x \in S$, avec quel agent $a_k \in N_x$ doit-il interagir afin de maximiser la qualité de ce service ?

7. Exponential weight algorithm for Exploration and Exploitation.
8. Upper-Confidence Bound

Considérons un service $s_x \in S$. L'agent a_i peut modéliser les différents agents fournissant ce service par un système de bandit manchot, en associant à chaque agent $a_k \in N_x$ un bras $b_{k,x}$ dont la fonction de récompense suit une distribution de probabilité $\theta_{k,x}$ d'espérance $\varepsilon_{k,x}$ (a priori inconnue de a_i). La récompense observée par a_i lorsqu'il demande le service s_x à l'agent a_k est ainsi équivalent à la récompense observée par a_i s'il tire le bras $b_{k,x}$.

Contrairement aux modèles classiques de bandits manchots où l'agent est seul et utilise donc uniquement ces propres observations pour calculer l'espérance de récompense de chaque bras, les agents d'un système d'échange de service peuvent partager leurs observations afin de permettre aux autres agents de calculer une réputation. Contrairement aux bandits manchots où les observations directes guident la décision, c'est la réputation du fournisseur de service qui influe sur la décision d'interaction.

Cependant, les modèles de bandits manchots considèrent la récompense comme une valeur réelle. Si cela est aussi le cas dans de nombreux systèmes de réputation comme BetaReputation [Jøsang et Ismail, 2002], certains systèmes comme Repage [Sabater *et al.*, 2006] utilisent une réputation discrète. Nous pouvons ramener ce cas à une approche continue en considérant une fonction $g_i : \mathcal{R}_i \to \mathbb{R}$ associant à la valeur de réputation une valeur réelle. Pour deux agents a_{k_1} et a_{k_2}, si $g_i(f_i(a_{k_1},s_x,\mathcal{F}_i)) > g_i(f_i(a_{k_2},s_x,\mathcal{F}_i))$ alors la réputation de a_{k_1} doit être supérieure à la réputation de a_{k_2}.

Exemple 7.1.1 - Considérons un système d'échange de services où l'agent a_i utilise la fonction de réputation du système Repage [Sabater *et al.*, 2006] où la réputation des agents est définie sur un ensemble (w_1,w_2,w_3,w_4,w_5) représentant la probabilité que la future interaction se déroule respectivement « très mal », « mal », « moyennement », « bien »et « très bien ». Considérons les deux agents a_{k_1} et a_{k_2} tels que :

$$f_i(a_{k_1},s_x,\mathcal{F}_i) = (0.5,0.3,0.2,0.0,0.0)$$
$$f_i(a_{k_2},s_x,\mathcal{F}_i) = (0.1,0.1,0.2,0.3,0.3)$$

En associant la qualité des très mauvais services à une valeur de 1, de 2 pour les mauvais services et ainsi de suite, nous pouvons définir la fonction g_i suivante :

$$g_i(\ (w_1,w_2,w_3,w_4,w_5)\ ) = 1 \times w_1 + 2 \times w_2 + 3 \times w_4 + 4 \times w_5 + 5 \times w_5$$

L'application de ce morphisme donne :

$$\begin{aligned} g_i(f_i(a_{k_1},s_x,\mathcal{F}_i)) &= 0.5 + 0.6 + 0.6 + 0 + 0 \\ &= 1.7 \\ g_i(f_i(a_{k_2},s_x,\mathcal{F}_i)) &= 0.1 + 0.2 + 0.6 + 1.2 + 1.5 \\ &= 3.6 \end{aligned}$$

Ainsi, ce morphisme permet de considérer que l'espérance de récompense d'une interaction entre a_i et a_{k_1} pour le service s_x est de 1.7 (soit entre très mauvais et mauvais) ; tandis que ce même service fourni par a_{k_2} à une espérance de récompense de 3.6 (soit entre moyen et bon).

Par simplicité, nous considérons dans toute la suite que la fonction g_i est implicite et notons par $f_i(a_{k_1},s_x,\mathcal{F}_i)$ l'application de g_i sur la valeur de réputation. Nous faisons alors l'hypothèse que la réputation d'un agent est corrélée avec l'espérance de récompense du service fourni.

Hypothèse 7.1.1 - : Pour deux agents a_{k_1},$a_{k_2} \in N_x$, l'inégalité $f_i(a_{k_1},s_x,\mathcal{F}_i) > f_i(a_{k_2},s_x,\mathcal{F}_i)$ signifie que l'espérance de récompense de l'agent a_i en demandant s_x à l'agent a_{k_1} est supérieure à l'espérance de récompense lors d'une interaction avec l'agent a_{k_2}.

Le tableau 7.2 résume l'analogie que nous faisons entre notre modèle générique d'interaction et bandits manchots. Remarquons que, bien que nous associons dans cette analogie les agents malveillants aux problèmes des adversaires dans les bandits manchots, ces adversaires différents du modèle de [Auer *et al.*, 1995]. En effet, chaque agent étant associé à un sous-ensemble de bras correspondant aux services qu'ils peuvent fournir, l'espérance de récompense de chaque bras ne peut être fixée a priori que par l'agent correspondant. Par ailleurs, dans notre modèle, la récompense reçue lorsqu'un service est rendu ne peut être définie volontairement comme mauvaise qu'une fois que l'agent a_i a décidé avec qui interagir.

	Modèle générique d'interaction	**Bandit manchot**
Objectif	Maximiser la qualité des services reçus	Maximiser son gain
Acteurs	Consommateurs de services Fournisseurs de services Fournisseur de témoignages	Agents Bras Agents
Communication	Témoignages sur le comportement des autres agents	Témoignages sur les bras
Capacité	Expertise	Fonction de distribution des gains
Gains	Qualité des services	Récompense d'un bras
Observations	Évaluation de la qualité	Récompense reçue
Réputation	Comportement futur espéré	Récompense future espérée
Politique de sélection	Déterminer le futur fournisseur de service	Déterminer le futur bras à tirer
Manipulations	Agents malveillants	Adversaire

Tableau 7.2 – Analogie entre notre modèle générique d'interaction et bandits manchots

Cette analogie nous permet alors de considérer qu'un agent de notre modèle générique d'interaction peut utiliser les politiques de sélection des bandits manchots pour décider avec qui interagir.

7.2 Utilisations des politiques de sélection

7.2.1 Politiques de sélection considérées

Une politique de sélection est la procédure que suit un agent pour déterminer à quels agents demander le service dont il a besoin. Pour respecter le troisième axiome de [Resnick *et al.*, 2000], la politique de sélection doit être guidée par la valeur de réputation des agents.

Définition 7.2.1 - *Politique de sélection* : La *politique de sélection* de l'agent $a_i \in N$ (notée π_i) désigne la procédure par laquelle l'agent a_i décide à quel agent $a_k \in N_x$ demander le service $s_x \in S$.

Classiquement dans les systèmes de réputation, les agents choisissent l'agent ayant la meilleure valeur de réputation pour leur fournir le service s_x [Kamvar *et al.*, 2003,Whitby *et al.*, 2004,Malik et Bouguettaya, 2009, Su *et al.*, 2013]. Cette stratégie est appelée l'élitisme pur.

Définition 7.2.2 - *Élitisme pur* : Soit un agent $a_i \in N$ désirant recevoir le service $s_x \in S$. L'agent a_i suit une politique d'*élitisme pur* s'il le demande à l'agent $a_k \in N_x$ qui maximise $f_i(a_k,s_x,\mathcal{F}_i)$.

Exemple 7.2.1 - Considérons un système d'échange de services $\langle N,S\rangle$ avec l'agent $a_1 \in N$ désirant le service $s_1 \in S$ tel que $N_1 = \{a_2,a_3,a_4,a_5\}$. Supposons que la qualité du service s_1 soit définie sur l'ensemble $[-1,1]$ et que la réputation des agents soit définie par la qualité moyenne des services s_1 qu'ils ont fournis précédemment. Considérons enfin les valeurs de réputation suivantes :

a_k	a_2	a_3	a_4	a_5
$f_1(a_k,s_1,\mathcal{F}_i)$	0.85	0.3	-0.5	0

Si l'agent a_1 suit une politique d'élitisme pur, il doit alors demander le service s_1 à l'agent a_2.

Cette politique élitisme a intuitivement deux désavantages. Premièrement, elle ferme le système aux agents nouveaux entrants même s'ils ont un facteur d'expertise élevé. Par ailleurs, en cas de manipulation (promotion et diffamation), elle peut amener un agent à interagir continuellement avec les agents malveillants. Quand est-il des politiques présentant un compromis entre exploitation et exploration ? Pour répondre à cette question nous proposons d'étudier trois politiques de sélection : les politiques ε-élitistes, ε-gloutonnes et UCB.

La politique ε-élitiste consiste à sélectionner le fournisseur d'un service en le choisissant aléatoirement parmi un sous-ensemble des fournisseurs ayant la meilleure valeur de réputation.

Définition 7.2.3 - *ε-élitisme* : Soit un agent $a_i \in N$ désirant le service $s_x \in S$. Soit $\varepsilon \in [0,1]$ et $N_{x,\varepsilon} \subseteq N_x$ tels que $|N_{x,\varepsilon}| = \lceil \varepsilon \times |N_x| \rceil$ et que $\forall a_k \in N_{x,\varepsilon}, \nexists a_{k_2} \in N_x \setminus N_{x,\varepsilon} : f_i(a_k,s_x,\mathcal{F}_i) < f_i(a_{k_2},s_x,\mathcal{F}_i)$. L'agent a_i suit une politique *ε-élitiste* s'il choisit l'agent a_k à qui demander le service s_x en le sélectionnant aléatoirement uniformément parmi $N_{x,\varepsilon}$.

Exemple 7.2.2 - Reprenons l'exemple 7.2.1 où l'agent a_1 suit une politique 0.3-élitiste. Comme $\lceil 0.3 \times |N_1| \rceil = 2$, $N_{1,0.3} = \{a_2,a_3\}$. L'agent a_1 demande donc le service s_1 en sélectionnant respectivement avec une probabilité 0.5 et 0.5 les agents a_2 et l'agent a_3.

Cette politique ne garantit pas l'ouverture du système, car des agents nouveaux entrants ont peu de chance de faire partie du sous-ensemble des meilleurs fournissant sans avoir déjà interagi. Cependant, cette politique peut permettre d'éviter de toujours interagir avec un agent malveillant si ce dernier est efficacement promu par une collusion.

La politique ε-gloutonne que nous considérons est un représentant canonique de sa famille [Kuleshov et Precup, 2014].

Définition 7.2.4 - *ε-gloutonne* : Soit un agent $a_i \in N$ désirant le service $s_x \in S$ et $\varepsilon \in [0,1]$. L'agent a_i suit une politique *ε-gloutonne* si il demande le service à l'agent $a_k \in N_x$ qui :

- avec une probabilité $p = 1 - \varepsilon$, maximise $f_i(a_k,s_x,\mathcal{F}_i)$;
- avec une probabilité $p = \varepsilon$, est sélectionné aléatoirement uniformément parmi N_x.

Exemple 7.2.3 - Reprenons l'exemple 7.2.1 où l'agent a_1 suit une politique 0.2-élitiste. Ainsi, avec une probabilité $p = 0.8$ il demande à l'agent a_2 de lui servir le service s_1 ou, avec une probabilité $p = 0.2$, il sélectionne aléatoirement uniformément parmi $N_1 = \{a_2,a_3,a_4,a_5\}$ l'agent à qui demander le service s_1.

Cette politique de sélection est un bon compromis entre l'exploration et l'exploitation : elle garantit par ailleurs l'ouverture du système en permettant d'interagir occasionnellement avec de nouveaux agents tout en garantissant d'interagir avec le meilleur agent régulièrement. Cependant, elle peut amener à interagir avec des agents clairement identifiés comme fournissant de mauvais services.

Enfin, nous considérons la politique UCB qui tient compte dans une certaine mesure de l'historique des interactions entre agents.

Définition 7.2.5 - *UCB* : Soit un agent $a_i \in N$ désirant le service $s_x \in S$. Soit $n_{k,x}$ le nombre de fois que l'agent $a_k \in N_x$ a fourni s_x à a_i et n_x le nombre de fois que a_i a demandé à un agent quelconque de lui fournir s_x. L'agent a_i suit la politique de sélection *UCB* s'il demande le service s_x à l'agent a_k qui maximise :

$$f_i(a_k,s_x,\mathcal{F}_i) + \sqrt{\frac{2\ln(1+n_x)}{1+n_{k,x}}}$$

Exemple 7.2.4 - Reprenons l'exemple 7.2.1 où l'agent a_1 suivant la politique de sélection UCB. Soient les paramètres suivants :

a_k	a_2	a_3	a_4	a_5
(1) $f_1(a_k,s_1,\mathcal{F}_i)$	0.85	0.3	-0.5	0
$n_{k,1}$	10	2	2	0
(2) $\sqrt{\frac{2\ln(1+n_1)}{1+n_{k,1}}}$	0.702	1.344	1.344	2.327
(1)+(2)	1,542	1.644	0.844	2.327

En suivant la politique de sélection UCB, a_1 choisit a_5 comme fournisseur du service s_1 car il ne possède pas d'information sur ce dernier.

Cette politique offre un compromis a priori intéressant car elle permet à un agent de sélectionner occasionnellement les agents avec lesquels il a le moins interagi tout en sélectionnant régulièrement les agents les plus fiables.

7.2.2 Évaluer les politiques

L'objectif des agents honnêtes est de maximiser leurs gains, la somme des récompenses des services qu'ils reçoivent (définition 6.1.5). Ceci ne permet cependant pas de mesurer la qualité de la décision de l'agent. Pour cela, nous utilisons une mesure classique des bandits manchots : le regret. Il s'agit de la différence entre le gain qu'aurait obtenu l'agent s'il avait suivi une politique de sélection optimale et son gain réel.

Définition 7.2.6 - *Regret* : Soit $a_i \in N$ un agent ayant à l'instant $t_1 \in [1,t]$ demandé le service $s_x^{t_1} \in S$ à un agent $a_k^{t_1}$. Soit $N_x^{t_1}$ l'ensemble des agents pouvant fournir le service $s_x \in S$ au temps t_1. Soit $\varepsilon_{\star,x}^{t_1}$ l'expertise de l'agent de $N_x^{t_1}$ le plus performant. Le *regret* de l'agent a_i est défini par :

$$r_i^t = \sum_{t_1=1}^{t} \varepsilon_{\star,x}^{t_1} - v_{i,k,x}^{t_1} \text{ où } \varepsilon_{\star,x}^{t_1} = \max_{a_k \in N_x^{t_1}} \varepsilon_{k,x}$$

Exemple 7.2.5 - Considérons un système d'échange de service $\langle N,S \rangle$ avec : $N = \{a_1,a_2,a_3,a_4,a_5\}$ et $S = \{s_1,s_2,s_3\}$. Soit les facteurs d'expertise suivants :

	a_1	a_2	a_3	a_4	a_5
s_1	0.6	0.3	$\emptyset$	$\emptyset$	$\emptyset$
s_2	$\emptyset$	0.8	$\emptyset$	$\emptyset$	0.75
s_3	0.7	$\emptyset$	$\emptyset$	0.95	0.6

Le tableau 7.3 illustre l'évolution du regret de a_3 au cours du temps. Dans cet exemple, nous supposons que l'agent a_4 quitte le système entre $t = 5$ et $t = 6$. Au temps $t = 10$, l'agent a_2 effectue une traîtrise en fournissant volontairement le service s_2 avec une qualité de 0. Dans les autres cas, nous supposons que la qualité des services fournis correspond au facteur d'expertise de l'agent sélectionné.

t	1	2	3	4	5	6	7	8	9	10
s_x^t	s_1	s_2	s_2	s_3	s_3	s_1	s_2	s_1	s_3	s_2
a_k^t	a_1	a_2	a_5	a_1	a_4	a_2	a_5	a_2	a_5	a_2
(1) $\varepsilon_{\star,x}^t$	0.6	0.8	0.8	0.95	0.95	0.6	0.8	0.6	0.7	0.75
(2) $v_{3,k,x}^t$	0.6	0.8	0.75	0.7	0.95	0.3	0.75	0.3	0.6	0
(1) - (2)	0	0	0.05	0.25	0	0.3	0.05	0.3	0.1	0.75
r_3^t	0	0	0.05	0.3	0.3	0.6	0.65	0.95	1.05	1.8

Tableau 7.3 – Évolution du regret de a_3 au cours du temps

Du point de vue des agents honnêtes, interagir avec l'agent qui minimise leur regret est équivalent à maximiser leur gain. Nous évaluons donc l'efficacité d'une politique de sélection sur notre modèle d'interaction en considérant le regret moyen de tous les agents honnêtes qui suivent cette politique. De plus, comme l'objectif des agents malveillants est de maximiser le regret des agents honnêtes, nous considérons cette mesure comme une mesure d'efficacité d'une manipulation : plus le regret des agents est important, plus la manipulation est efficace.

Définition 7.2.7 - *Regret moyen* : Soit un système d'échange de services $\langle N,S \rangle$ et $M \subset N$ un ensemble d'agents malveillants en collusion. Le regret moyen des agents de $N \setminus M$ est :

$$\text{regret}(N) = \frac{\sum_{a_i \in N \setminus M} r_i^t}{|N \setminus M|}$$

Comme certaines stratégies de manipulation telle que la traîtrise (définition 6.2.8) nécessitent que les agents malveillants fournissent des services de bonne qualité, nous considérons aussi que fournir de bons services est un coût associé à la manipulation.

Définition 7.2.8 - *Coût de la manipulation* : Soit $M \subset N$ un ensemble d'agents malveillants en collusion. Soit $n_{i,k,x}$ le nombre de fois que l'agent $a_i \in N \setminus M$ a demandé à un agent $a_k \in M$ de lui fournir le service $s_x \in S_k$ et $n^+_{i,k,x}$ le nombre de fois où a_k a présenté un comportement honnête (définition 6.2.2). Le *coût* de la manipulation est donné par :

$$\text{coût}(M) = \frac{\sum_{a_k \in M} \sum_{s_x \in S_k} n^+_{i,k,x}}{\sum_{a_k \in M} \sum_{s_x \in S_k} n_{i,k,x}}$$

Ainsi, une manipulation est intuitivement intéressante si elle est efficace et peu coûteuse. Par conséquent, nous pouvons étudier la robustesse d'un système de réputation en fonction de l'efficacité des manipulations qu'il subit et du coût qu'il impose à ces manipulations.

7.3 Étude empirique

7.3.1 Protocole expérimental

Étudions maintenant empiriquement l'influence des différentes politiques de sélection sur la robustesse du système. Pour cela, nous considérons un système d'échange de services $\langle N,S \rangle$ où initialement $|N| = 100$ et $|S| = 10$ et chaque agent a_k peut fournir entre 0 et 5 services avec des facteurs d'expertise $\varepsilon_{k,x}$ tirés aléatoirement uniformément dans $[-1,1]$. À l'initialisation du système, les agents n'ont aucune connaissance des autres et vont interagir durant 200 pas de temps. À chaque pas de temps, chaque agent sélectionne aléatoirement uniformément un service dans S et, s'il ne peut pas le réaliser lui-même, il décide à l'aide de sa politique π_i à quel agent le demander et interagit avec ce dernier. Nous supposons que chaque service est fourni en un pas de temps et qu'un agent peut fournir simultanément autant de services que demandé. Lorsque un agent reçoit un service, il met ensuite à jours ses observations et les partagent avec tous les autres agents du système, puis recalcule les valeurs de réputation. Nous considérons ici que 10 % des agents du système sont des agents malveillants en collusion et, comme le système est ouvert, chaque agent honnête a une probabilité de 0.01 de quitter le système à chaque pas de temps. Avec la même probabilité, un nouvel agent honnête le rejoint.

Nous considérons deux scénarios aux stratégies malveillantes différentes. Ces deux scénarios nous permettent d'étudier l'influence des politiques de sélection sur la robustesse du système en considérant un cas favorable (scénario 1) où les agents malveillants n'effectuent pas de manipulation et un cas défavorable (scénario 2) où les agents malveillants mettent en œuvre une manipulation complexe.

Scénario 1 : les agents malveillants se contentent de toujours fournir de mauvais services. Un agent malveillant a_k fournit volontairement un mauvais service s_x à un agent honnête a_i avec une expertise $\overline{\varepsilon_{k,x}} = -1$. L'agent honnête l'évalue toujours comme tel avec $v^t_{i,k,x} = -1$.

Scénario 2 : les agents malveillants mettent en œuvre une attaque oscillante qui combine traîtrises (définition 6.2.8), faux témoignages (définition 6.2.5) et blanchiments (définition 6.2.4).

Nous considérons quatre fonctions de réputation qui recouvrent les principaux types de fonctions de réputation continues (globales, personnalisées, symétriques ou asymétriques) classiquement étudiées dans la littérature.

Estimation collective : La réputation d'un agent est la récompense moyenne observée par l'ensemble des agents lors des interactions passées (définition 7.3.1). Intuitivement, cette fonction est particulièrement sensible à toutes les manipulations, mais des variantes de cette fonction sont fréquemment utilisées sur les systèmes de réputation en ligne telles que Epinions [9] ou eBay [Resnick *et al.*, 2006]. Il est important d'étudier cette fonction pour mettre en lumière le comportement des politiques de sélection lorsque la fonction est fortement biaisée.

BetaReputation [Jøsang et Ismail, 2002] : La réputation d'un agent est une estimation de l'espérance de récompense selon une loi de densité beta où les observations des agents sont pondérées par leur réputation. La définition formelle de cette fonction est donnée en section 2.2.2. Intuitivement, cette fonction est très robuste aux faux témoignages.

EigenTrust [Kamvar *et al.*, 2003] : La réputation d'un agent est la probabilité qu'un marcheur aléatoire se déplaçant sur le graphe de confiance atteigne cet agent. Cette fonction permet de ranger les agents par ordre de fiabilité, mais la valeur n'exprime pas l'estimation de la récompense, contrairement à un modèle de bandit manchot. La définition formelle de cette fonction est donnée en section 2.2.2. Il a été montré que cette fonction est sensible aux collusions [Cheng et Friedman, 2006].

FlowTrust [Cheng et Friedman, 2005] : La réputation d'un agent est le flot maximum allant de l'agent évaluateur à l'agent évalué. La définition formelle de cette fonction est donnée en section 2.2.2. Comme précédemment, elle ne fait qu'ordonner les agents par ordre de fiabilité. Cependant, il est intéressant de noter que cette fonction est insensible aux diffamations.

Définition 7.3.1 - *Estimation collective* : Soit $\langle N,S\rangle$ un système d'échange de service. Soit $s_x \in S$ un service quelconque, $a_i \in N$ et $a_k \in N_x$ deux agents. L'*estimation collective* désigne la fonction de réputation f_i où, pour $\mu_{\mathcal{F}_i,k,x}$ la moyenne de l'ensemble $\cup_{a_j \in N} F_{i,j,k,x}$, la réputation de a_k est $f_i(a_k,s_x,\mathcal{F}_i) = \mu_{\mathcal{F}_i,k,x}$.

Pour chacune de ces fonctions de réputation, nous étudions les politiques de sélection présentée en section 7.2.1 : l'élitisme pur (définition 7.2.2), la politique ε-élitiste (définition 7.2.3), la politique ε-gloutonne (définition 7.2.4) et la politique UCB (définition 7.2.5) où ε varie par pas de 0,1 dans [0,1]. Pour chaque politique, nous mesurons le regret moyen des agents honnêtes (définition 7.2.7) et le coût de la manipulation (définition 7.2.8) sur 50 simulations. Afin de montrer que malgré les manipulations, les agents ont un intérêt à utiliser un système de réputation, nous comparons ces résultats avec un cas classique de bandit manchot où les agents utilisent uniquement leurs observations personnelles et suivent la politique UCB.

Dans un premier temps, nous présentons les résultats illustrant la dynamique du système. Pour des raisons de lisibilité des résultats, notre analyse se porte sur les cas où $\varepsilon = 0{,}1$ qui est le paramètre où le regret est minimisé dans la majorité des cas, puis nous présentons le coût de la manipulation sur ces cas. L'influence de la variation de ε est étudiée dans un troisième temps.

9. http ://www.epinions.com/

7.3.2 Influence sur le regret

Dans le scénario 1, les agents malveillants ne mettent pas en œuvre de manipulation et sont petit à petit identifiés comme malveillants et ne sont plus sollicités pour fournir les services. La figure 7.1 montre l'évolution du regret moyen des agents agents honnêtes, en fonction du système de réputation et de la politique de sélection utilisés.

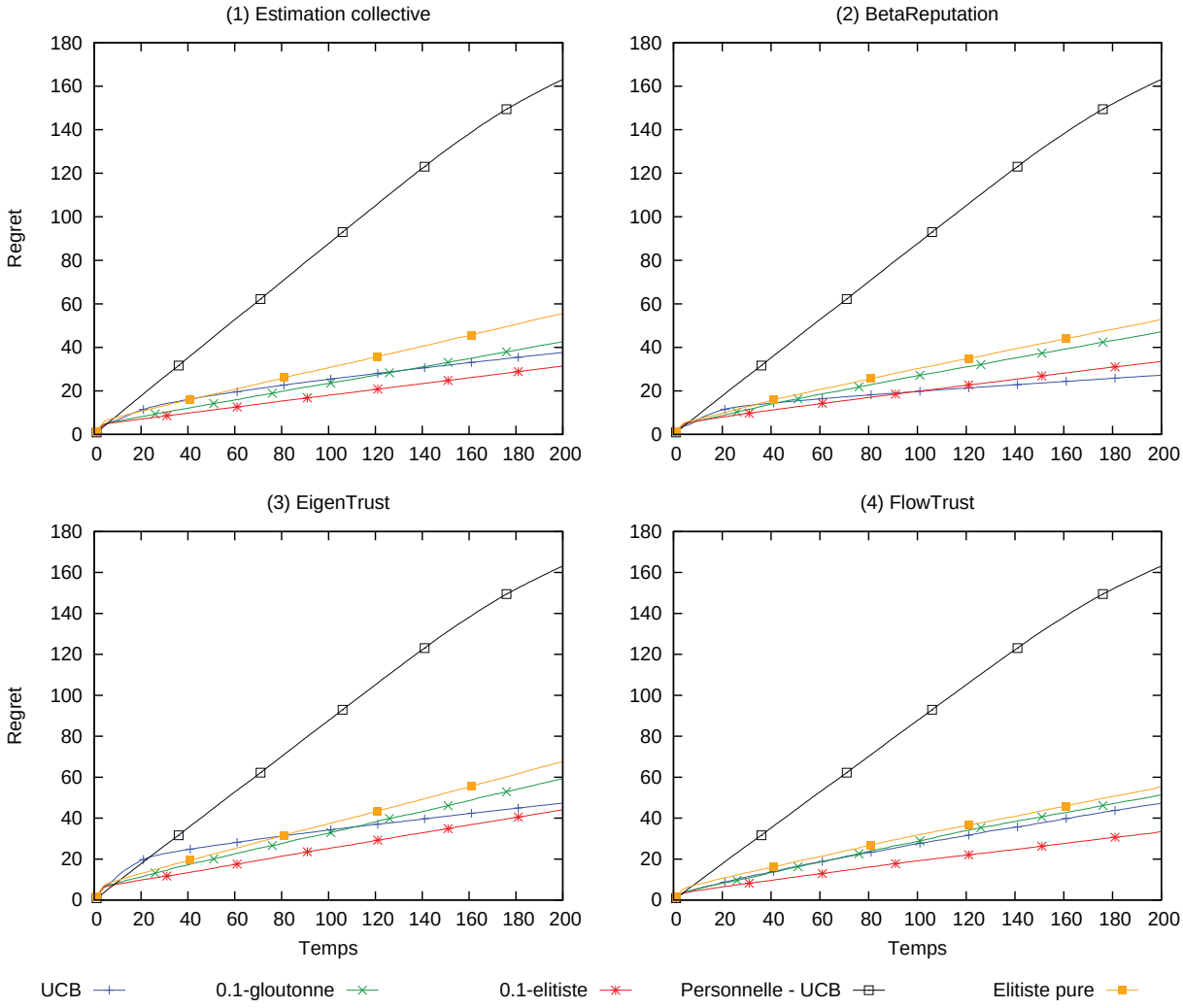

FIGURE 7.1 – Regret moyen des agents honnêtes sur le scénario 1

Nous pouvons constater qu'en l'absence de stratégie de manipulation, le système de réputation permet dans tous les cas de réduire fortement le regret des agents comparé à un cadre de bandits manchots classiques. Bien que cela s'explique par l'échelle du système, car nous considérons ici un grand nombre d'agents et de services qui doivent être chacun évalués, cela illustre l'intérêt que les agents ont à coopérer en utilisant un système de réputation. Il est toutefois intéressant de constater que l'utilisation sur un système de réputation des politiques issues des bandits manchots (UCB, ε-gloutonne, et ε-élitiste) est en moyenne inférieure à celui obtenu en suivant une politique élitiste pure classiquement utilisée dans les systèmes de réputation.

Sans surprise, la politique UCB est la plus performante sur BetaReputation (figure 7.1.2) et sur EigenTrust [10] (figure 7.1.3). Cependant, c'est la politique 0,1-élitiste qui est la plus performante sur FlowTrust (figure 7.1.4). Ceci vient du fait que l'échelle des valeurs de réputation sur BetaReputation et EigenTrust est petite rendant ainsi le facteur d'exploration d'UCB plus important. Cela permet à UCB d'avoir une meilleur estimation des agents plus rapidement. Au contraire, sur FlowTrust, l'échelle des valeurs de réputation est plus grande et le facteur d'exploration d'UCB n'a qu'un faible poids. Ainsi, le facteur d'exploration de la politique 0,1-élitiste est plus efficace. Dans les quatre cas, l'utilisation de la politique 0,1-gloutonne est la moins performante car les agents qui ont une très faible valeur de réputation sont toujours occasionnellement sollicités.

Dans le scénario 2, les agents malveillants mettent en œuvre une attaque oscillante. La figure 7.2 montre l'évolution du regret moyens des agents agents honnêtes en fonction du système de réputation et de la politique de sélection utilisés.

Sans surprise, certaines des fonctions de réputation sont plus robustes que d'autres à l'attaque oscillante. L'estimation collective (figure 7.2.1) est particulièrement sensible et aucune politique n'est plus performante que l'estimation personnelle. Ainsi, ce type de système de réputation est inefficace. Les autres fonctions de réputation sont plus robustes et il est toujours plus intéressant de les utiliser malgré la présence d'agents manipulateurs. Nous pouvons remarquer des oscillations (plus ou moins grandes en fonction de la politique) dans la croissance du regret qui coïncident avec les blanchiments et les traîtrises.

De manière intéressante, les performances de la politique UCB sont grandement dégradées sur BetaReputation et sur FlowTrust. Ceci est dû au facteur d'exploration qui, dans le cas de BetaReputation, incite UCB à interagir avec les agents malveillants qui viennent de se blanchir, leurs permettant ainsi d'augmenter leur valeur de réputation et être à même d'effectuer une traîtrise. Dans le cas de FlowTrust, le facteur d'exploration n'a pas assez de poids et conduit à interagir majoritairement avec les agents malveillants. À l'inverse, UCB devient clairement la politique la plus performante sur EigenTrust car l'échelle des valeurs de réputation sur EigenTrust est très petite : les agents malveillants obtiennent donc des valeurs de réputation suffisantes pour ne pas effectuer de blanchiment alors que le facteur d'exploration prend suffisamment de poids pour permettre d'interagir régulièrement avec des agents honnêtes.

Les politiques 0,1-élitiste et 0,1-gloutonne sont très proches mis à part dans le cas de BetaReputation (figure 7.2.2) et leur intérêt dépend essentiellement de leurs performances relatives par rapport à UCB. Leur facteur d'exploration fonctionne différemment de celui d'UCB et ne tient pas compte des échelles des valeurs de réputation. Ainsi, ces politiques permettent d'interagir régulièrement avec des agents honnêtes. Dans le cas de BetaReputation, les diffamations amènent de nombreux agents malveillants à avoir les plus hautes valeurs de réputation et donc à affaiblir la politique élitiste. Ce n'est pas le cas avec FlowTrust qui est insensible aux diffamations et EigenTrust pour lequel promotion et diffamation sont formellement identiques.

7.3.3 Coût de la manipulation

Cependant pour que l'attaque oscillante soit efficace, les agents malveillants doivent présenter régulièrement un comportement honnête. La figure 7.3 nous montre le coût de la manipulation (définition 7.2.8) en fonction du système de réputation et de la politique de sélection utilisés.

Sur toutes les courbes, le coût des manipulations présente des oscillations régulières, plus ou moins grandes en fonction du système de réputation et de la politique de sélection. Ces

10. Bien que l'élitisme semble plus efficace sur cette courbe, une prolongation de quelques pas temps montre UCB plus performant.

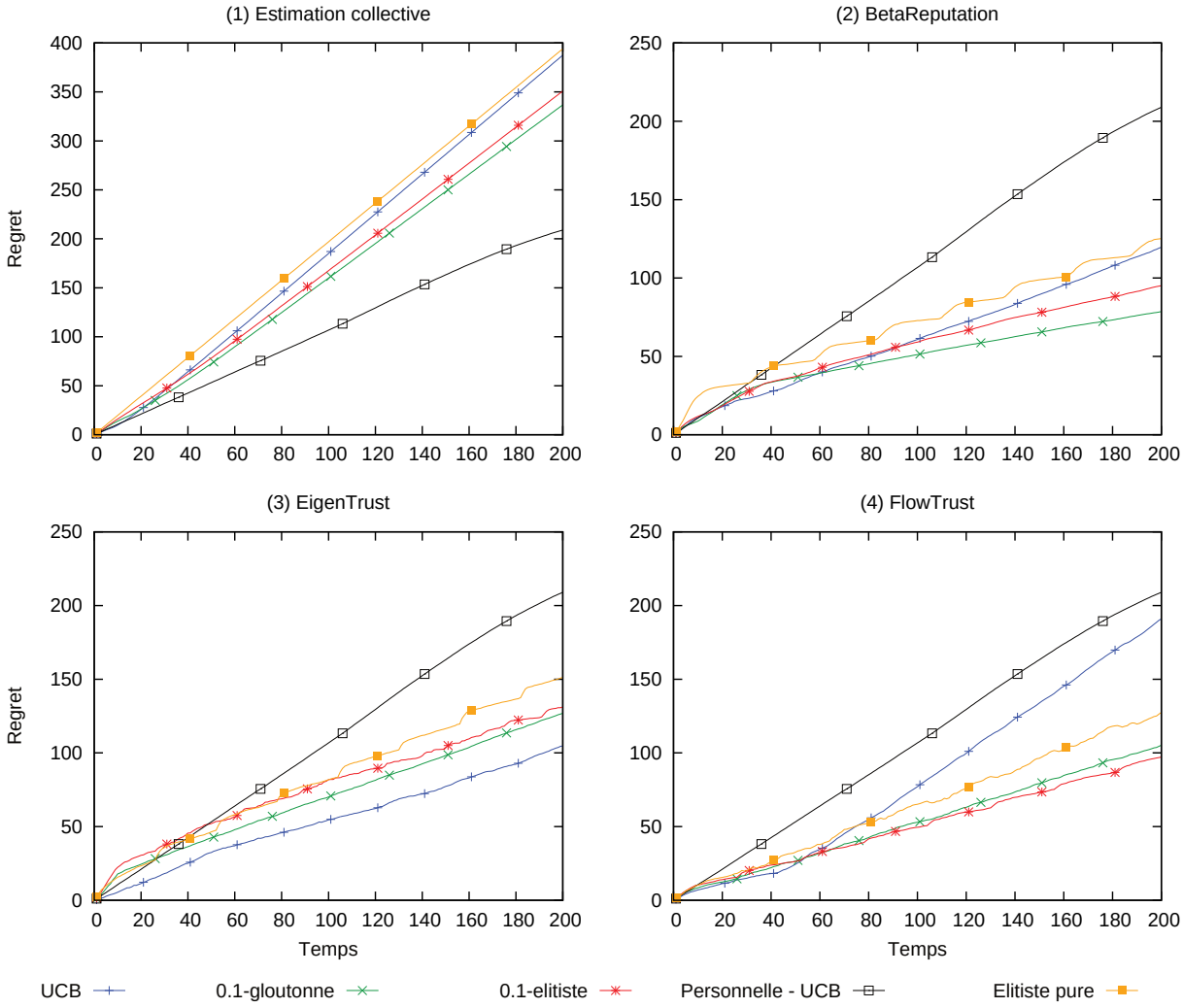

FIGURE 7.2 – Regret moyen des agents honnêtes sur le scénario 2

oscillations correspondent aux périodes où les agents de M_2 voient leur réputation décroître pour devenir inférieure à celle des agents de M_1 qui viennent de rejoindre le système suite à un blanchiment. Les agents de M_1 sont donc sélectionnés par les agents honnêtes, ce qui leur permet d'acquérir suffisamment de réputation pour pouvoir effectuer une promotion des agents de M_2.

Nous pouvons remarquer que le coût initial est très élevé avec UCB puis chute rapidement. En effet, les agents honnêtes commencent par explorer et interagir avec les agents malveillants qui ne sont pas promus et qui fournissent de bons services puis, très rapidement, le facteur d'exploration perd de son importance et les agents honnêtes vont interagir avec les agents malveillants qui fournissent de mauvais services. Remarquons qu'utiliser UCB sur l'estimation personnelle implique un coût de manipulation presque nul. Cependant, si UCB est la politique la moins performante sur BetaReputation, EigenTrust et FlowTrust, elle est aussi la plus coûteuse pour les agents malveillants car ils doivent se comporter honnêtement dans un tiers des interactions.

Les politiques 0,1-gloutonne et 0,1-élitiste implique un coût de manipulation globalement

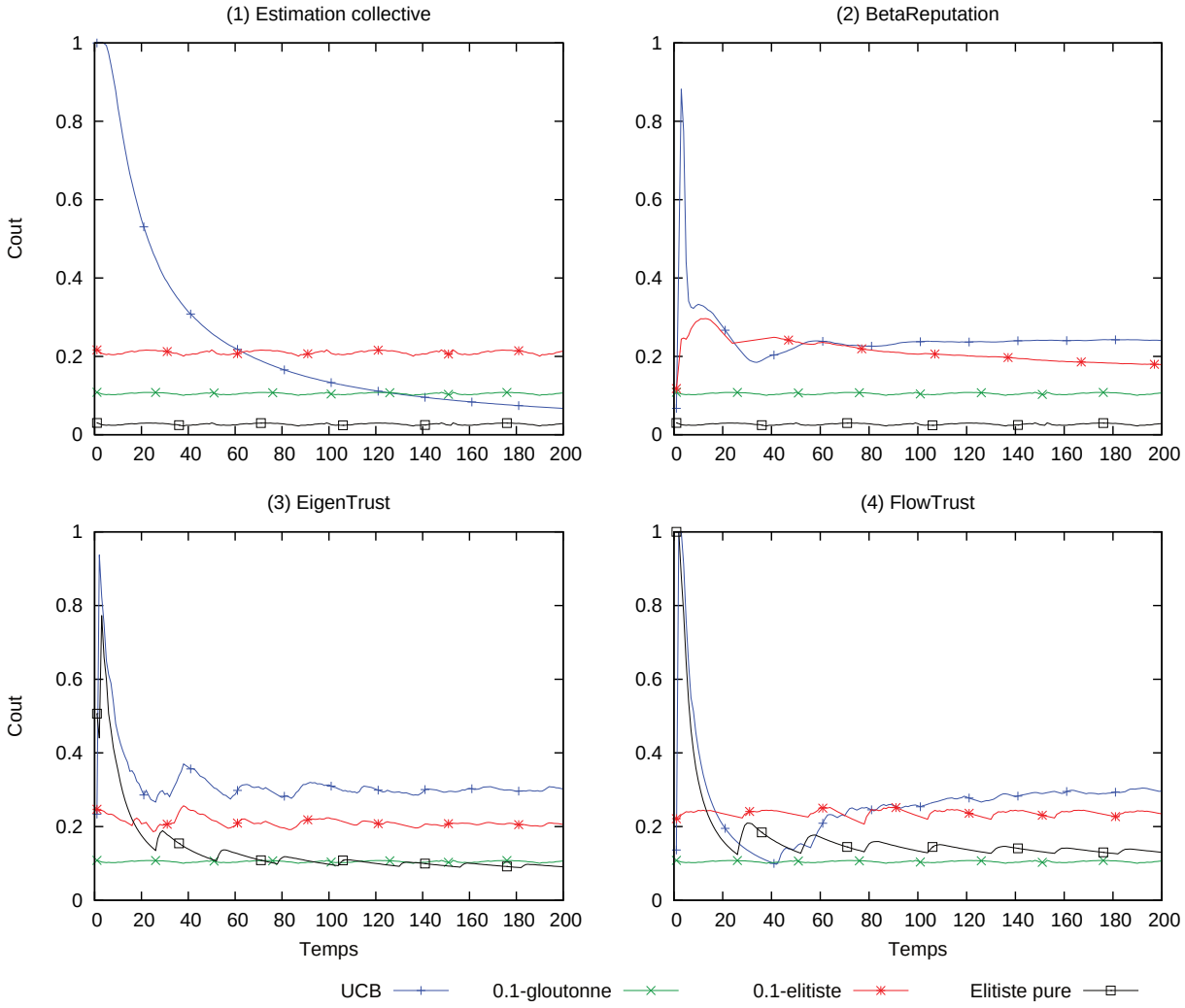

FIGURE 7.3 – Coût de la manipulation pour les agents malveillants sur le scénario 2

constant pour toutes les fonctions de réputation. Encore une fois, le facteur d'exploration permet d'interagir occasionnellement avec des agents malveillants qui doivent présenter un comportement honnête. La politique élitiste pure, elle, présente deux profils de comportements bien que, de manière générale, ce soit celle dont le coût de manipulation est le plus faible. Sur l'estimation collective et BetaReputation, elle est constante comme les politiques 0,1-gloutonne et 0,1-élitiste. Sur EigenTrust et FlowTrust, elle présente la même forme que la politique UCB. Dans le premier cas, les agents malveillants qui promeuvent n'ont pas les meilleures réputations et ne sont donc pas sélectionnés. S'ils le sont, c'est au moment d'effectuer leur traîtrise. Dans le second cas, cela est dû au fait que la réputation représente un ordre sur les agents et que de petites variations de réputation induisent des inversions dans cet ordre, l'agent ayant la meilleure réputation alterne entre un agent malveillant de M_1, de M_2 et occasionnellement des agents honnêtes.

7.3.4 Influence du facteur d'exploration

De manière intéressante, les sections précédentes nous amènent à considérer le facteur d'exploration comme étant un élément essentiel de la robustesse du système de réputation aux manipulations. C'est pourquoi nous nous intéressons ici à l'influence des valeurs de ε sur les politiques ε-gloutonne et ε-élitiste. La figure 7.4 représentons le regret moyen des agents honnêtes après 200 pas de temps selon les fonctions de réputation et politiques de sélection utilisées. Les histogrammes présentent ce regret en fonction de la valeur de ε. Les lignes constantes représentent le regret avec les autres politiques qui ont un facteur d'exploration nul ou non paramétré.

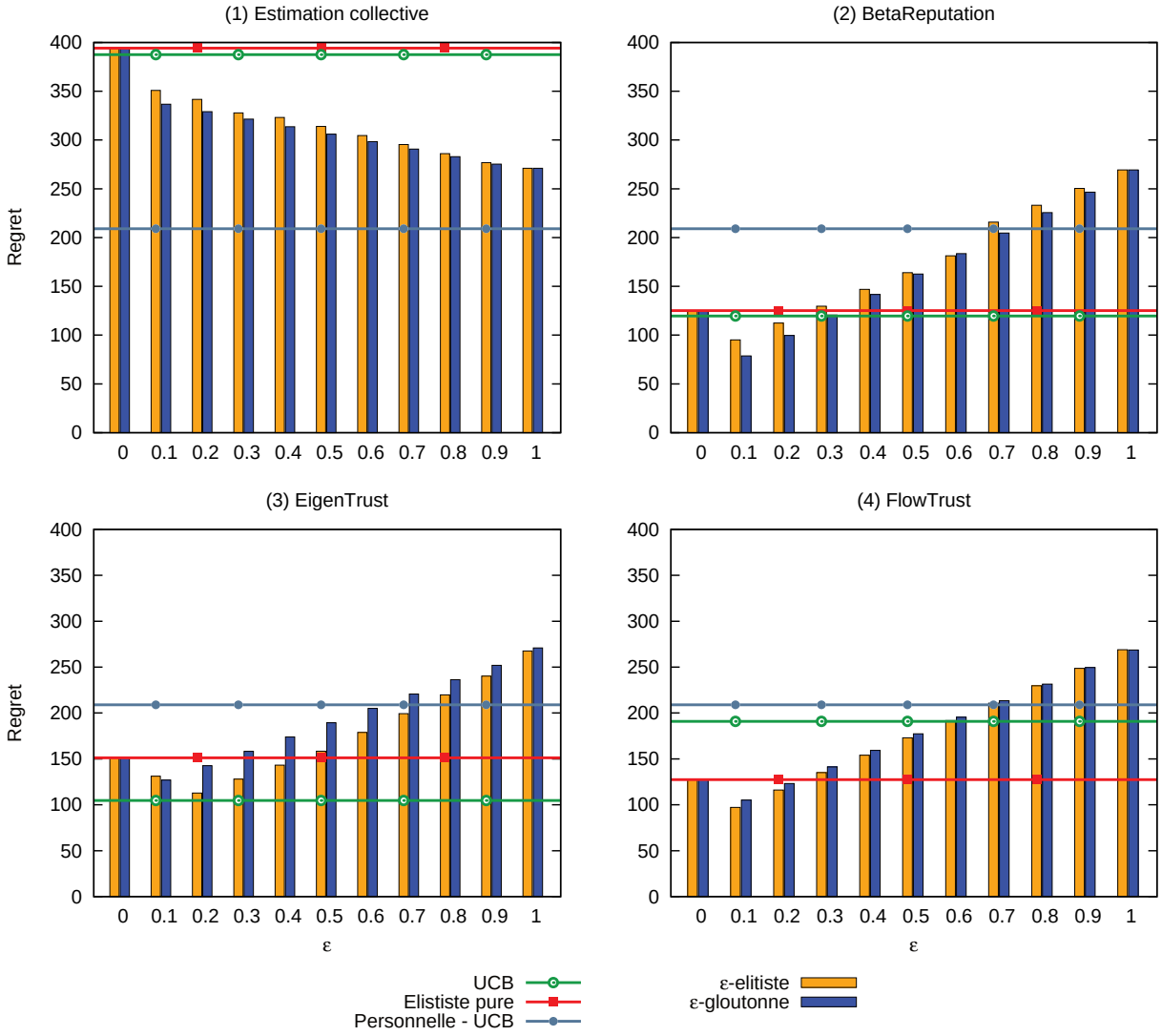

FIGURE 7.4 – Influence du facteur d'exploration des politiques ε-elitiste et ε-gloutonne

À de petites variations près, les politiques ε-gloutonne, et ε-élitiste produisent les mêmes résultats lorsque ε varie. Dans le cas de l'estimation collective, bien qu'augmenter la valeur de ε diminue le regret, cette fonction de réputation est si sensible aux manipulations que l'estimation personnelle reste toujours plus intéressante. Sur les autres fonctions de réputation, le regret

décroît pour $\varepsilon \in]0;0{,}1]$ puis augmente pour $\varepsilon \in]0{,}1;1]$. Globalement, les politiques ε-gloutonne, et ε-élitiste sont plus performantes que l'élitisme pur lorsque $\varepsilon \in]0;0{,}2]$ et plus performantes qu'UCB pour $\varepsilon \in]0;0{,}5]$ sur FlowTrust et pour $\varepsilon \in]0;0{,}2]$ sur BetaReputation. Comme vu précédemment, la politique UCB reste la plus performante sur EigenTrust. Le paramètre optimal est de 0,1 pour toutes les politiques sur toutes les fonctions de réputation, sauf sur EigenTrust où il est de 0,2 pour la politique ε-élitiste car les agents malveillants arrivent aisément à tous se trouver parmi les agents ayant les plus hautes valeurs de réputation : un facteur d'exploration plus important permet d'éviter de toujours interagir avec eux.

Ainsi, un faible facteur d'exploration non nul est suffisant pour rendre plus robuste la fonction de réputation. Toutefois, si le facteur devient trop important alors les interactions avec les meilleurs des agents honnêtes sont plus rares et le regret augmente.

7.4 Conclusion

Nous avons montré dans ce chapitre que le problème de décision d'un agent de notre modèle d'interaction peut être associé par analogie à un problème de bandits manchots. Décider en fonction d'une valeur de réputation avec quel agent interagir peut en effet être considéré comme décider à l'aide d'observations quel bras d'une machine à sous tirer afin de maximiser un gain. Nous avons présenté le modèle classique des bandits manchots ainsi que ses principales politiques de sélection qui sont des compromis entre l'exploitation des connaissances et l'exploration. Nous avons adapté ces politiques dans le contexte des systèmes de réputation où il est classiquement admis que les agents suivent une politique élitiste pure pour décider avec qui interagir.

Pour mesurer l'influence des politiques de sélection sur la robustesse aux manipulations du système de réputation, nous avons procédé à une étude empirique en nous servant d'une mesure classique de regret et d'une notion de coût de la manipulation. Si, sans surprise, l'utilisation d'un système de réputation est toujours plus intéressante dans un contexte sans manipulation qu'une approche individuelle, ce n'est pas toujours le cas lorsqu'il y a des manipulations. Bien que l'estimation collective [11] soit si peu robuste qu'une approche individuelle est toujours préférable, nos résultats prennent leur sens sur des systèmes de réputation moins triviaux tels que BetaReputation, EigenTrust et FlowTrust. Nous avons montrés que si la politique élitiste pure est classiquement utilisée dans les systèmes de réputation, celle-ci n'est pas nécessairement la plus performante.

Lorsque les agents se limitent à exhiber des comportements malveillants, c'est-à-dire fournir de mauvais services, la politique UCB est naturellement celle qui minimise le regret. Cependant, la politique UCB est très sensible aux blanchiments car son facteur d'exploration repose sur la sélection des agents pour lesquels il y a eu le moins d'interaction (de fait les agents malveillants qui viennent se blanchir). En revanche, cette exploration entraîne un fort coût de manipulation car les agents malveillants doivent régulièrement fournir de bons services. Selon la fonction de réputation, la perte de performance de la politique UCB est plus ou moins importante et la politique la plus intéressante dépend alors de chaque système. Ainsi :

1. la politique ε-gloutonne est la plus performante sur BetaReputation ;
2. la politique ε-élitiste est la plus performante sur FlowTrust ;
3. la politique UCB reste la plus performante sur EigenTrust.

La raison de ces différences repose sur l'*interaction entre certaines caractéristiques des fonctions de réputation et la manière dont l'exploration s'effectue*. Par exemple, lorsque la fonc-

11. Archétype de nombreux systèmes de réputation tels qu'eBay et Epinions.

tion de réputation est robuste aux diffamations, la politique ε-élitiste devient plus performante. Lorsque les valeurs de réputation représentent non pas une estimation de la récompense mais un simple ordre entre les agents, les coûts de manipulation des politiques élitistes pures augmentent. Lorsque les valeurs de réputation présentent un large écart entre les unes et les autres, la politique UCB est dégradée alors que la politique ε-gloutonne devient plus performante.

Dans tous les cas, le paramétrage du facteur d'exploration a une influence sur la robustesse du système mais ce paramétrage doit être fin : ni trop important, ni trop faible pour que la robustesse augmente. Cependant, utiliser un facteur d'exploration ne prévient pas toutes les manipulations. En effet, la valeur de réputation reste faussée par les faux témoignages et il convient de s'en prémunir.

Chapitre 8

Crédibilité des témoignages

Sommaire

Résumé.
L'utilisation des politiques de sélection présentée au chapitre précédent permet de réduire l'influence des comportements malveillants, mais reste sensible aux manipulations faussant les valeurs de réputation, en particulier aux *faux témoignages* (définition 6.2.5). Dans ce chapitre, nous nous intéressons alors à une approche permettant de lutter contre ce type de manipulation. Pour ce faire, nous proposons une nouvelle mesure de *crédibilité des témoignages* fondée sur la *divergence de Kullback-Leibler*. Cette mesure consiste à calculer la divergence entre les observations de l'agent et les témoignages qu'il reçoit, et à ne considérer que les informations offrant un compromis entre nouveauté et cohérence. Contrairement aux approches de la littérature qui se servent de la mesure de crédibilité pour pondérer les témoignages, nous proposons d'utiliser des *fonctions de filtrage* qui, en fonction de la valeur de crédibilité, décident ou non de retirer un témoignage du calcul de la réputation. Cette proposition a fait l'objet de publications [Vallée et Bonnet, 2015, Vallée *et al.*, 2015]. Dans une première section, nous présentons pourquoi la divergence de Kullback-Leibler peut être considérée comme une mesure de crédibilité. Nous présentons ensuite trois fonctions de filtrage utilisant cette mesure : le filtrage par *KL-credibilité*, le filtrage par *k-fautes* et la *k-stochocratie*. Pour conclure, nous étudions empiriquement leur efficacité sur différents systèmes de réputation.

Afin de simplifier la lecture de ce chapitre, le tableau 8.1 résume les principales notations utilisées.

Crédibilité et filtrage	
$\mu_{i,k,x}$	Moyenne des valeurs de $O_{i,k,x}$
$\sigma_{i,k,x}$	Écart-type des valeurs de $O_{i,k,x}$
$\mathcal{N}(\mu_{i,k,x},\sigma^2_{i,k,x})$	Approximation par la loi normale de $\varepsilon_{k,x}$ à partir de $O_{i,k,x}$
$\mu_{i,j,k,x}$	Moyenne des valeurs de $F_{i,j,k,x}$
$\sigma_{i,j,k,x}$	Écart-type des valeurs de $F_{i,j,k,x}$
$\mathcal{N}(\mu_{i,j,k,x},\sigma^2_{i,j,k,x})$	Approximation par la loi normale de $\varepsilon_{k,x}$ à partir de $F_{i,j,k,x}$
$D_{i,j,k,x}$	Divergence de Kullback-Leibler entre $\mathcal{N}(\mu_{i,k,x},\sigma^2_{i,k,x})$ et $\mathcal{N}(\mu_{i,j,k,x},\sigma^2_{i,j,k,x})$
$KL_i(F_{i,j,k,x})$	Le témoignage $F_{i,j,k,x}$ est crédible
$KL^k_i(N) \subseteq N$	Ensemble des agents k-crédible
$L^k_i(F_{i,j,k',x})$	Le témoignage $F_{i,j,k',x}$ est crédible par k-stochocratie
ϕ_i	Fonction de filtrage de l'agent a_i

Tableau 8.1 – Principales notations de la crédibilité des témoignages

8.1 Une mesure de crédibilité

L'efficacité d'un système de réputation repose sur le fait que les agents partagent le résultat de leurs observations passées afin de pouvoir les utiliser lors du calcul des valeurs de réputation. La figure 7.1 nous a permis de mettre en évidence le gain apporté par un tel partage. Cependant, des agents malveillants peuvent fournir de *faux témoignages* (définition 6.2.5) afin d'altérer le calcul de la valeur de réputation à leur avantage. Pour lutter contre cette manipulation, nous proposons une technique permettant à un agent d'estimer si un témoignage doit être utilisé ou non.

8.1.1 Divergence d'un témoignage

Dans notre modèle d'interaction, la capacité d'un agent $a_k \in N$ à fournir un service $s_x \in S_k$ de bonne qualité est définie par son facteur d'expertise $\varepsilon_{k,x}$ qui correspond à l'espérance d'une distribution de probabilité $\theta_{k,x}$ de paramètres inconnus. Comme les observations d'un agent a_i correspondent à l'ensemble des évaluations de ses interactions passées, il peut utiliser $O_{i,k,x}$ pour estimer le gain moyen de ces interactions. Dans toute la suite, nous notons par $\mu_{i,k,x}$ la moyenne des valeurs de $O_{i,k,x}$ et par $\sigma_{i,k,x}$ son écart-type. Ces estimations permettent ensuite à l'agent a_i d'approximer la fonction $\theta_{k,x}$ en présupposant qu'elle suit une loi normale.

Définition 8.1.1 - *Approximation par loi normale de l'expertise* : Soit $a_i,a_k \in S$ deux agents et $s_x \in S_k$ un service. L'*approximation par loi normale de l'expertise* de l'agent a_k pour le service s_x est la loi de probabilité $\mathcal{N}(\mu_{i,k,x},\sigma^2_{i,k,x})$ où $\mu_{i,k,x}$ désigne la moyenne et $\sigma_{i,k,x}$ l'écart-type des valeurs de $O_{i,k,x}$.

Comme le témoignage d'un agent a_j correspond aux observations que a_j prétend avoir eu, un agent a_i peut utiliser les témoignages de a_j vis-à-vis d'un agent a_k pour le service s_x pour calculer une autre approximation de $\varepsilon_{k,x}$. Ainsi, a_i peut approximer la fonction $\theta_{k,x}$ par la loi

normale $\mathcal{N}(\mu_{i,j,k,x},\sigma^2_{i,j,k,x})$ où $\mu_{i,j,k,x}$ et $\sigma_{i,j,k,x}$ désigne respectivement la moyenne et l'écart-type des valeurs du témoignage $F_{i,j,k,x}$.

Exemple 8.1.1 - Considérons un service s_1 et les agents a_1,$a_2 \in N$ et $a_3 \in N_1$. Supposons que $D_1 = [-1,1]$ et que l'agent a_1 a les informations suivantes :

$$O_{1,3,1} = \{0.25, -0.5, 0.5, 0, -0.25\}$$
$$F_{1,2,3,1} = \{0.25, -0.25, 0.6\}$$

À partir de ces informations, a_1 peut calculer les moyennes et écarts-types suivants :

$$\mu_{1,3,1} = 0 \text{ et } \mu_{1,2,3,1} = 0.2$$
$$\sigma_{1,3,1} \simeq 0.354 \text{ et } \sigma_{1,2,3,1} \simeq 0.349$$

Ainsi, a_1 peut approximer l'expertise de a_3 pour le service s_1 par les lois normales $\mathcal{N}(0,0.354^2)$ et $\mathcal{N}(0.2,0.349^2)$.

Ces deux estimations sont supposées être deux estimations de la même distribution de probabilité. Ainsi, sous l'hypothèse que la qualité des services fournis est indépendante de l'agent recevant le service, a_i et a_j doivent obtenir les mêmes approximations par loi normale de l'expertise pour un grand nombre d'observations.

Hypothèse 8.1.1 - : Si $O_{i,k,x}$ et $F_{i,j,k,x}$ sont des observations du service $s_x \in S$ fourni par a_k alors ces observations proviennent de la même fonction de distribution de probabilité et les approximations par lois normales $\mathcal{N}(\mu_{i,k,x},\sigma^2_{i,k,x})$ et $\mathcal{N}(\mu_{i,j,k,x},\sigma^2_{i,j,k,x})$ doivent converger. Pour $n = |O_{i,k,x}|$ et $m = |F_{i,j,k,x}|$, nous devons avoir :

$$\lim_{n,m\to\infty} \mathcal{N}(\mu_{i,k,x},\sigma^2_{i,k,x}) = \mathcal{N}(\mu_{i,j,k,x},\sigma^2_{i,j,k,x})$$

À l'inverse, si $F_{i,j,k,x}$ est un faux témoignage alors :

$$\lim_{n,m\to\infty} \mathcal{N}(\mu_{i,k,x},\sigma^2_{i,k,x}) \neq \mathcal{N}(\mu_{i,j,k,x},\sigma^2_{i,j,k,x})$$

Ainsi, si le témoignage de a_j est faux alors l'approximation par la loi normale de l'expertise construite à partir de ce témoignage diffère fortement de l'approximation construite à partir des observations de a_i. Cela est aussi le cas lorsque des erreurs d'observation faussent l'estimation de l'agent témoin ou que ce dernier n'évalue pas la qualité des services sur les mêmes critères. Cependant, dans tous les cas, utiliser ce témoignage n'a pas d'intérêt.

Comme les agents ne disposent que d'un nombre fini d'observations, leurs estimations diffèrent nécessairement. Nous proposons alors de mesurer leur différence en utilisant la divergence de Kullback-Leibler qui est une mesure de dissimilarité entre deux distributions de probabilité $f(x)$ et $g(x)$ [Kullback, 1968] :

$$D_{KL}(f||g) = \int f(x) \log \frac{f(x)}{g(x)} d(x)$$

Intuitivement, plus la divergence entre les distributions $f(x)$ et $g(x)$ est importante plus elles différent. Ainsi, nous utilisons cette mesure pour calculer l'écart entre les observations d'un agent et le témoignage d'un autre.

Définition 8.1.2 - *Divergence de Kullback-Leibler* : La *divergence de Kullback-Leibler* entre les observations de a_i et les témoignages de a_j vis-à-vis de $\varepsilon_{k,x}$ est :

$$D_{i,j,k,x} = D_{KL}(\mathcal{N}(\mu_{i,k,x},\sigma^2_{i,k,x})||\mathcal{N}(\mu_{i,j,k,x},\sigma^2_{i,j,k,x}))$$

Exemple 8.1.2 - Reprenons les agents a_1, a_2 et a_3 de l'exemple 8.1.1 où les observations de a_1 vis-à-vis de a_3 pour le service s_1 sont $O_{1,3,1} = \{0.25, -0.5, 0.5, 0, -0.25\}$ et le témoignage de a_2 est $F_{1,2,3,1} = \{0.25, -0.25, 0.6\}$. Les deux approximations par loi normale de l'expertise sont respectivement $\mathcal{N}(0,0.354^2)$ et $\mathcal{N}(0.2,0.349^2)$. La divergence de Kullback-Leibler entre les observations de a_1 et le témoignage de a_2 vis-à-vis de $\varepsilon_{3,1}$ est donc :

$$\begin{aligned} D_{1,2,3,1} &= D_{KL}(\mathcal{N}(0,0.354^2),\mathcal{N}(0.2,0.349^2)) \\ &\simeq 0.165 \end{aligned}$$

Comme $D_{KL}(f||g) = 0$ implique que $f(x)$ et $g(x)$ sont les mêmes distributions de probabilité alors l'hypothèse 8.1.1 se traduit par :

$$\lim_{n,m\to\infty} D_{i,j,k,x} = 0$$

Par ailleurs, la divergence de Kullback-Leibler est utilisée dans les problèmes d'apprentissage par renforcement pour mesurer le gain apporté par une nouvelle observation [Hershey *et al.*, 2007]. En effet, la divergence de Kullback-Leibler est fortement liée à l'entropie de Shanon et permet de calculer la quantité d'informations nouvelles apportée par un ensemble d'observations [Shannon, 1951]. Dans d'un système de réputation, la divergence entre les observations d'un agent a_i et le témoignage d'un agent a_j peut être considérée comme l'apport de ce témoignage aux connaissances de l'agent a_i. Cependant, une valeur de divergence importante peut également signifier que les observations de a_i et le témoignage de a_j ne proviennent pas de la même distribution de probabilité. C'est pourquoi nous proposons ici d'utiliser la divergence de Kullback-Leibler pour estimer la crédibilité d'un témoignage.

8.1.2 Crédibilité d'un témoignage

Si le témoignage de a_j est similaire aux observations de a_i alors $D_{i,j,k,x} \simeq 0$. Inversement, si $D_{i,j,k,x}$ est supérieure à un seuil δ, cela signifie que l'agent a_i et l'agent a_j n'ont pas la même estimation de $\varepsilon_{k,x}$ car :

1. les agents a_i ou a_j n'ont pas suffisamment d'observations pour bien estimer $\varepsilon_{k,x}$;
2. les agents a_i et a_j évaluent la qualité des services sur des critères différents ($v_i \neq v_j$) ;
3. les agents a_i et a_j subissent des bruits différents sur leurs observations ;
4. le témoignage de a_j est volontairement faux (définition 6.2.5).

Dans le premier cas, après quelques interactions supplémentaires, la divergence entre les observations et les témoignages $D_{i,j,k,x}$ diminuera pour ensuite tendre vers 0. Dans les trois autres cas, l'agent a_i ne doit pas considérer comme crédible ce témoignage car il est soit inutile (second cas et troisième cas), soit faux (quatrième cas). En effet, bien qu'il y ait une distinction de sens entre un faux témoignage et un témoignage inutile, se servir de ce témoignage amène dans tous les cas à une mauvaise estimation de l'expertise. Il convient donc pour l'agent a_i de les considérer de la même manière, c'est-à-dire *non crédible*.

Exemple 8.1.3 - Reprenons l'exemple 8.1.1 où les observations de a_1 vis-à-vis de a_3 pour le service s_1 sont $O_{1,3,1} = \{0.25, -0.5, 0.5, 0, -0.25\}$. Considérons les agents a_2, a_4, a_5 et a_6 où a_2 et a_4 sont honnêtes et ont la même fonction d'évaluation que a_1, où a_4 dispose d'un grand nombre d'observations, où a_5 est un agent honnête mais possède une autre fonction d'évaluation et où a_6 est un agent malveillant qui diffame a_3. Le tableau 8.2 récapitule la divergence entre les observations de a_1 et les témoignages des autres agents :

a_j	$\mu_{1,j,3,1}$	$\sigma_{1,j,3,1}$	$D_{1,j,3,1}$
a_2	0.2	0.349	0.165
a_4	0.5	0.6	0.55
a_5	0.9	0.1	44.987
a_6	−0.8	0.2	8.49

Tableau 8.2 – Moyenne, écart-type et divergence des témoignages reçus vis-à-vis de $\varepsilon_{3,1}$

Dans cet exemple, le témoignage de a_2 a une faible divergence : il semble donc être le plus crédible. Le témoignage de a_4 a une divergence de 0.55 et semble donc légèrement moins crédible. Les témoignages de a_5 et de a_6 ont tous deux une très grande valeur de divergence et ne sont donc pas crédibles. Supposons maintenant que a_1 obtienne 6 nouvelles observations :

$$O_{1,3,1} = \{0.25, -0.5, 0.5, 0, -0.25, 0.5, 0.75, -0.25, 0.5, -0.25, 0.75\}$$

Nous avons alors $\mu_{1,3,1} \simeq 0.182$, $\sigma_{1,3,1} \simeq 0.428$ et $D_{1,4,3,1} \simeq 0.093$. Les divergences entre les observations de a_1 et les témoignages reçus sont alors de :

$$D_{1,2,3,1} \simeq 0.049 \ , \ D_{1,4,3,1} \simeq 0.233$$
$$D_{1,5,3,1} \simeq 32.978 \ , \ D_{1,6,3,1} \simeq 13.075$$

Ainsi, avec un plus grand nombre d'observations, les témoignages de a_2 et a_4 divergent moins tandis que les témoignages de a_5 et de a_6 conservent une forte valeur de divergence.

L'exemple 8.1.3 illustre que le nombre d'observations joue un rôle important dans la crédibilité. Par ailleurs, si un témoignage n'est pas crédible à un instant t_1, ce même témoignage peut le devenir dans le futur. Ainsi, pour fixer le seuil δ à partir duquel un agent considère comme non crédible un témoignage, nous proposons d'utiliser l'erreur type de l'estimateur. Intuitivement, l'erreur type de la moyenne correspond à la confiance de l'agent a_i dans son estimation de $\mu_{i,k,x}$ en fonction de son échantillonnage, c'est-à-dire du nombre d'observations.

Définition 8.1.3 - *Erreur type de la moyenne* : Soit $a_i \in N$ un agent ayant n observations $O_{i,k,x}$ et $\mathcal{N}(\mu_{i,k,x}, \sigma^2_{i,k,x})$ son approximation d'une expertise $\varepsilon_{k,x}$ par la loi normale. L'*erreur type de la moyenne* de son approximation est :

$$SEM(\mathcal{N}(\mu_{i,k,x}, \sigma^2_{i,k,x})) = \frac{\sigma_{i,k,x}}{\sqrt{n}}$$

L'utilisation de l'erreur type de la moyenne permet à l'agent a_i de déterminer avec un intervalle de confiance quelle est la récompense réelle espérée. Dans toute la suite, fixons un intervalle de confiance à 95 % où la récompense réelle espérée se trouve dans l'intervalle :

$$[\mu_{i,k,x} - \frac{1.96 \times \sigma_{i,k,x}}{\sqrt{n}}, \mu_{i,k,x} + \frac{1.96 \times \sigma_{i,k,x}}{\sqrt{n}}]$$

L'agent a_i se sert alors de sa propre erreur d'estimation pour fixer δ et décider si un témoignage est crédible, comme illustré sur la figure 8.1.

Définition 8.1.4 - *Témoignage crédible* : Soit $F_{i,j,k,x}$ le témoignage que a_j a fourni à a_i vis-à-vis de l'expertise de a_k pour le service s_x. $F_{i,j,k,x}$ est *crédible* si $D_{i,j,k,x} \leq \delta$ où :

$$\delta = D_{KL}(\mathcal{N}(\mu_{i,k,x},\sigma^2_{i,k,x})||\mathcal{N}(\mu_{i,k,x} + \frac{1.96 \times \sigma_{i,k,x}}{\sqrt{n}},\sigma^2_{i,k,x}))$$

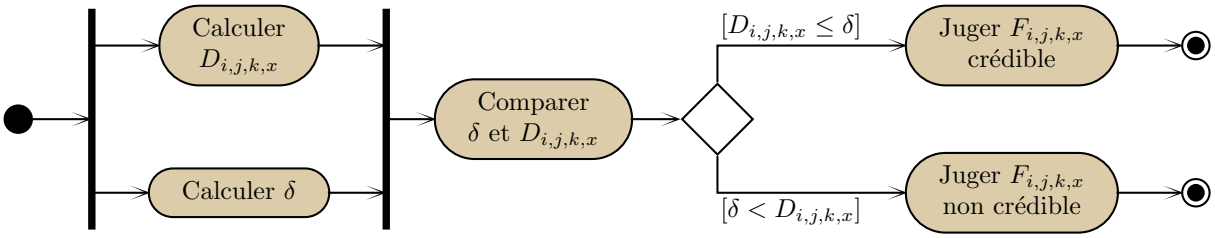

FIGURE 8.1 – Diagramme d'activité du jugement d'un témoignage

Utiliser la divergence de Kullback-Leibler comme mesure de crédibilité et l'erreur type de la moyenne pour fixer dynamiquement le seuil présente plusieurs avantages. Comme la divergence de Kullback-Leibler est fortement liée à l'entropie, un témoignage divergent apporte de nouvelles informations utiles lorsque l'agent ne dispose que de peu d'observations. À l'inverse, plus l'agent dispose d'informations, moins un nouveau témoignage est supposé apporter une information utile. Comme l'erreur type de la moyenne dépend du nombre d'observations, plus l'agent en dispose, moins un témoignage divergent est supposé crédible, et inversement. Ainsi, cette notion de crédibilité est dynamique car elle peut être remise en cause au cours du temps au fur et à mesure que l'agent obtient de nouvelles informations.

L'asymétrie de la divergence de Kullback-Leibler nous permet de représenter le fait que si un agent a_i considère comme non crédible le témoignage d'un agent a_j, a_j peut quant à lui considérer le témoignage de a_i comme crédible, car lui-même ne dispose pas du même nombre d'observations. Enfin, la prise en compte de l'erreur type de la moyenne dans la définition du seuil de crédibilité permet à un agent de considérer que ses observations sont en partie imparfaites et un agent qui saurait qu'il est en proie à un bruit d'observation pourrait réévaluer ce seuil.

Enfin, remarquons qu'un agent a_i considère nécessairement comme crédible ses propres observations car $D_{i,i,k,x} = 0$ pour tout service $s_x \in S$ et tout agent $a_k \in N_x$.

Exemple 8.1.4 - Reprenons l'exemple 8.1.3. Supposons que les observations de a_1 soient telles que $\mu_{1,3,1} = 0$ et $\sigma_{1,3,1} = 0.4$. La figure 8.2 représente les valeurs de divergence entre les observations de a_1 et un témoignage $F_{1,2,3,1}$ en fonction de $\mu_{1,2,3,1}$ et $\sigma_{1,2,3,1}$. Sur cette figure, le blanc correspond à une configuration où $D_{1,2,3,1} > 1$.
Considérons le témoignage de a_2 tel que $\mu_{1,2,3,1} = 0.2$ et $\sigma_{1,2,3,1} = 0.6$. Le tableau 8.3 indique la crédibilité de ce témoignage en fonction du nombre d'observations de l'agent a_1. Nous pouvons constater que a_1 ayant peu d'observations ($|O_{1,3,1}| = 5$) considère le témoignage de a_2 comme crédible initialement, car son seuil de crédibilité δ est de 0.41. Ce seuil décroît au fur et à mesure

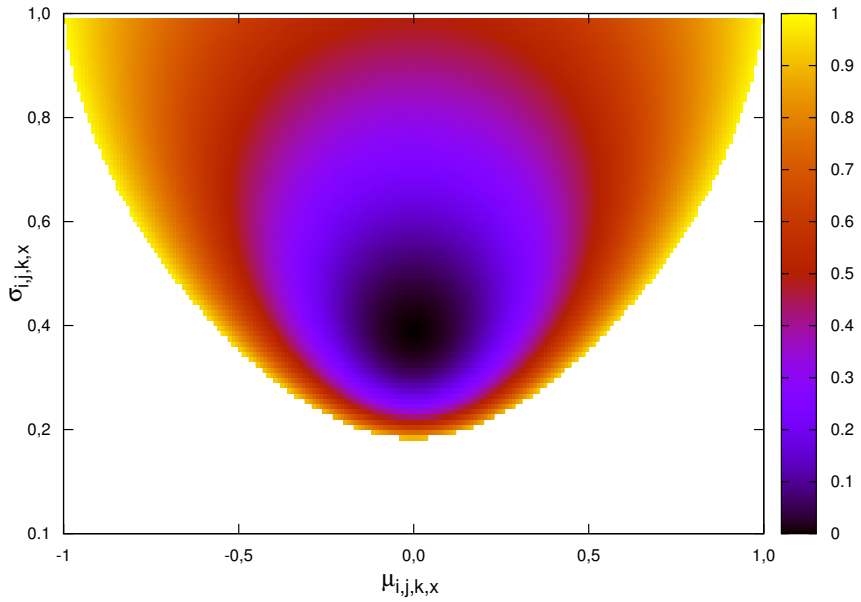

FIGURE 8.2 – $D_{1,2,3,1}$ en fonction de $\mu_{1,2,3,1}$ et de $\sigma_{1,2,3,1}$

que son nombre d'observations augmente ($|O_{1,3,1}| = 20$). Cependant, son erreur standard de la moyenne reste dans un premier temps suffisamment élevée pour qu'il continue à considérer le témoignage de a_2 comme crédible. Lorsque a_1 dispose enfin d'un nombre suffisant d'observations ($|O_{1,3,1}| = 50$), il devient suffisamment confiant dans son estimation pour considérer le témoignage de a_2 comme non crédible.

$\|O_{1,3,1}\|$	$D_{1,2,3,1}$	SEM	δ	Crédibilité de $F_{1,2,3,1}$
5	0.183	0.179	0.41	$\surd$
20	0.183	0.089	0.196	$\surd$
50	0.183	0.057	0.134	$\times$

Tableau 8.3 – Crédibilité du témoignage de l'agent a_2 en fonction de $|O_{1,3,1}|$

Dans toute la suite, nous notons par $KL_i(F_{i,j,k,x})$ (respectivement $\overline{KL_i}(F_{i,j,k,x})$) le fait qu'un témoignage $F_{i,j,k,x}$ est crédible (respectivement non crédible) du point de vue de a_i.

8.2 Filtrer les témoignages non crédibles

En se fondant sur la divergence de Kullback-Leibler et l'erreur standard de la moyenne, l'agent a_i est capable de décider s'il considère comme crédible un témoignage reçu. Dans cette section, nous présentons trois fonctions permettant à l'agent a_i d'utiliser cette mesure de crédibilité. L'objectif de ces fonctions est de réduire l'influence des faux témoignages (définition 6.2.5) lors du calcul des valeurs de réputation afin de renforcer la robustesse du système contre les manipulations.

8.2.1 Des fonctions de KL-filtrage

Pour décider quels témoignages l'agent a_i doit utiliser pour obtenir une estimation du comportement futur d'un agent a_k sans être biaisé par des faux témoignages, nous proposons un processus qui filtre l'ensemble des témoignages reçus afin de ne conserver qu'un sous-ensemble de témoignages crédibles. Ce processus est assuré par une *fonction de filtrage* propre à l'agent a_i. Par abus de notation, nous notons $2^{\mathcal{F}}$ l'ensemble des matrices $\mathcal{F}_i$ possibles, c'est-à-dire l'ensemble des témoignages possibles, car nous les représentons par une matrice $N \times N \times S$.

Définition 8.2.1 - *Fonction de filtrage* : La *fonction de filtrage* de l'agent $a_i \in N$ est la fonction $\phi_i : 2^{\mathcal{F}} \rightarrow 2^{\mathcal{F}}$ qui retourne pour un ensemble de témoignages le sous-ensemble que a_i considère comme crédible.

Désormais, la fonction de réputation (définition 6.1.8) d'un agent a_i reçoit en entrée non plus tous les témoignages mais seuls les témoignages crédibles. Ainsi, la réputation d'un agent a_k selon un agent a_i est donnée par :

$$f_i(a_k,s_x,\phi_i(\mathcal{F}_i))$$

Une première fonction de filtrage intuitive est la fonction qui filtre tout témoignage qui n'est pas crédible au sens de la définition 8.1.4. Nous appelons cette fonction le KL-filtrage.

Définition 8.2.2 - *Fonction de KL-filtrage* : La fonction de *KL-filtrage* est la fonction ϕ_i définie par :

$$\phi_i(\mathcal{F}_i) = \{F_{i,j,k,x} \in \mathcal{F}_i | KL_i(F_{i,j,k,x})\}$$

Exemple 8.2.1 - Considérons un système d'échange de services $\langle N,S \rangle$ tel que $N = \{a_1,a_2,a_3, a_4,a_5\}$, $S = \{s_1\}$ et que $N_1 = \{a_4,a_5\}$. Considérons la matrice de témoignages de l'agent a_1 comme présenté dans la table 8.4 sachant que a_2 est un agent honnête et a_3 un agent malveillant qui diffame a_4 et promeut a_5.

a_j \ a_k	a_4	a_5
a_1	$\{0.5,0.75,0.75,0.5\}$	$\{-0.5,0.25,0\}$
a_2	$\{0.75,0.9,0.75,0.25\}$	$\{-0.25,0.5,-0.5\}$
a_3	$\{-0.75,-0.5,-0.5,-0.75\}$	$\{0.75,0.75,0.75,0.9,0.9\}$

Tableau 8.4 – Matrice de témoignages de l'agent a_1

Supposons que a_1 utilise comme fonction de réputation l'estimation collective (définition 7.3.1). Sans fonction de filtrage, nous avons :

$$f_1(a_4,s_1,\mathcal{F}_i) \simeq 0.220$$
$$f_1(a_5,s_1,\mathcal{F}_i) \simeq 0{,}323$$

Ainsi, les faux témoignages fournis par a_3 permettent à a_5 d'obtenir la meilleure réputation. Cependant, les témoignages de a_2 sont crédibles pour a_1 contrairement aux témoignages de a_3.

Ainsi, l'application d'une fonction de KL-filtrage donne :

$$\phi_1(\mathcal{F}_i) = \begin{pmatrix} \{0.5,0.75,0.75,0.5\} & \{-0.5,0.25,0\} \\ \{0.75,0.9,0.75,0.25\} & \{-0.25,0.5,-0.5\} \end{pmatrix}$$

a_1 calcule alors les valeurs de réputation suivantes :

$$f_1(a_4,s_1,\phi_1(\mathcal{F}_i)) \simeq 0.644$$
$$f_1(a_5,s_1,\phi_1(\mathcal{F}_i)) \simeq -0{,}083$$

Toutefois, dans le cas général, le KL-filtrage n'exclut pas nécessairement tous les faux témoignages. En effet, le KL-filtrage considère la crédibilité dans chaque témoignage d'un agent a_j indépendamment de la crédibilité de ses autres témoignages. Or, si un agent a_i n'a pas toujours suffisamment d'observations lui permettant de juger correctement un témoignage $F_{i,j,k,x}$, il peut en avoir pour juger un autre témoignage F_{i,j,k_2,x_2}, c'est-à-dire pour juger un autre témoignage provenant du même agent mais portant sur un autre couple d'agent et de service. Si nous faisons l'hypothèse qu'un agent qui délivre un faux témoignage a une forte probabilité d'en délivrer un autre, nous pouvons considérer que si a_j n'est pas crédible sur un sous-ensemble de ses témoignages alors aucun de ses témoignages ne doit l'être.

Cette hypothèse nous permet de définir une notion de crédibilité portant non plus sur chaque témoignage mais sur l'agent fournissant les témoignages. Cette mesure consiste à considérer que si l'agent a_i considère que k témoignages d'un agent a_j ne sont pas crédibles alors a_j n'est globalement pas crédible.

Définition 8.2.3 - *Agent k-crédible* : Soit un système d'échange de services $\langle N,S\rangle$, $a_i,a_j \in N$ deux agents et $k \in \mathbb{N}$. L'agent a_j est *k-crédible* du point de vue de a_i si :

$$|\{F_{i,j,k',x} \in \mathcal{F}_i | \overline{KL_i}(F_{i,j,k',x})\}| \leq k$$

Dans toute la suite, nous notons par $KL_i^k(N) \subseteq N$ l'ensemble des agents considérés comme k-crédible par a_i. Cette notion d'agents k-crédibles permet alors de définir une seconde fonction de filtrage plus drastique qui rejette non seulement l'ensemble des témoignages qui ne sont pas crédibles mais aussi les témoignages provenant des agents non k-crédibles. Cette fonction de filtrage est appelée le *filtrage par k-fautes*.

Définition 8.2.4 - *Fonction de filtrage par k-fautes* : Soit un système d'échange de services $\langle N,S\rangle$, $a_i \in N$ un agent et $k \in \mathbb{N}$. La *fonction de filtrage par k fautes* est la fonction ϕ_i définie par :

$$\phi_i(\mathcal{F}_i) = \{F_{i,j,k',x} \in \mathcal{F}_i | a_i \in KL_i^k(N) \wedge KL_i(F_{i,j,k',x})\}$$

Remarquons que même si l'agent a_j est k-crédible, le sous-ensemble de ses témoignages qui ne sont pas crédibles sont tout de même filtrés. Ainsi, le filtrage par k fautes est une généralisation du KL-filtrage. En effet, plus k est proche de 0, moins un agent accepte de témoignages crédibles car l'agent qui les fournit ne l'est pas. Inversement, plus k est grand, plus le filtrage par k fautes est proche du KL-filtrage.

Exemple 8.2.2 - Considérons $\langle N,S\rangle$ avec $N = \{a_1,a_2,a_3,a_4,a_5,a_6\}$ et $S = \{s_1,s_2,s_3\}$ tel que $N_1 = \{a_3,a_2,a_6\}$, $N_2 = \{a_2,a_4,a_5\}$ et $N_3 = \{a_1,a_4,a_5\}$. Supposons que la crédibilité des témoignages et des agents du point de vue de l'agent a_i soit comme présentée sur le tableau 8.5.

	s_1			s_2			s_3			Agent 4-crédible
a_j \ a_k	a_2	a_3	a_6	a_2	a_4	a_5	a_1	a_4	a_5	
a_1	√	√	√	√	√	√	√	√	√	√
a_2	√	×	×	√	×	√	√	√	√	√
a_3	×	×	×	√	√	√	×	√	√	×
a_4	√	×	√	√	√	√	×	√	×	√
a_5	×	×	√	√	×	√	×	×	√	×
a_6	√	×	√	√	√	×	√	√	×	√

Tableau 8.5 – Crédibilité des témoignages et des agents selon l'agent a_1

Supposons que a_1 ne dispose que de peu d'observations vis-à-vis du service s_2 fourni par a_2 et accepte tous les témoignages portant sur $\varepsilon_{2,2}$. À l'inverse, supposons que a_1 dispose d'un grand nombre d'observations vis-à-vis du service s_1 fournit par a_3 et rejette tous les témoignages sur $\varepsilon_{1,3}$. Supposons enfin que a_1 effectue du filtrage par 4-fautes. Ainsi, les agents a_3 et a_5 ayant fourni respectivement 4 et 5 témoignages non crédibles ne sont pas 4-crédible et nous obtenons $KL_1^4(N) = \{a_1,a_2,a_4,a_6\}$. Donc, aucun des témoignages crédibles de a_3 et a_5 n'est intégré dans le calcul de la réputation. Aucun des témoignages non crédibles de a_2, a_4 et a_6 ne sont aussi intégrés.

8.2.2 La stochocratie : un vote aléatoire sur la crédibilité

Si le KL-filtrage (définition 8.2.2) et par k-fautes (définition 8.2.4) permettent à un agent de détecter et filtrer une partie des faux témoignages, la crédibilité de chaque témoignage reste uniquement fondée sur les observations personnelles de l'agent qui la calcule. Or, il est possible que les témoignages des autres agents ne soient pas faux mais que les observations personnelles de celui qui les juge soient erronées. Ceci est le cas en présence de *discrimination*, c'est-à-dire lorsqu'un agent malveillant choisi son comportement en fonction de l'agent avec qui il interagit, de *traîtrises* car elle n'est observée que par ceux avec qui l'agent malveillant a interagi ou plus simplement de bruit dans les observations de l'agent qui juge le témoignage.

C'est pourquoi nous proposons ici une troisième fonction de filtrage permettant à un agent de remettre en cause ses propres observations. Cette fonction s'inspire du système de vote par stochocratie, aussi appelé suffrage par le sort. En politique, la stochocratie désigne un État dont le gouvernement est sélectionné aléatoirement. L'objectif est d'introduire de l'aléa dans la constitution d'instances de décision afin de limiter les risques de collusion a priori et d'assurer une diversité des points de vue [Delannoi et Dowlen, 2010]. Dans notre contexte, nous proposons d'utiliser un mécanisme de stochocratie dans le processus de jugement de chaque témoignage. Il s'agit intuitivement de soumettre chaque témoignage face à un jury de k agents sélectionnés aléatoirement uniformément. Ce jury décide alors par vote majoritaire si le témoignage est crédible.

Définition 8.2.5 - *Crédibilité par k-stochocratie* : Soit un système d'échange de services $\langle N,S\rangle$, $a_i \in N$ un agent, $k \in \mathbb{N}$ et un témoignage $F_{i,j,k',x}$. Le témoignage $F_{i,j,k',x}$ est dit *crédible par k-stochocratie* si, pour un sous-ensemble $N' \subseteq N \setminus \{a_j,a_{k'}\}$ de k agents tirés aléatoirement uniformément, au moins $\lceil k/2 \rceil$ agents de N' jugent $F_{i,j,k,x}$ comme crédible.

Dans la suite, nous appelons *juge* un agent sélectionné par stochocratie pour juger de la crédibilité d'un témoignage. Afin de ne pas inciter certains agents honnêtes à voter stratégiquement, les agents a_j et a_k ne sont pas sélectionnés en tant que juges. En effet, a_j est supposé considérer comme crédible son propre témoignage et a_k peut être tenté de voter pour ou contre le témoignage pour se promouvoir ou éviter une mauvaise réputation. Dans la suite, nous notons $L_i^k(F_{i,j,k',x})$ un témoignage $F_{i,j,k',x}$ jugé crédible par k-stochocratie. Nous pouvons alors définir la fonction de filtrage qui rejette tous les témoignages qui ne sont pas crédibles par k-stochocratie.

Définition 8.2.6 - *Fonction de filtrage par k-stochocratie* : Soit un système d'échange de services $\langle N,S\rangle$, $a_i \in N$ un agent et $k \in \mathbb{N}$. La *fonction de filtrage par k-stochocratie* est la fonction ϕ_i définie par :

$$\phi_i(\mathcal{F}_i) = \{F_{i,j,k',x} \in \mathcal{F}_i | L_i^k(F_{i,j,k',x})\}$$

Notons que les observations de l'agent a_i sont elles aussi soumises au filtrage par k-stochocratie. Ainsi, si a_i a peu d'observations ou une trop grande incertitude sur ses observations, ces dernières peuvent ne pas être utilisées lors du calcul de la réputation.

Cependant, la crédibilité par k-stochocratie est un processus stochastique et deux jugements successifs peuvent rendre des résultats différents. De plus, comme elle repose sur un vote majoritaire, elle peut être influencée par un vote stratégique (définition 1.2.1). En effet, un juge malveillant peut volontairement déclarer comme non crédible un témoignage qu'il considère en réalité comme crédible ou inversement déclarer crédible un témoignage qu'il considère ne pas l'être, afin de promouvoir ou de diffamer un autre agent. Toutefois, nous considérons la crédibilité par k-stochocratie comme robuste car :

1. Nous pouvons simuler les votes pour éviter les votes stratégiques et réduire dans le même temps les coûts de communication. L'hypothèse faite par la stochocratie est qu'un juge honnête $a_{i'}$ vote pour un témoignage $F_{i,j,k,x}$ en mesurant la crédibilité de $F_{i,j,k,x}$ au regard de ses propres observations $O_{i',k,x}$. Or, comme les témoignages sont censés être les observations des agents, si le juge est honnête alors $O_{i',k,x} = F_{i,i',k,x}$. Par conséquent, un agent a_i peut lui-même calculer $D_{i',j,k,x}$ et décider si le témoignage de $F_{i,j,k,x}$ serait jugé comme crédible par $a_{i'}$. Ainsi, bien que nous ayons décrit le filtrage par k-stochocratie comme un vote, il n'en est pas un au sens strict du terme car un agent peut utiliser les témoignages qu'il a reçus pour calculer a priori quel serait le vote de chacun des juges qu'il considère. Cette méthode empêche alors un agent malveillant de voter stratégiquement puisque son vote est formulé à partir de témoignages qu'il a lui-même préalablement fournis et qui sont utilisés à son insu.
2. La probabilité que la majorité des juges soient malveillants est faible. En effet, pour qu'un autre agent malveillant $a_{i'}$ puisse accorder sa voix pour rendre un témoignage (promotion ou diffamation) $F_{i,j,k',x}$ fourni par un agent a_j en collusion avec $a_{i'}$, il faut que $a_{i'}$ soit sélectionné en tant que juge, que $D_{i',j,k',x} \simeq 0$ et qu'au moins $\lceil k/2 \rceil$ des juges aient fourni des témoignages similaires. Cette probabilité est alors caractérisée par la propriété 8.2.1.

Propriété 8.2.1 : Soit $\langle N,S\rangle$ un système d'échange de services, $a_i \in N$ et $F_{i,j,k',x} \in \mathcal{F}_i$ un témoignage. Notons $l \in [0,|N|-2]$ le nombre d'agents $a_z \in N \setminus \{a_j,a_{k'}\}$ tels que $KL_z(F_{i,j,k',x})$ (resp. $\overline{KL_z}(F_{i,j,k',x})$). Le témoignage $F_{i,j,k',x}$ est *jugé par erreur* par k-stochocratie comme crédible (resp. non crédible) alors qu'il est faux (resp. vrai) avec une probabilité $p \in [0,1]$ telle que :

$$p = \sum_{K=\lceil k/2 \rceil}^{k} \frac{\binom{|N'|}{K} \times \binom{|N|-|N'|-2}{k-K}}{\binom{|N|-2}{k}}$$

où $N \setminus \{a_j,a_k\}$ est l'ensemble des agents pouvant être sélectionnés comme juges, k est le nombre de juges sélectionnés, N' l'ensemble des agents en erreur sur le témoignage $F_{i,j,k',x}$, et K le nombre d'agents de N' sélectionnés comme juges.

Démonstration : Considérons le cas où le témoignage $F_{i,j,k',x}$ est un faux témoignage fourni par l'agent malveillant a_j. Soit $N' \subseteq N \setminus \{a_j,a_{k'}\}$ l'ensemble des agents a_z tels que $KL_z(F_{i,j,k',x})$. La probabilité que le témoignage $F_{i,j,k',x}$ soit jugé par k-stochocratie comme crédible correspond à la probabilité qu'au moins $\lceil k/2 \rceil$ parmi les k juges appartiennent à N'. Or, la probabilité que exactement $K \in [0,|N'|]$ des agents de N' soient sélectionnés aléatoirement uniformément parmi les k juges suit une loi hypergéométrique $\mathcal{H}(N \setminus \{a_j,a_k\},k,|N'|)$, soit :

$$\mathbb{P}_K = \frac{\binom{|N'|}{K} \times \binom{|N|-|N'|-2}{k-K}}{\binom{|N|-2}{k}}$$

Pour un ensemble N' fixé, la probabilité $p \in [0,1]$ minimale pour qu'au moins $\lceil k/2 \rceil$ agents de N' soit sélectionnés parmi les k juges est :

$$p = \sum_{K=\lceil k/2 \rceil}^{k} \mathbb{P}_K = \sum_{K=\lceil k/2 \rceil}^{k} \frac{\binom{|N'|}{K} \times \binom{|N|-|N'|-2}{k-K}}{\binom{|N|-2}{k}}$$

Un raisonnement identique peut être tenu dans le cas où le témoignage $F_{i,j,k',x}$ est un vrai témoignage. Il suffit de considérer cette fois N' comme l'ensemble des agents a_z tels que $\overline{KL_z}(F_{i,j,k',x})$. □

Le tableau 8.6 donne la probabilité qu'un faux témoignage soit jugé par k-stochocratie comme crédible en fonction de $|N|$, $|N'|$ et de k. Dans les cas où $k > |N|-2$ nous considérons que l'agent a_i simule un vote majoritaire parmi tous les agents de $N \setminus \{a_j,a_{k'}\}$.

Bien qu'il y ait des cas particuliers où la probabilité d'erreur soit nulle ou égale à 1, la probabilité d'erreur est faible dans un contexte réaliste où $|N|/2 > |N'| > k/2$. Par exemple $|N| = 100$, $|N'| = 10$, $k = 10$, cette probabilité est de 0,0008. Il est intéressant de remarquer que lorsque le nombre d'agents dans l'erreur est majoritaire (troisième colonne pour chaque valeur de $|N|$) la parité de k influe sur la probabilité d'erreur puisqu'il faut réussir un vote majoritaire. Cela est en particulier visible sur lorsque $|N| = 100$ et $|N'| = 50$. Ainsi, la k-stochocratie est une fonction de filtrage qui permet à un agent de réduire le taux de faux témoignages utilisés lors du calcul de la valeur de réputation tout en remettant en cause les observations de l'agent, et ce avec une faible probabilité d'erreur.

	$\|N\| = 25$		$\|N\| = 50$			$\|N\| = 100$		
k \ $\|N'\|$	5	15	10	20	30	10	25	50
1	0,217	0,652	0,208	0,417	0,625	0,102	0,255	0,51
10	0,007	0,963	0,022	0,401	0,899	0,0007	0,073	0,656
25	0	1	0	0,111	0,97	0	0,0008	0,547
50	0	1	0	0	1	0	9×10^{-9}	0,658
75	0	1	0	0	1	0	0	0,642
98	0	1	0	0	1	0	0	1

Tableau 8.6 – Probabilité d'erreur du filtrage par k-stochocratie en fonction de $|N|$, $|N'|$ et k

8.3 Évaluation des fonctions de filtrage

8.3.1 Protocole d'évaluation

Le protocole de cette étude empirique est similaire à celui utilisé pour évaluer l'influence des politiques de sélection sur la robustesse du système de la section 7.3.1. Nous considérons toujours $|N| = 100$ et $|S| = 10$ où l'expertise de chaque agent est tirée aléatoirement uniformément dans $[-1; 1]$ pour 0 à 5. Nous considérons ici que la politique de sélection utilisée par les agents du système est la politique UCB.

Comme précédemment, nous considérons les agents malveillants mettent en œuvre une attaque oscillante afin de maximiser le regret des agents honnêtes et considérons comme fonction de réputation l'estimation collective, BetaReputation et FlowTrust. Nous ne considérons pas ici EigenTrust car il suffit d'un chemin de confiance entre un *agent de confiance* et un agent malveillants pour que la réputation des agents soit faussée [Cheng et Friedman, 2006]. Ainsi, excepté le cas où tous les témoignages des agents malveillants sont filtrés, la fonction de filtrage ne permet pas d'accroître la robustesse du système.

Sur ces trois fonctions de réputation, nous étudions l'influence des fonctions filtrage en comparant des agents qui n'utilisent pas de fonctions de filtrage, du KL-filtrage, du filtrage par 10-fautes et par 10-stochocratie. Pour des raisons de lisibilité des résultats, notre analyse se porte sur les cas où $k = 10$ qui est le paramètre où les résultats sont les plus significatifs. Nous comparons nos résultats avec l'estimation personnelle. Enfin, nous réitérons 50 fois les simulations et mesurons le regret moyen des agents honnêtes. Comme l'utilisation de la KL-divergence comme mesure de crédibilité dépend du nombre d'observations dont les agents disposent, nous considérons ici deux scénarios.

Scénario 1 : les 100 agents sont tous nouveaux dans le système et n'ont donc aucune information pour juger de la crédibilité des témoignages lors des premiers pas de temps. Nous mesurons l'évolution de leur regret lors des 200 premiers pas de temps. Ce scénario est appelé fonctionnement initial du système.

Scénario 2 : les 100 agents ont déjà interagi durant 100 pas de temps lorsque 20 nouveaux agents honnêtes rejoignent le système. Ces nouveaux agents n'ont donc aucune connaissance a priori sur les autres agents et utilisent les témoignages qu'ils reçoivent pour calculer les valeurs de réputation. Nous mesurons l'évolution du regret moyen de ces 20 agents durant les 200 pas de temps suivant. Ce scénario est appelé fonctionnement nominal du système.

Comme nous l'avons montré dans les exemples de la section 8.2, la crédibilité et la crédibilité par k-stochocratie peuvent filtrer de vrais témoignages et ne pas filtrer de faux témoignages. Nous proposons donc de prendre en compte deux mesures classiques dans le domaine de la classification et du filtrage collaboratif : le *rappel* et la *précision* [Bramer *et al.*, 2007]. Le rappel correspond au taux de faux témoignages filtrés et la précision désigne le ratio de faux témoignages filtrés parmi l'ensemble de tous les témoignages filtrés.

Définition 8.3.1 - *Rappel* : Soit $a_i \in N$ un agent du système. Soit TP l'ensemble des faux témoignages filtrés et FN l'ensemble des faux témoignages considérés comme crédibles par a_i. Le *rappel* de la fonction de filtrage ϕ_i est :

$$\text{rappel}(\phi_i) = \frac{|TP|}{|TP| + |FN|}$$

Définition 8.3.2 - *Précision* : Soit $a_i \in N$ un agent du système. Soit TP l'ensemble des faux témoignages filtrés et TN l'ensemble des vrais témoignages considérés comme non crédibles par a_i. La *précision* de la fonction de filtrage ϕ_i est :

$$\text{precision}(\phi_i) = \frac{|TP|}{|TP| + |TN|}$$

Ces deux mesures nous permettent de déterminer si nos propositions évaluent correctement la crédibilité des témoignages. Nous présentons ici l'évolution du rappel et de la précision des fonctions de filtrage sur nos systèmes lors de la phase d'initialisation afin de mettre en lumière le nombre d'observations nécessaires aux agents pour détecter les faux témoignages.

8.3.2 Évaluation du regret

Les figures 8.3, 8.4, et 8.5 montrent l'évolution du regret des agents sur le scénario 1. Globalement, les fonctions de KL-filtrage et par 10-fautes donnent lors des premiers pas de temps le même regret. En revanche, après quelques pas de temps, les agents malveillants ayant produit plus de 10 témoignages non crédibles ne sont plus pris en compte et nous pouvons constater une diminution de regret avec le filtrage par 10-fautes. Nous pouvons aussi observer des oscillations qui coïncident avec les périodes de blanchiments et traîtrises.

L'estimation collective (figure 8.3) est une fonction de réputation triviale très sensible aux manipulations, et plus particulièrement aux faux témoignages. Ainsi, en l'absence de fonction de filtrage, il est alors plus intéressant d'utiliser une estimation personnelle. En revanche, les fonctions de filtrage réduisent fortement l'influence des faux témoignages. Ici, la 10-stochocratie est la fonction de filtrage qui minimise le regret des agents, car les agents peuvent remettre en cause rapidement leurs observations qui sont peu nombreuses à l'initialisation.

BetaReputation (figure 8.4) est un système robuste et par conséquent le regret des agents, avec ou sans fonction de filtrage, est presque identique. Cependant, les filtrages par 10-fautes et par 10-stochocratie sont plus efficace que le KL-filtrage qui tend à être légèrement supérieur à l'absence de filtrage. Ceci s'explique par le fait que le KL-filtrage est uniquement fondée sur des approximations par la loi normale de l'expertise alors que la réputation des agents est calculée ici en considérant une fonction de densité beta.

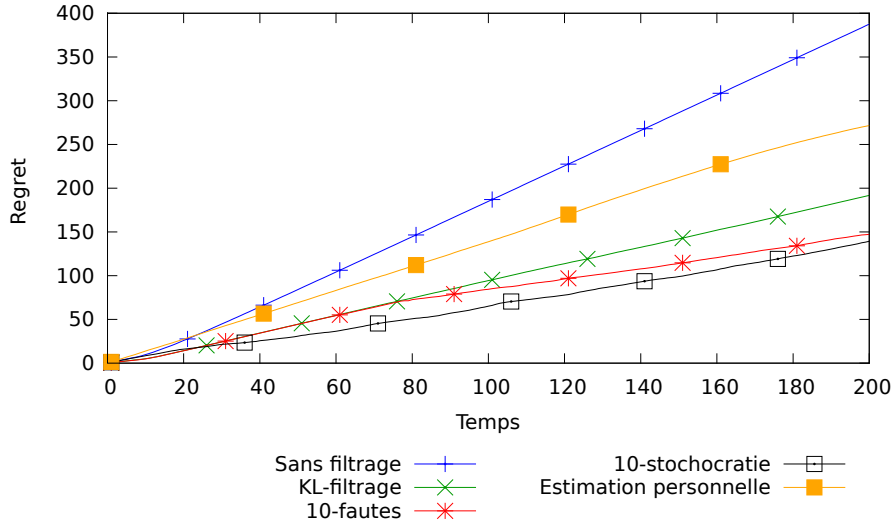

FIGURE 8.3 – Scénario 1 : regret moyen des agents pour l'estimation collective

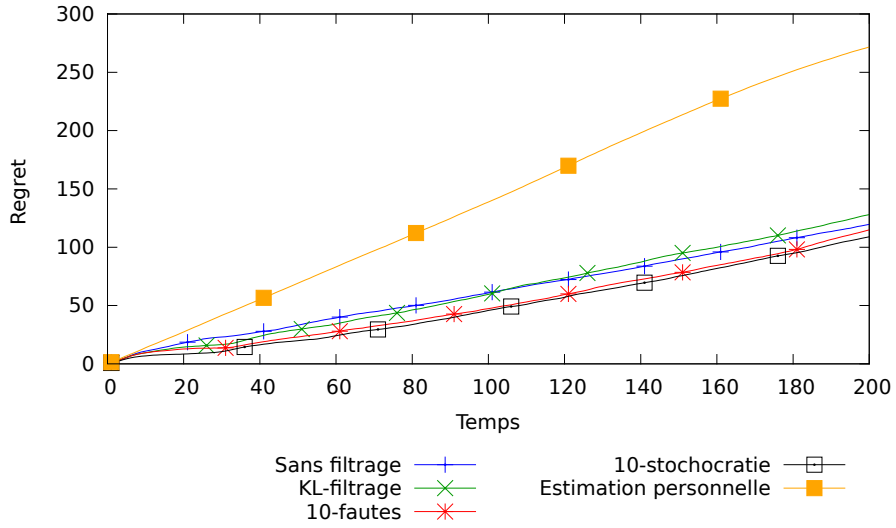

FIGURE 8.4 – Scénario 1 : regret moyen des agents pour BetaReputation

FlowTrust (figure 8.5) est sensible aux promotions. Ainsi, l'utilisation des fonctions de filtrage permet de réduire le regret des agents. Notons cependant que contrairement aux cas précédents, la 10-stochocratie n'est pas la fonction de filtrage la plus performante et le filtrage par 10 fautes est le plus efficace. En effet, comme la valeur de réputation sur FlowTrust n'est pas une estimation de l'espérance de gains, UCB n'est pas aussi efficace dans son exploration et tend interagir avec les mêmes agents et les promotions visant ces agents sont plus rapidement détecté par le filtrage par 10 fautes, car chaque agent dispose d'un plus grand nombre d'observations.

Sur les figures 8.6, 8.7, et 8.8 nous présentons l'évolution du regret moyen des agents sur le scénario 2. Globalement, le filtrage est plus efficace que dans le scénario 1. De plus, la 10-

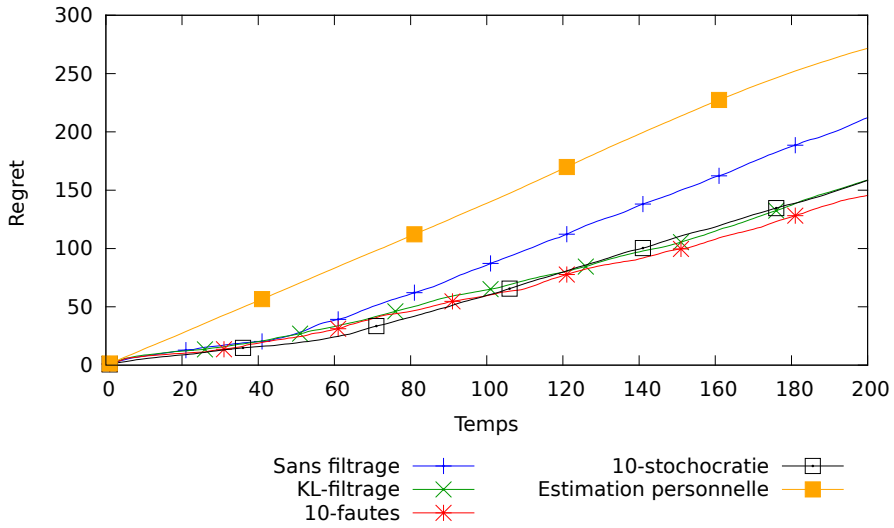

FIGURE 8.5 – Scénario 1 : regret moyen des agents pour FlowTrust

stochocratie est clairement la fonction de filtrage qui minimise le regret sur ce scénario. En effet, les agents qui viennent de rejoindre le système n'ont pas besoin d'observations propres pour juger de la crédibilité d'un témoignage et peuvent utiliser les expériences passées des autres agents qui disposent alors en moyenne de suffisamment d'observations pour juger de la crédibilité des différents témoignages.

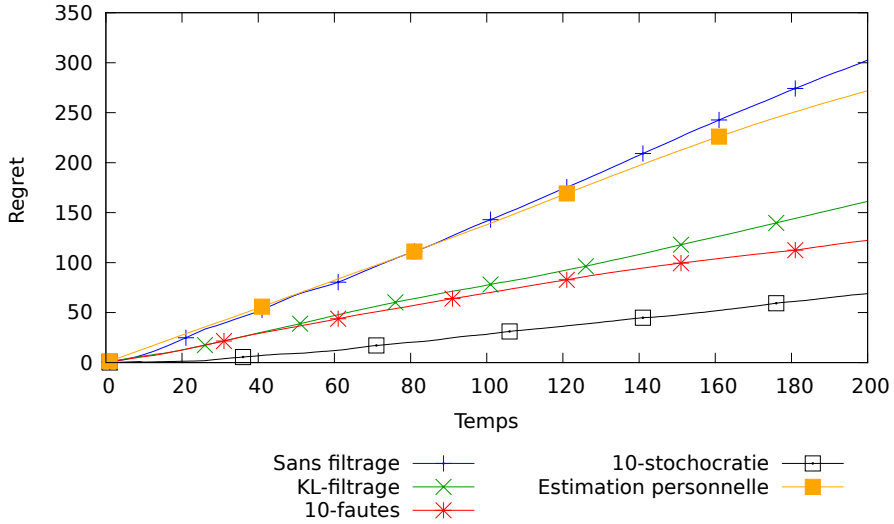

FIGURE 8.6 – Scénario 2 : regret moyen des agents pour l'estimation collective

L'estimation collective (figure 8.6) présente un comportement intéressant : elle est aussi robuste sans fonction de filtrage que l'estimation personnelle. Ceci s'explique par le fait que les agents honnêtes présents lors de l'initialisation fournissent suffisamment de témoignages pour compenser en partie les faux témoignages fournis par les agents malveillants.

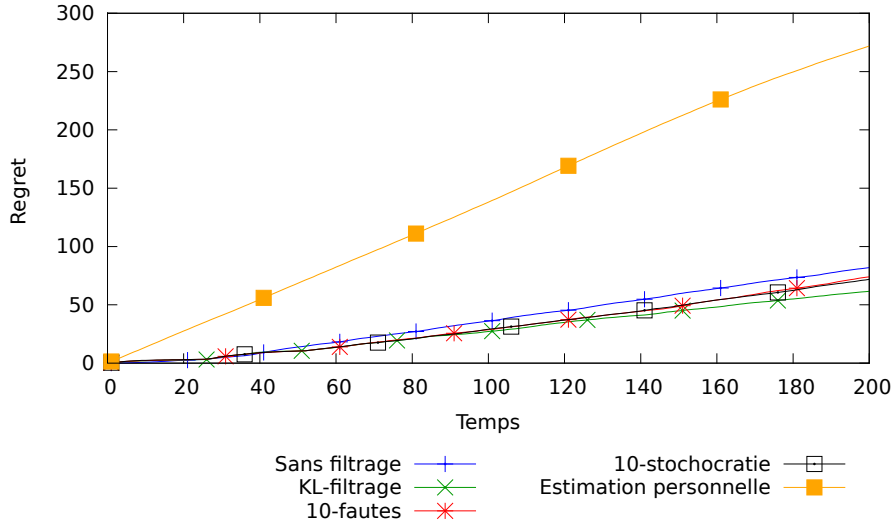

FIGURE 8.7 – Scénario 2 : regret moyen des agents pour BetaReputation

Si BetaReputation (figure 8.7) est plus performante sur le scénario 2 que sur le scénario 1, cela ne semble pas lié à l'utilisation des fonctions de filtrage mais simplement au fait que les agents utilisent les expériences des autres pour calculer les valeurs de réputation. Cependant, le filtrage par 10-stochocratie reste plus efficace dans les premiers pas de temps.

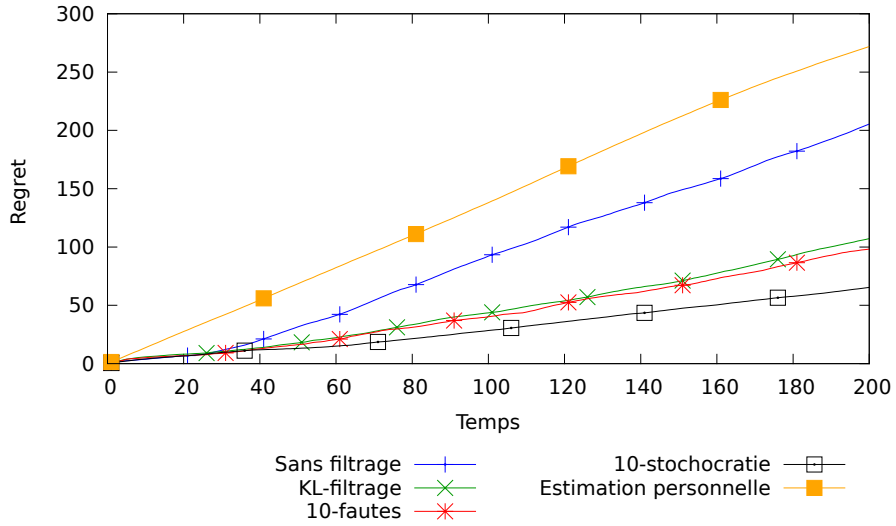

FIGURE 8.8 – Scénario 2 : regret moyen des agents pour FlowTrust

FlowTrust (figure 8.8) bénéficie clairement sur le scénario 2 de l'utilisation des fonctions de filtrage. Contrairement au scénario 1 où le filtrage par 10-fautes était le plus performant, la 10-stochocratie est très efficace sur FlowTrust.

8.3.3 Rappel et précisions des fonctions de filtrage

Le figure 8.9 présente l'évolution du rappel moyen des fonctions de filtrage sur le scénario 1. Nous pouvons remarquer de manière générale que la majorité des faux témoignages sont détectés en 40 et 60 pas de temps. Auparavant, les agents n'ont que peu d'observations et leurs erreurs standards de la moyenne sont suffisamment élevées pour qu'un faux témoignage soit considéré comme crédible. Initialement, le K-filtrage et le filtrage par 10-fautes ont des résultats similaires mais, autour des pas de temps 20 à 40, le filtrage par 10-fautes permet de détecter très rapidement la majorité des faux témoignages. Le filtrage par 10-stochocratie est particulièrement performant car il détecte les faux témoignages même durant le période d'initialisation. Nous pouvons cependant remarquer que, sur le long terme, le filtrage par 10-fautes tend à être le plus efficace.

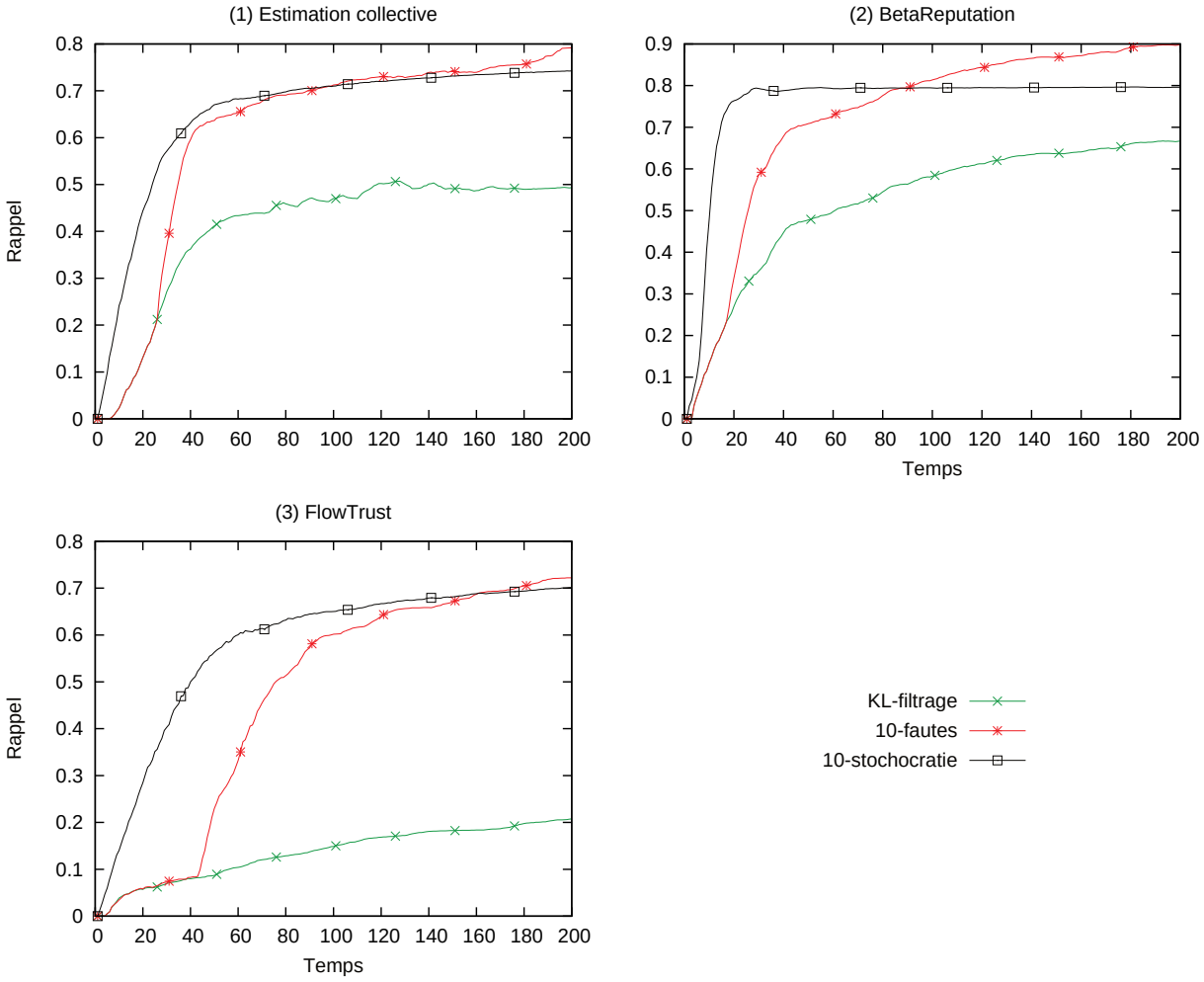

FIGURE 8.9 – Rappel moyen des fonctions de filtrage

D'un point de vue spécifique, le KL-filtrage est très peu efficace sur FlowTrust avec à peine 20 % de faux témoignages détectés. Cela est dû au fait qu'UCB incite sur FlowTrust les agents à interagir avec les agents malveillants fournissant de bons services (ceux du groupe M_1) et plus

rarement avec les autres agents honnêtes. En conséquence, les agents manquent d'observation pour détecter les diffamations (produites par les agents de M_2). Sur BetaReputation, les fonctions de filtrage sont très efficaces, en particulier le filtrage par 10-fautes qui dépasse rapidement le filtrage par 10-stochocratie. Cela est du au fait que BetaReputation est plus robuste aux manipulations. Ainsi, le facteur d'exploration prend un poids plus important dans les premiers de temps et permet d'interagir avec les agents honnêtes. Ainsi, les agents obtiennent plus rapidement des observations correctes leur permettant de filtrer les témoignages.

La figure 8.10 présente l'évolution de la précision des fonctions de filtrage sur le scénario 1. De manière générale, la précision des fonctions de filtrage décroît initialement pour augmenter à nouveau. Ceci est dû au fait que les agents utilisent dans les premiers pas de temps leurs propres observations pour juger de la crédibilité des témoignages : si deux agents honnêtes ont des observations opposées alors ils jugent non crédibles leurs témoignages. Le filtrage par 10-fautes est plus efficace que le KL-filtrage. Cependant, il convient de remarquer que le filtrage par 10-stochocratie n'est que peu sujet à ce manque initial d'informations car les erreurs de jugement ont une faible probabilité d'occurence.

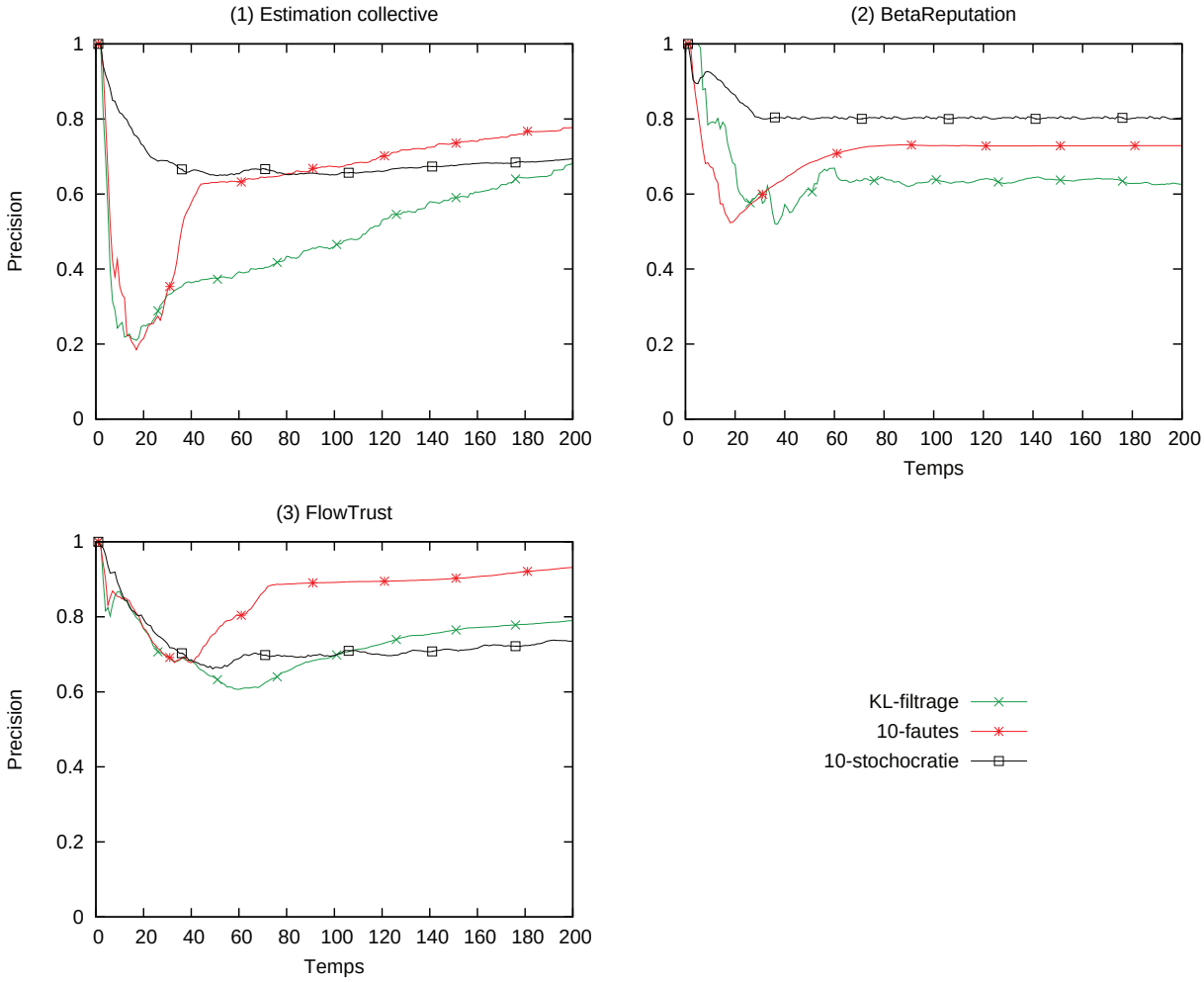

FIGURE 8.10 – Precision moyenne des fonctions de filtrage

Sur les trois fonctions de filtrage, une partie des vrais témoignages sont donc considérés comme non crédibles. Il s'agit en effet des témoignages portant sur des agents honnêtes ayant une faible expertise car peu d'agents interagissent avec eux et obtiennent suffisamment d'observations pour considérer comme crédible des témoignages divergents mais pourtant corrects. D'un point plus spécifique, les fonctions de filtrage ont une meilleure précision sur FlowTrust car moins de témoignages sont globalement filtrés et le KL-filtrage met de nombreux pas de temps à être efficace sur l'estimation collective.

Ainsi, notre mesure de crédibilité et nos fonctions de filtrage permettent de détecter et filtrer efficacement les faux témoignages fournis par les agents malveillants car en moyenne sur toutes les fonctions de filtrage et toutes les fonctions de réputation 80 % des faux témoignages sont filtrés et que 80 % des témoignages filtrés sont faux.

8.4 Conclusion

Dans ce chapitre, nous avons proposé une approche pour lutter contre les *faux témoignages*. Nous avons proposé une nouvelle mesure de crédibilité des témoignages fondée sur la *divergence de Kullback-Leibler* et l'*erreur standard de la moyenne*. Les témoignages considérés comme non crédibles sont alors écartés par des *fonctions de filtrage* lors du calcul de la réputation.

Notre notion de crédibilité repose sur la divergence entre les témoignages et les observations propres d'un agent lorsque les facteurs d'expertise sont approximés par une loi normale. En effet, deux agents honnêtes évaluant la qualité des services sur des critères communs doivent alors avoir des approximations similaires. Un témoignage est alors considéré comme crédible si la divergence entre les observations de l'évaluateur et le témoignage évalué est inférieure à un seuil fondé sur l'erreur standard de la moyenne.

À partir de la KL-crédibilité, nous avons proposé trois *fonctions de filtrage* qui écartent les témoignages non crédibles lors du calcul de la réputation. Le *KL-filtrage* consiste simplement à n'utiliser que les témoignages crédibles. Le *filtrage par k-fautes* est une généralisation du précédent permettant d'écarter tous les témoignages d'un agent ayant fourni au moins k témoignages non crédibles et ainsi compenser un manque d'information du à l'absence d'interaction. Enfin, le *filtrage par k-stochocratie* consiste à sélectionner aléatoirement des agents chargés de juger de la crédibilité d'un témoignage. Cette dernière fonction permet ainsi à un agent de remettre en cause ses propres observations et compenser un éventuel bruit.

Nous avons évalué empiriquement les performances de ce filtrage sur trois fonctions de réputation, l'estimation collective, BetaReputation et FlowTrust, en prise avec une attaque oscillante. Les trois fonctions de filtrage permettent de réduire significativement le regret des agents honnêtes. De plus, l'évaluation du rappel et de la précision du filtrage nous permet de dire que nos propositions filtrent efficacement les faux témoignages.

Chapitre 9

Conclusion et perspectives

Sommaire

Résumé.
Les travaux présentés dans ce manuscrit portent sur l'utilisation de stratégies permettant à des agents pleinement autonomes et rationnels de manipuler un système multi-agent afin d'atteindre leurs objectifs aux dépens des autres agents. Nous avons étudié ce problème en considérant deux familles de systèmes multi-agents : les *jeux de coalitions hédoniques* et les *systèmes de réputations*. Dans ce dernier chapitre, nous récapitulons l'ensemble de nos contributions et nous proposons une liste de perspectives qui nous paraissent intéressantes à approfondir.

9.1 Contributions

Dans ce manuscrit, nous nous sommes intéressés aux problèmes de manipulation dans les systèmes multi-agents. Plus spécifiquement, nous nous sommes intéressés aux systèmes multi-agents ouverts (des agents hétérogènes peuvent rejoindre le système) et décentralisés (sans autorité centrale de contrôle). Dans ces systèmes, des agents pleinement autonomes prennent des décisions rationnelles, qu'elles soient collectives ou individuelles. Cependant, un agent peut ne pas se satisfaire des décisions prises par les autres agents et être incité à les manipuler.

De manière générique, une manipulation est une stratégie mise en œuvre par un agent (ou des agents en collusion) permettant d'influencer le processus de décision d'un ensemble d'agents du système à l'aide de fausses informations afin que ces derniers prennent des décisions favorables pour le manipulateur. Des travaux antérieurs tendent à montrer qu'il n'existe pas de système parfaitement robuste à toute manipulation sans remettre en cause des propriétés désirables du

système. Par ailleurs, malgré des preuves de complexité dans le pire cas, en pratique, la mise en œuvre de certaines des manipulations est simple. C'est pourquoi nous avons étudié dans ce manuscrit les manipulations au regard des propriétés dont elles ont besoin pour être mises en œuvre de manière efficace ainsi que les stratégies de défense qui viennent renforcer ou affaiblir ces propriétés en fonction de leur influence sur les manipulations. En effet, la robustesse aux manipulations d'un système multi-agent passe par :

1. l'identification des conditions permettant de rendre inefficaces des manipulations simples en pratique ;
2. la proposition de stratégies de défense à adjoindre à un système multi-agent qui n'affaiblissent pas ses propriétés fondamentales.

Pour cela, nous avons considéré deux catégories de systèmes, les *jeux de coalitions hédoniques* et les *systèmes de réputation*, aux propriétés complémentaires : les jeux de coalitions hédoniques modélisent des problèmes statiques de décision collective, les systèmes de réputation des problèmes dynamiques de décision individuelle.

9.1.1 Contributions dans les jeux hédoniques

Pour étudier la robustesse des jeux hédoniques aux manipulations, nous avons proposé un *modèle générique de manipulation* fondé sur la construction de faux profils de préférence et l'introduction d'agents Sybil (chapitre 3). Dans ce modèle, une manipulation n'a d'intérêt à être mise en œuvre que si et seulement si elle est *rationnelle*, c'est-à-dire qu'elle amène à un nouveau jeu et que ce dernier est plus favorable à l'agent manipulateur.

Nous avons alors caractérisé à l'aide de ce modèle les *conditions minimales nécessaires à la rationalité de toutes manipulations* sur un jeu hédonique utilisant la *stabilité au sens de Nash* comme concept de solution (chapitre 4). Nous avons alors montré que ce concept de solution est *robuste aux manipulations* au sens où décider de la rationalité d'une manipulation est un *problème NP-complet*. La robustesse des jeux hédoniques utilisant la stabilité au sens de Nash a été renforcée par une étude empirique montrant que les conditions nécessaires à la rationalité d'une manipulation sont rarement satisfaites. Par ailleurs, cette robustesse aux manipulations a été prouvée dans un contexte favorable pour un agent manipulateur, et la relaxation de l'une de nos hypothèses nous a permis de montrer des conditions nécessaires à la rationalité de la manipulation plus restrictive renforçant nos résultats.

Nous avons appliqué la même méthodologie aux concepts de solution de *stabilité individuelle et de stabilité au sens du cœur* (chapitre 5). Pour ces deux concepts de solution , nous avons exhibé des manipulations simples à mettre en œuvre et dont les conditions nécessaires à la rationalité sont fréquentes. Ainsi, ces deux concepts ne sont *pas robustes aux manipulations*.

Le tableau 9.1 récapitule les résultats de cette étude.

9.1.2 Contributions dans les systèmes de réputation

Afin d'étudier la robustesse des systèmes de réputation, nous avons proposé un modèle générique d'interaction entre agents où ces derniers utilisent une fonction de réputation afin d'estimer le comportement des autres agents (chapitre 6). Les études antérieures portant sur les manipulations dans les systèmes de réputation nous ont permis d'identifier les principales propriétés

12. Ces résultats de complexité n'ont pas été prouvés dans ce manuscrit mais se fondent sur les travaux de [Peters et Elkind, 2015].

Concept de solution	Nash	Individuel	Cœur
Manipulations simples	constructive, destructive	constructive	constructive, destruction du cœur
Condition de la rationalité	Minimales, rares	Fréquentes	Fréquentes
Complexité de décision	NP-complet	NP-complet[12]	NP-complet[12]
Robustesse du système	√	×	×

Tableau 9.1 – Robustesse aux manipulations des concepts de solution

sur lesquels se fondent les agents manipulateurs. Nombre de stratégies de défense issues de la littérature amènent alors à affaiblir certaines de ces propriétés.

Dans ce manuscrit, nous avons proposé de renforcer l'un de ces axiomes des systèmes de réputation, stipulant que le processus de décision des agents doit être guidé par les résultats d'interaction. Cette stratégie de défense consiste à utiliser les politiques de sélection de *bandits manchots* en tant que processus de décision (chapitre 7) et de considérer la réputation d'un agent comme une estimation de la récompense apportée lors d'une future interaction avec ce dernier. Nous avons montré empiriquement que l'introduction d'un *facteur d'exploration* adapté à la fonction de réputation permet de renforcer sa robustesse aux manipulations, évaluée par une réduction du *regret* des agents et une augmentation du *coût des manipulations*.

Nous avons aussi proposé de renforcer la robustesse des fonctions de réputation aux faux témoignages. Pour cela, nous avons proposé une nouvelle *mesure de crédibilité* fondée sur la *divergence de Kullback-Leibler* et l'*erreur standard de la moyenne* (chapitre 8). Cette mesure de crédibilité nous permet de considérer d'un côté le gain d'information apporté par un témoignage par rapport aux connaissances des agents et, d'un autre côté, une incertitude sur ces connaissances. Nous avons ensuite intégré cette notion de crédibilité dans des *fonctions de filtrage* qui écartent de manière efficace les témoignages non crédibles du calcul de la réputation.

Le tableau 9.2 positionne nos propositions selon les critères donnés par [Hoffman *et al.*, 2009] pour catégoriser les stratégies de défense dans les systèmes de réputation.

De manière générique, l'utilisation de ces deux stratégies nous permet d'améliorer la robustesse des fonctions de réputation issues de la littérature en renforçant leurs propriétés fondamentales identifiées par [Resnick *et al.*, 2000]. L'utilisation d'un facteur d'exploration permet de garantir la propriété d'ouverture du système tout en réduisant les mauvaises interactions. L'utilisation d'une fonction de filtrage s'attaque directement à l'utilisation de faux témoignages en ne prenant pas en compte ces derniers lors du calcul des valeurs de réputation.

9.2 Perspectives

Notre travail nous amène à nous poser de nouvelles questions auxquelles il serait intéressant de répondre. Le tableau 9.3 présente nos perspectives selon trois axes identiques pour les deux types de système que nous avons considérés : une généralisation de nos résultats afin d'améliorer leur robustesse, une articulation des deux modèles en construisant les profils de préférence des agents à partir d'un système de réputation et enfin l'utilisation de nos résultats pour manipuler les manipulateurs.

	Stratégies de défense						
	Centralisation	**Statistique**	**Heuristique**	**Redondance**	**Aléatoire**	**Cryptographie**	**Preuve formelle**
UCB		√					√
ε-gloutonne		√	√		√		
ε-élitiste		√	√		√		
KL-filtrage		√					
k-fautes		√	√				
k-stochocratie		√		√	√		√

Tableau 9.2 – Classification des stratégies de défense selon les critères de [Hoffman *et al.*, 2009]

	Jeux hédoniques	**Systèmes de réputation**
Généralisation des résultats	un modèle de connaissances incomplètes	une méta-fonction de filtrage
Articulation des modèles	utiliser des fonctions de réputation pour construire les profils de préférence	
Manipuler les manipulateurs	poser un dilemme du prisonnier	fournir de faux témoignages honnêtes

Tableau 9.3 – Une vision synthétique de nos perspectives

9.2.1 Généralisation des résultats

Connaissances incomplètes

Dans notre étude des jeux hédoniques, nous avons fait l'hypothèse que les agents malhonnêtes ont une connaissance complète des préférence des autres agents. Cette hypothèse forte leur permet de décider s'il est rationnel de mettre en œuvre une manipulation. Qu'en est-il lorsque cette hypothèse est remise en cause ?

Pour cela, il serait intéressant de modéliser explicitement les connaissances d'un agent a_i par un ensemble de tuples $\langle a_j, \succeq_{a_j}, C_1, C_2 \rangle$ représentant le fait que l'agent a_i sait que $C_1 \succeq_j C_2$. L'introduction de cette *représentation partielle des connaissances* nous amène alors à redéfinir la notion de rationalité d'une manipulation. Une intuition consiste à considérer la rationalité des manipulations non plus de manière absolue mais selon une probabilité de rationalité. Dans ce cas, la robustesse d'un jeu hédonique reposerait sur une caractérisation des connaissances minimales nécessaires pour qu'un agent malhonnête puisse décider de la rationalité d'une manipulation avec une certaine probabilité. Une telle caractérisation permettrait d'étudier de manière plus fine la robustesse des concepts de solution comme par exemple la stabilité individuelle que nous avons identifiée comme sensible aux manipulations sous l'hypothèse de connaissances complètes.

Méta-fonction de filtrage

Dans le cadre des systèmes de réputation, nous avons montré que l'efficacité des fonctions de filtrage n'est pas la même selon le contexte. Par exemple, le filtrage par k-stochocratie est efficace lors des premiers pas de temps car il permet de compenser le manque d'informations directes. Par contre, après quelques pas de temps, le filtrage par k-fautes permet d'obtenir un meilleur rappel et une meilleure précision. Il serait alors intéressant de généraliser nos fonctions de filtrage en une unique fonction prenant en paramètres un ensemble de témoignages, un ensemble d'agents devant juger de la crédibilité, une mesure de crédibilité et une règle de vote.

Cette *méta-fonction* nous permettrait de capturer les différentes fonctions de filtrage présentées dans ce manuscrit comme des instances d'une même fonction aux paramètres différents. Par exemple, le filtrage par KL-divergence d'un agent a_i est la fonction ayant le paramétrage :

- ensemble des témoignages : $\mathcal{F}_i$;
- ensemble des juges potentiels : $\{a_i\}$;
- mesure de crédibilité : divergence de Kullback-Leibler et erreur standard de la moyenne ;
- règle de vote : vote majoritaire.

Cette approche permettrait aux agents d'adapter dynamiquement la fonction de filtrage en variant ses paramètres selon le contexte et de considérer de nouvelles fonctions de filtrage agrégeant par exemple du filtrage par k-stochocratie et du filtrage par k-fautes où seuls les agents k-crédibles peuvent être sélectionnés comme juges. De plus, une généralisation de la fonction de filtrage permettrait une meilleure exploration expérimentale de l'espace des fonctions de filtrage et ainsi déterminer quels paramètres ont une influence positive et significative sur la robustesse du système aux manipulations.

9.2.2 Articulation des modèles

Dans le contexte des jeux hédoniques, nous nous sommes limité à un contexte statique qui est le contexte canonique de tels jeux. Il serait alors intéressant de généraliser ces résultats dans un contexte dynamique de jeux hédoniques répétés. Ceci nous permettrait de faire un lien explicite entre jeux hédoniques et systèmes de réputation. Nous pourrions considérer une succession de jeux hédoniques où les agents mettent à jour leur profil de préférence en fonction du résultat des interactions passées entre les membres d'une coalition. L'introduction d'un système de réputation dans ce jeu de coalitions serait alors naturelle. Intuitivement, la répétition des jeux amène une stabilité dans les profils de préférence de chaque agent. Un changement important dans ces derniers serait alors signe de la mise en œuvre d'une manipulation et pourrait être rapidement détecté.

L'articulation des jeux hédoniques et des systèmes de réputation peut également être considérée d'un autre point de vue : est-il possible de construire une fonction de réputation robuste aux manipulations en utilisant des concepts de solution telle que la stabilité au sens de Nash ? Intuitivement, il s'agirait de construire dans le système d'échange de service des profils de préférences en se fondant sur les valeurs de confiance et utiliser comme fonction d'agrégation un concept de solution plutôt qu'une fonction de réputation. Chaque agent pourrait alors décider avec quel autre agent interagir en fonction de la coalition à laquelle il appartient. Cependant, comment construire un profil de préférence collectif à partir d'une structure de coalitions stable ? Comment construire un tel profil de préférence collectif en l'absence de structure stable ? Quelle est la complexité algorithmique de cette approche ?

9.2.3 Manipuler les manipulateurs

Les agents manipulateurs se fondent toujours sur certaines propriétés du système pour construire leurs manipulations. Les agents honnêtes ne pourraient-ils eux aussi utiliser ces propriétés et mettre en œuvre des *manipulations honnêtes* afin de perturber les agents manipulateurs ?

Poser un dilemme du prisonnier

Comme nous avons caractérisé les conditions nécessaires à la rationalité d'une manipulation, il serait possible de piéger un agent manipulateur dans le contexte de jeux hédoniques répétés. Par exemple, un agent honnête ou une autorité de régulation pourrait créer artificiellement un *pot de miel*, c'est-à-dire un jeu où il serait rationnel de manipuler et de détecter ainsi si l'un des participants met en œuvre une manipulation. Sachant cela, un agent manipulateur serait confronté à une variante du *dilemme du prisonnier* : comment décider s'il s'agit effectivement d'un jeu manipulable ou s'il s'agit d'un pot de miel ? La mise en œuvre de cette stratégie de défense pose cependant d'autres questions :

- quel est le coût de ce dilemme sur les agents honnêtes ?
- quelles doivent être les pénalités à mettre en œuvre pour que la meilleure stratégie d'un agent manipulateur lors d'un tel dilemme soit une stratégie mixte ?
- sans autorité de certification, permettre une telle stratégie ne peut-il pas amener un agent malveillant à construire lui aussi un pot de miel pour pénaliser les agents honnêtes ?

Fournir de faux témoignages honnêtes

Certaines manipulations se fondent sur les connaissances des agents manipulateurs pour décider quand changer de comportement. Par exemple, un agent manipulateur à intérêt à se blanchir lorsque sa valeur de réputation est trop faible. Il serait alors intéressant de permettre aux agents honnêtes de diffuser des *faux témoignages honnêtes* pour contrecarrer ce type de manipulation. Intuitivement, une diffamation diminue la valeur de réputation de sa victime et, s'il s'agit d'un agent manipulateur, cela peut l'inciter à se blanchir prématurément. De même, il serait possible de piéger un agent manipulateur en partageant des informations et en vérifiant à l'aide d'un agent Sybil si l'agent soupçonné de malveillance diffuse correctement cette information. Comme pour les jeux de coalitions, quelle serait alors l'influence de ces *manipulations honnêtes* sur l'efficacité du système ?

Annexe A

Notations

Dans cette annexes, nous rappelons les notations utilisées dans ce manuscrit. L'annexe A.1 présente les notations utilisées dans les chapitres 3, 4 et 5. L'annexe A.2 présente les notations utilisées dans les chapitres 6, 7 et 8.

A.1 Notations dans les jeux hédoniques

Notation d'un jeu hédonique

$HG = \langle N, \succeq, \mathbb{P} \rangle$	Un jeu hédonique ;
$N = \{a_1, \dots a_n\}$	Ensemble des agents du jeu ;
$\succeq = \{\succeq_1, \dots \succeq_n\}$	Ensemble des profils de préférence ;
$\mathbb{P}$	Protocole de sélection ;
$\mathcal{P}_N$	Ensemble des structures de coalitions possibles de N ;
$C_{a_i}^{\Pi}$	Coalition de l'agent a_i dans la structure $\Pi \in \mathcal{P}_N$;
$\mathcal{C}_{a_i}^N \subseteq 2^N$	Sous-ensemble des coalitions auxquelles l'agent a_i appartient.

Concepts de solution

$S_{HG} \subseteq \mathcal{P}_N$	Structures de coalitions satisfaisant le concept de soluions S ;
$NS_{HG} \subseteq \mathcal{P}_N$	Structures de coalitions stables au sens de Nash ;
$IS_{HG} \subseteq \mathcal{P}_N$	Structures de coalitions individuellement stables ;
$ICS_{HG} \subseteq \mathcal{P}_N$	Structures de coalitions individuellement contractuellement stables ;
$CS_{HG} \subseteq \mathcal{P}_N$	Structures de coalitions stables au sens du cœur ;
$PO_{HG} \subseteq \mathcal{P}_N$	Structures de coalitions optimales au sens de Pareto.

Solution du jeu

$\mathbb{P}(HG) \in \mathcal{P}_N$	Solution du jeu hédonique HG définie par le protocole $\mathbb{P}$;
$AP_{a_i}^x(HG) \subseteq \mathcal{P}_N$	Structures de coalitions acceptables par a_i à une profondeur x ;
$\mathrm{P}^*(a_i, x \mid HG)$	Probabilité que $\mathbb{P}(HG)$ appartiennent à $AP_{a_i}^x(HG)$.

Manipulations

$\{a_m,s_1,\ldots,s_x\}$	Agent manipulateur et ses x agents Sybil ;
HG^M	Jeu résultant de l'application de la manipulation M sur HG ;
$\succeq_{a_i}^M$	Profil de préférence de l'agent a_i dans le jeu HG^M ;
M_C,M_D,M_K	Manipulations constructive, destructive et destruction du cœur ;
$UR_{a_m}^{HG} \subseteq \mathcal{P}_N$	Structures de coalitions dont a_m est l'unique responsable de la non stabilité ;
$f(\Pi,s,C_0) \in \mathcal{P}_{N\cup\{s\}}$	Structure de coalitions construite par l'ajout de l'agent s dans la coalition $C_0 \in \Pi \cup \{\emptyset\}$;
$f^{-1}(\Pi') \in \mathcal{P}_N$	Structure de coalitions ayant servie à construire $\Pi' \in \mathcal{P}_{N\cup\{s\}}$;
$card_{M_C}(\Pi\|HG)$	Nombre de structures de coalitions stables dans le jeu HG^{M_C} construites à partir de la structure de coalitions $\Pi \in \mathcal{P}_N$.

A.2 Notations dans les systèmes de réputation

Échange de services

$N = \{a_1,\ldots,a_n\}$	Ensemble des agents du système ;
$S = \{s_1,\ldots,s_m\}$	Ensemble des services fournis par les agents ;
$N_x \subseteq N$	Ensembles des agents fournissant le service s_x ;
$S_k \subseteq S$	Ensemble des services fournis par l'agent a_k ;
$\varepsilon_{k,x}$	Expertise de l'agent $a_k \in N_x$ pour le service $s_x \in S_k$.

Évaluation de l'agent

v_i	Fonction d'évaluation de l'agent $a_i \in N$;
$v_{i,k,x}^t$	Évaluation de a_i de la qualité du service s_x fourni par l'agent a_k à l'instant t.
g_i^t	Gains totaux observés par l'agent $a_i \in N$ à l'instant t ;
$c_{i,k,x}$	Confiance de l'agent a_iN envers l'agent $a_k \in N$ pour fournir le service $s_x \in S$;
f_i	Fonction de réputation l'agent $a_i \in N$.

Observations et témoignages

$O_{i,k,x}$	Ensemble des observations de l'agent $a_i \in N$ pour les services $s_x \in S$ fournis par l'agent $a_k \in N_x$;
$F_{i,j,k,x}$	Ensemble des témoignages que l'agent $a_j \in N$ a fournis à l'agent $a_i \in N$ vis-à-vis des services $s_x \in S$ fournis par l'agent $a_k \in N_x$;
$\mathcal{F}_i$	Ensemble des témoignages et des observations de l'agent $a_i \in N$.

Bandits manchots

$B = \{b_1,\ldots,b_k\}$	Ensemble des bras d'un bandits manchots ;
θ_i	Distribution de probabilité de la récompense associée aux bras b_i ;
$b_i^t \in M$	Bras sélectionné à l'instant t ;
π_i	Politique de sélection de l'agent a_i ;
r_i^t	Regret de l'agent a_i à l'instant t.

Crédibilité et filtrage

$\mu_{i,k,x}$	Moyenne des valeurs de $O_{i,k,x}$;
$\sigma_{i,k,x}$	Écart-type des valeurs de $O_{i,k,x}$;
$\mathcal{N}(\mu_{i,k,x},\sigma^2_{i,k,x})$	Approximation par la loi normale de $\varepsilon_{k,x}$ à partir de $O_{i,k,x}$;
$\mu_{i,j,k,x}$	Moyenne des valeurs de $F_{i,j,k,x}$;
$\sigma_{i,j,k,x}$	Écart-type des valeurs de $F_{i,j,k,x}$;
$\mathcal{N}(\mu_{i,j,k,x},\sigma^2_{i,j,k,x})$	Approximation par la loi normale de $\varepsilon_{k,x}$ à partir de $F_{i,j,k,x}$;
$D_{i,j,k,x}$	Divergence de Kullback-Leibler entre $\mathcal{N}(\mu_{i,k,x},\sigma^2_{i,k,x})$ et $\mathcal{N}(\mu_{i,j,k,x},\sigma^2_{i,j,k,x})$;
$KL_i(F_{i,j,k,x})$	Le témoignage $F_{i,j,k,x}$ est KL-crédible ;
$KL^k_i(N) \subseteq N$	Ensemble des agents k-crédibles ;
$L^k_i(F_{i,j,k',x})$	Le témoignage $F_{i,j,k',x}$ est crédible par k-stochocratie ;
ϕ_i	Fonction de filtrage de l'agent a_i.

Annexe B

Démonstrations des propriétés

Dans cet annexe, nous rappelons les différentes propriétés du manuscrit et présentons leurs démonstrations.

Propriété 4.1.5 *(Page 71)* :

Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique tel que $NS_{HG} \neq \emptyset$. Soit $a_m \in N$ un agent malhonnête ayant le profil de préférence $\succeq_{a_m} = C_{a_m,1} \succeq_{a_m} C_{a_m,2} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. La manipulation constructive mise en œuvre par l'agent a_m est k-rationnelle si :

$$\forall i \in [1,k[, \frac{|\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,i}\}|}{|NS_{HG}|} = \frac{(1)_i + (2)_i}{(3)_i}$$
$$\frac{|\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\}|}{|NS_{HG}|} < \frac{(1)_k + (2)_k}{(3)}$$

avec
- $(1)_k = card_{M_C}(\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\} | HG)$;
- $(2)_k = card_{M_C}(\{\Pi \in UR^{HG}_{a_m} | \exists C_0 \in \Pi \cup \{\emptyset\} : (2.1)_k \wedge (2.2)\} | HG)$;
- $(2.1)_k = C_0 \cup \{a_m\} \sim_{a_m} C_{a_m,k}$;
- $(2.2) = \forall C \in \Pi, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}$;
- $(3) = card_{M_C}(NS_{HG} | HG) + card_{M_C}(UR^{HG}_{a_m} | HG)$.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et $a_m \in N$ un agent malhonnête ayant le profil de préférence $\succeq_{a_m} = C_{a_m,1} \succeq_{a_m} C_{a_m,2} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Fixons $k \in [1,2^{n-1}[$ tel que $C_{a_m,k} \succ_{a_m} \{a_m\}$. Par définition de la k-rationalité (définition 3.2.5), la manipulation constructive est k-rationnelle si :

$$\begin{aligned} \forall i : 1 \leq i < k, \mathrm{P}^*(a_m,i|HG) &= \mathrm{P}^*(a_m,i|HG^{M_C}) \\ \mathrm{P}^*(a_m,k|HG) &< \mathrm{P}^*(a_m,k|HG^{M_C}) \end{aligned} \tag{B.1}$$

Montrons uniquement les conditions portant sur l'inégalité car celle portant sur l'égalité pour tout $i \in [1,k[$ suit le même raisonnement. Par définition des probabilités de satisfaction (définition 3.1.6), et de la propriété 3.1.2, nous avons :

$$\mathrm{P}^*(a_m,k|HG) < \mathrm{P}^*(a_m,k-1|HG) + \sum_{\Pi \in AP^k_{a_m}(HG) \setminus AP^{k-1}_{a_m}(HG)} \mathrm{P}(\Pi|HG)$$
$$\text{et } \mathrm{P}^*(a_m,k|HG^{M_C}) < \mathrm{P}^*(a_m,k-1|HG^{M_C}) + \sum_{\Pi \in AP^k_{a_m}(HG^{M_C}) \setminus AP^{k-1}_{a_m}(HG^{M_C})} \mathrm{P}(\Pi|HG^{M_C})$$

Pour $k = 1$, nous avons $\mathrm{P}^*(a_m,k-1|HG) = \mathrm{P}^*(a_m,k-1|HG^{M_C}) = 0$ et, pour tout $k > 1$, si $\mathrm{P}^*(a_m,k|HG) \neq \mathrm{P}^*(a_m,k|HG^{M_C})$ alors la première condition à la k-rationalité n'est pas satisfaite est la manipulation n'est pas k-rationnelle. L'inégalité B.1 peut ainsi être récrite par :

$$\sum_{\Pi \in AP^k_{a_m}(HG) \setminus AP^{k-1}_{a_m}(HG)} \mathrm{P}(\Pi|HG) < \sum_{\Pi \in AP^k_{a_m}(HG^{M_C}) \setminus AP^{k-1}_{a_m}(HG^{M_C})} \mathrm{P}(\Pi|HG^{M_C}) \tag{B.2}$$

Comme nous faisons l'hypothèse que le protocole de sélection définit la solution aléatoirement uniformément parmi l'ensemble des structures de coalitions stables au sens de Nash, nous avons les deux égalités suivantes :

$$\sum_{\Pi \in AP^k_{a_m}(HG) \setminus AP^{k-1}_{a_m}(HG)} \mathrm{P}(\Pi|HG) = \frac{|NS_{HG} \cap (AP^k_{a_m}(HG) \setminus AP^{k-1}_{a_m}(HG))|}{|NS_{HG}|} \tag{B.3}$$

$$\sum_{\Pi \in AP^k_{a_m}(HG^{M_C}) \setminus AP^{k-1}_{a_m}(HG^{M_C})} \mathrm{P}(\Pi|HG^{M_C}) = \frac{|NS_{HG^{M_C}} \cap AP^k_{a_m}(HG^{M_C}) \setminus AP^{k-1}_{a_m}(HG^{M_C})|}{|NS_{HG^{M_C}}|} \tag{B.4}$$

Par définition des concepts d'acceptation (définition 3.1.3) et la propriété 3.1.1, nous avons :

$$NS_{HG} \cap (AP^k_{a_m}(HG) \setminus AP^{k-1}_{a_m}(HG)) = \{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\}$$

L'égalité B.3 peut ainsi être simplifiée par :

$$\frac{|NS_{HG} \cap (AP^k_{a_m}(HG) \setminus AP^{k-1}_{a_m}(HG))|}{|NS_{HG}|} = \frac{|\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\}|}{|NS_{HG}|} \tag{B.5}$$

De même, par définition du concept d'acceptation de l'agent malhonnête, nous avons :

$$\begin{aligned}
& NS_{HG^{M_C}} \cap AP^k_{a_m}(HG^{M_C}) \setminus AP^{k-1}_{a_m}(HG^{M_C}) \\
= & \{ \Pi' \in NS_{HG^{M_C}} | C^{\Pi'}_{a_m} \sim_{a_m} C_{a_m,k} \succeq_{a_m} (C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\} \\
& \qquad \vee (C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\} \sim_{a_m} C_{a_m,k} \succ_{a_m} C^{\Pi'}_{a_m} \} \\
= & \{\Pi' \in NS_{HG^{M_C}} | C^{\Pi'}_{a_m} \sim_{a_m} C_{a_m,k} \succeq_{a_m} (C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\}\} \\
& \cup \{\Pi' \in NS_{HG^{M_C}} | (C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\} \sim_{a_m} C_{a_m,k} \succ_{a_m} C^{\Pi'}_{a_m}\}
\end{aligned}$$

Cet ensemble correspond aux structures de coalitions stables au sens de Nash de HG^{M_C} telles que l'agent manipulateur forme la coalition $C_{a_m,k}$ et que l'agent Sybil ne forme pas une coalition préférée à $C_{a_m,k}$ (ou inversement). Il s'agit ici d'une union d'ensemble disjoint et, pour calculer sa cardinalité, nous pouvons donc calculer la cardinalité des ensembles suivants :

$$\begin{aligned}
& \{\Pi' \in NS_{HG^{M_C}} | C^{\Pi'}_{a_m} \sim_{a_m} C_{a_m,k} \succeq_{a_m} (C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\}\} \\
& \{\Pi' \in NS_{HG^{M_C}} | (C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\} \sim_{a_m} C_{a_m,k} \succ_{a_m} C^{\Pi'}_{a_m}\}
\end{aligned}$$

Considérons une structure de coalitions $\Pi' \in NS_{HG^{M_C}} \cap AP^k_{a_m}(HG^{M_C}) \setminus AP^{k-1}_{a_m}(HG^{M_C})$ et la coalition $\Pi = f(\Pi')^{-1}$. Par construction de partition, nous avons $C^{\Pi}_{a_m} = C^{\Pi'}_{a_m}$. Notons par $C_0 \in \Pi \cup \{\emptyset\}$ telle que coalition $C^{\Pi'}_s = C_0 \cup \{s\}$. Rappelons que par la propriété 4.1.1, nous avons nécessairement $C_0 \cup \{a_m\} \neq C^{\Pi}_{a_m}$.

Regardons dans un premier temps le cas où $\Pi' \in \{\Pi' \in NS_{HG^{M_C}} | C^{\Pi'}_{a_m} \sim_{a_m} C_{a_m,k} \succeq_{a_m} C_0 \cup \{a_m\}\}$. Comme $C^{\Pi'}_{a_m} \succeq_{a_m} C_0 \cup \{a_m\}$, par propriété 4.1.1, nous avons $\Pi \in NS_{HG}$. Ainsi, toute structure de coalitions $\Pi' \in \{\Pi' \in NS_{HG^{M_C}} | C^{\Pi'}_{a_m} \sim_{a_m} C_{a_m,k} \succeq_{a_m} C_0 \cup \{a_m\}\}$ est construite à partir d'une structure de coalitions Π telle que $\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\}$. Par conséquent, nous avons l'égalité :

$$\begin{aligned} &|\{\Pi' \in NS_{HG^{M_C}} | C^{\Pi'}_{a_m} \sim_{a_m} C_{a_m,k} \succeq_{a_m} (C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\}\}| \\ &= card_{M_C}(\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\} | HG) \end{aligned} \tag{B.6}$$

Supposons maintenant que $\Pi' \in \{\Pi' \in NS_{HG^{M_C}} | C_0 \cup \{a_m\} \sim_{a_m} C_{a_m,k} \succ_{a_m} C^{\Pi'}_{a_m}\}$. Comme $C_0 \cup \{a_m\} \succ_{a_m} C^{\Pi'}_{a_m}$, par la propriété 4.1.3, nous avons $\Pi \in UR^{HG}_{a_m}$. Comme $\Pi' \in NS_{HGM_C}$, nous avons même $\forall C \in \Pi, C_0 \cup \{a_m\} \sim_{a_m} C_{a_m,k} \succeq_{a_m} C \cup \{a_m\}$. Ainsi, toute structure de coalitions $\Pi' \in \{\Pi' \in NS_{HG^{M_C}} | C_0 \cup \{a_m\} \sim_{a_m} C_{a_m,k} \succ_{a_m} C^{\Pi'}_{a_m}\}$ est construite à partir d'une structure de coalitions Π telle que $\Pi \in \{\Pi \in UR^{HG}_{a_m} | (1) \wedge (2)\}$ où :

- $(1) = \exists C_0 \in \Pi \cup \{\emptyset\} : C_0 \cup \{a_m\} \sim_{a_m} C_{a_m,k}$;
- $(2) = \forall C \in \Pi, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}$.

Par conséquent, nous avons l'égalité :

$$\begin{aligned} &|\{\Pi' \in NS_{HG^{M_C}} | C_0 \cup \{a_m\} \sim_{a_m} C_{a_m,k} \succ_{a_m} C^{\Pi'}_{a_m}\}| \\ &= card_{M_C}(\{\Pi \in UR^{HG}_{a_m} | (1) \wedge (2)\} | HG) \end{aligned} \tag{B.7}$$

La propriété 4.1.4, nous donne l'égalité :

$$|NS_{HG^{M_C}}| = card_{M_C}(NSHG|HG) + card_{M_C}(UR^{HG}_{a_m}|HG) \tag{B.8}$$

$$\tag{B.9}$$

À partir des égalités B.6 et B.7 et B.8 nous pouvons réécrire l'égalité B.4 par :

$$\begin{aligned} &\frac{|NS_{HG^{M_C}} \cap AP^k_{a_m}(HG^{M_C}) \setminus AP^{k-1}_{a_m}(HG^{M_C})|}{|NS_{HG^{M_C}}|} \\ &= \frac{card_{M_C}(\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\} | HG)}{card_{M_C}(NS_{HG}|HG) + card_{M_C}(UR^{HG}_{a_m}|HG)} \\ &+ \frac{card_{M_C}(\{\Pi \in UR^{HG}_{a_m} | (1) \wedge (2)\} | HG)}{card_{M_C}(NS_{HG}|HG) + card_{M_C}(UR^{HG}_{a_m}|HG)} \end{aligned} \tag{B.10}$$

Les égalités B.5 et B.10 nous permette ainsi de dire que l'inégalité B.1 est vraie si et seulement si :

$$\begin{aligned} \frac{|\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\}|}{|NS_{HG}|} < &\frac{card_{M_C}(\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\} | HG)}{card_{M_C}(NS_{HG}|HG) + card_{M_C}(UR^{HG}_{a_m}|HG)} \\ &+ \frac{card_{M_C}(\{\Pi \in UR^{HG}_{a_m} | (1) \wedge (2)\} | HG)}{card_{M_C}(NS_{HG}|HG) + card_{M_C}(UR^{HG}_{a_m}|HG)} \end{aligned} \tag{B.11}$$

avec :

- $(1) = \exists C_0 \in \Pi \cup \{\emptyset\} : C_0 \cup \{a_m\} \sim_{a_m} C_{a_m,k}$;
- $(2) = \forall C \in \Pi, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}$.

□

Propriété 4.1.7 *(Page 73)* :

Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succ_{a_m} \ldots \succ_{a_m} C_{a_m,2^{n-1}}$. Soit $(1)_k$ la condition $\forall i \in [1,k[, C_{a_m,i} \setminus \{a_m\} \notin \Pi$. La manipulation constructive est k-rationnelle si et seulement si :

$$\forall i \in [1,k[, \frac{|\{\Pi \in NS_{HG} | C_{a_m,i} \in \Pi\}|}{|NS_{HG}|} = \frac{|\{\Pi \in UR^{HG}_{a_m} | Ca_m,i \setminus \{a_m\} \in \Pi \wedge (1)_i\}|}{|UR^{HG}_{a_m}|}$$
$$\frac{|\{\Pi \in NS_{HG} | C_{a_m,k} \in \Pi\}|}{|NS_{HG}|} < \frac{|\{\Pi \in UR^{HG}_{a_m} | Ca_m,k \setminus \{a_m\} \in \Pi \wedge (1)_k\}|}{|UR^{HG}_{a_m}|}$$

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succ_{a_m} \ldots \succ_{a_m} C_{a_m,2^{n-1}}$.

Montrons uniquement les conditions portant sur l'inégalité car la démonstration portant sur l'égalité pour tout $i \in [1,k[$ suit le même raisonnement. Par la propriété 4.1.5, la manipulation constructive est k-rationnelle si :

$$\frac{|\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\}|}{|NS_{HG}|} < \frac{(1)_k + (2)_k}{(3)} \tag{B.12}$$

avec :
- $(1)_k = card_{M_C}(\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\} | HG)$;
- $(2)_k = card_{M_C}(\{\Pi \in UR^{HG}_{a_m} | \exists C_0 \in \Pi \cup \{\emptyset\} : (2.1)_k \wedge (2.2)\} | HG)$;
- $(2.1)_k = C_0 \sim_{a_m} C_{a_m,k}$;
- $(2.2) = \forall C \in \Pi, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}$;
- $(3) = card_{M_C}(NS_{HG} | HG) + card_{M_C}(UR^{HG}_{a_m} | HG)$.

Par le corollaire 4.1.1, comme le profil de préférence de a_m est un ordre strict, nous avons :

$$\begin{aligned}(1)_k &= card_{M_C}(\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\} | HG) \\ &= |\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\}|\end{aligned}$$

De même, par le corollaire 4.1.3 :

$$\begin{aligned}(2)_k &= card_{M_C}(\{\Pi \in UR^{HG}_{a_m} | (2.1)_k \wedge (2.2)\} | HG) \\ &= |\{\Pi \in UR^{HG}_{a_m} | (2.1)_k \wedge (2.2)\}|\end{aligned}$$

Nous avons par ailleurs par la propriété 4.1.4 :

$$\begin{aligned}(3) &= card_{M_C}(NS_{HG} | HG) + card_{M_C}(UR^{HG}_{a_m} | HG) \\ &= |NS_{HG}| + |UR^{HG}_{a_m}|\end{aligned}$$

L'inégalité B.12 peut ainsi être simplifiée par :

$$\frac{(1)_k}{|NS_{HG}|} < \frac{(1)_k + (2)_k}{|NS_{HG}| + |UR^{HG}_{a_m}|} \tag{B.13}$$

Or, nous avons :

$$\frac{(1)_k}{|NS_{HG}|} < \frac{(1)_k + (2)_k}{|NS_{HG}| + |UR^{HG}_{a_m}|}$$

SSI $(1)_k \times (|NS_{HG}| + |UR^{HG}_{a_m}|) < |NS_{HG}| \times ((1)_k + (2)_k)$

SSI $(1)_k \times |NS_{HG}| + (1)_k \times |UR^{HG}_{a_m}| < (1)_k \times |NS_{HG}| + (2)_k \times |NS_{HG}|$

SSI $(1)_k \times |UR^{HG}_{a_m}| < (2)_k \times |NS_{HG}|$

SSI $\frac{(1)_k}{|NS_{HG}|} < \frac{(2)_k}{|UR^{HG}_{a_m}|}$

SSI $\frac{|\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\}|}{|NS_{HG}|} < \frac{|\{\Pi \in UR^{HG}_{a_m} | (2.1)_k \wedge (2.2)\}|}{|UR^{HG}_{a_m}|}$

Comme le profil de préférence de l'agent a_m est strict, nous avons :

$$\begin{aligned}\{\Pi \in UR^{HG}_{a_m} | (2.1)_k \wedge (2.2)\} = \{&\Pi \in UR^{HG}_{a_m} | Ca_m,k \setminus \{a_m\} \in \Pi \\ &\wedge \forall i \in [1,k[, C_{a_m,i} \setminus \{a_m\} \notin \Pi\} \\ \{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\} = \{&\Pi \in NS_{HG} | C_{a_m,k} \in \Pi\}\end{aligned}$$

Ainsi, l'inégalité B.13 peut être simplifiée par :

$$\frac{|\{\Pi \in NS_{HG} | C_{a_m,k} \in \Pi\}|}{|NS_{HG}|} < \frac{|\{\Pi \in UR^{HG}_{a_m} | Ca_m,k \setminus \{a_m\} \in \Pi \wedge (1)_k\}|}{|UR^{HG}_{a_m}|}$$

où $(1)_k$ désigne la condition $\forall i \in [1,k[, C_{a_m,i} \setminus \{a_m\} \notin \Pi$ □

Propriété 4.3.1 *(Page 80)* :

Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique et $a_m \in N$ un agent malhonnête. Une manipulation quelconque M effectuée par a_m est k-rationnelle si et seulement si soit la manipulation constructive M_C, soit la manipulation destructive M_D est k-rationnelle.

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ et un agent malhonnête $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Considérons la manipulation $M = \langle \{a_m, s_1, \ldots, s_x\}, \{\succeq^M_{a_m}, \succeq^M_{s_1}, \ldots, \succeq^M_{s_x}\}, \succeq_{a_m} \rangle$. Supposons que M soit k-rationnelle et qu'il n'existe pas de $k' \in [1,2^{n-1}]$ tel que la manipulation constructive (définition 4.1.1) et la manipulation destructive (définition 4.2.1) est k'-rationnelle.

Par l'hypothèse que la manipulation M est k-rationnelles (définition 3.2.5), nous avons :

$$\begin{gathered}\forall i \in [1,k[, \mathrm{P}^*(a_m,i|HG) = \mathrm{P}^*(a_m,i|HG^M) \\ \mathrm{P}^*(a_m,k|HG) < 1\end{gathered}$$

Comme nous supposons que la manipulation destructive n'est pas rationnelle, par la propriété 4.2.6, nous avons :

$$\forall i \in [1,k[, \nexists \Pi \in NS_{HG} : C_{a_m,i} \in \Pi$$

En effet, s'il existe un $i \in [1,k[$ tel que $\Pi \in NS_{HG} : C_{a_m,i} \in \Pi$ alors la manipulation destructive serait i-rationnelle. Par conséquent, nous avons :

$$\forall i \in [1,k[, \mathrm{P}^*(a_m,i|HG) = 0$$

Supposons que $0 < \mathrm{P}^*(a_m,k|HG)$. Comme $\mathrm{P}^*(a_m,k|HG) < 1$, par définition de la probabilité de sélection, il existe $i \in [k,2^{n-1}]$ tel que $Ca_m,k \succ_{a_m} Ca_m,i$ et que $\exists \Pi \in NS_{HG} : C^{\Pi}_{a_m} \sim_{a_m} Ca_m,i$. Or par la propriété 4.2.6, la manipulation destructive est alors k-rationnelle. Par l'hypothèse de non-rationalité de cette manipulation, nous avons donc $\mathrm{P}^*(a_m,k|HG) = 0$. Comme nous faisons l'hypothèse que la manipulation M est k-rationnelle par définition, il existe une structure de coalitions $\Pi' \in NS_{HG^M}$ telle que :

$$\exists! a_i \in \{a_m,s_1,\ldots,s_x\} : (C_i^{\Pi'} \setminus \{a_i\}) \cup \{a_m\} \sim_{a_m} C_{a_m,k}$$

Fixons cette structure de coalitions Π' et la structure de coalitions $\Pi \in \mathcal{P}_N$ telle que $\forall C \in \Pi'$, nous avons $C \setminus \{s_1,\ldots,s_x\} \in \Pi \cup \{\emptyset\}$. Par construction de Π, nous avons $C^{\Pi'}_{a_m} = C^{\Pi'}_{a_m}$ et l'une des conditions suivantes est satisfaite :

1. $C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}$;
2. $\exists C \in \Pi : C \cup \{a_m\} \succeq_{a_m} C_{a_m,k}$.

Par les hypothèses d'indépendance des alternatives non-pertinentes et de bénéfice du doute (hypothèse 3.2.3 et 3.2.4), comme $\Pi' \in NS_{HG^M}$, nous avons :

$$\forall a_i \in N \setminus \{a_m\}, \nexists C \in \Pi : C \cup \{a_i\} \succ_{a_i} C_i^{\Pi}$$

Par conséquent, soit $\Pi \in NS_{HG}$, soit $\Pi \in UR^{HG}_{a_m}$. Supposons maintenant que $\Pi \in NS_{HG}$. Comme $\mathrm{P}^*(a_m,k|HG) = 0$, la condition (1) ne peut pas être satisfaite. De même, par définition de la stabilité, la condition (2) ne peut également pas être satisfaite. Par conséquent, aucune des deux conditions liées à Π n'est satisfaite si $\Pi \in NS_{HG}$. Ainsi, nous avons nécessairement $\Pi \in UR^{HG}_{a_m}$. Ainsi, nous avons une structure de coalitions $\Pi \in UR^{HG}_{a_m}$ telle que la condition (2) est satisfaite. Rappelons que pour tout $i \in [1,k[$, nous avons $\mathrm{P}^*(a_m,i|HG) = 0$.

Si $NS_{HG} = \emptyset$, par la propriété 4.1.6, soit $\exists i' \in [1,i[$ tel que la manipulation constructive est i'-rationnelle, soit la manipulation constructive est i-rationnelle. Dans les deux cas, cela contredit l'hypothèse de non-rationalité de la manipulation constructive. Par conséquent, nous avons $NS_{HG} \neq \emptyset$. Comme $\mathrm{P}^*(a_m,k|HG) = 0$ et que $NS_{HG} \neq \emptyset$, pour tout $i \in [1,k]$, nous avons l'égalité :

$$\frac{|\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,i}\}|}{|NS_{HG}|} = 0.$$

Par la propriété 4.1.5, pour que la manipulation constructive ne soit pas k-rationnelle, tout $i \in [1,k]$, nous devons donc avoir :

$$\frac{(1)_i + (2)_i}{(3)_i} = 0$$

avec pour rappel :

- $(1)_i = card_{M_C}(\{\Pi \in NS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,i}\}|HG)$;
- $(2)_i = card_{M_C}(\{\Pi \in UR^{HG}_{a_m} | \exists C_0 \in \Pi \cup \{\emptyset\} : (2.1)_i \wedge (2.2)\}|HG)$;
- $(2.1)_i = C_0 \cup \{a_m\} \sim_{a_m} C_{a_m,i}$;
- $(2.2) = \forall C \in \Pi, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}$;

– $(3) = card_{M_C}(NS_{HG}|HG) + card_{M_C}(UR^{HG}_{a_m}|HG)$.

Regardons enfin le cas de Π. Nous avons déjà montré que $\Pi \in UR^{HG}_{a_m}$ et que $\exists C \in \Pi : C \cup \{a_m\} \succeq_{a_m} C_{a_m,k}$. Par définition de l'agent unique responsable de la non-stabilité, il existe une coalition $C_0 \in \Pi \cup \{\emptyset\}$ telle que $\forall C \in \Pi, C_0 \cup \{a_m\} \succeq_{a_m} C \cup \{a_m\}$. Ainsi, $\exists i \in [1,k]$ tel que $\Pi \in \{\Pi \in UR^{HG}_{a_m} | \exists C_0 \in \Pi \cup \{\emptyset\} : (2.1)_i \wedge (2.2)\}$. Par conséquent, nous avons $(2)_i > 0$, ce qui est en contradiction avec la non-rationalité de la manipulation constructive (propriété 4.1.5) .

Ainsi, nous avons montré que :

– si $\mathrm{P}^*(a_m,k|HG) = 1$ alors la manipulation M n'est pas k rationnelle ;

– si $\mathrm{P}^*(a_m,k|HG) > 0$ alors la manipulation destructive est rationnelle ;

– si $\mathrm{P}^*(a_m,k|HG) = 0$ alors la manipulation constructive est rationnelle ;

Ainsi, pour que la manipulation M soit k-rationnelle, il faut nécessairement que soit la manipulation constructive, soit la manipulation destructive soit rationnelle. □

Propriété 5.2.8 *(Page 100)* :

Soit $HG = \langle N, \succeq, \mathbb{P} \rangle$ un jeu hédonique telle que $\mathbb{P}$ retourne la solution du jeu HG aléatoirement uniformément parmi CS_{HG} l'ensemble des structures de coalitions stables au sens du cœur dans HG et $a_m \in N$ un agent malhonnête ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} C_{a_m,2} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$. Soit $(1)_i$ le nombre de structures de coalitions $\Pi \in CS_{HG}$ telles que :

$$C_{a_m,i} \sim_{a_m} C^{\Pi}_{a_m} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$$

Soit $(2)_i$ le nombre de structures de coalitions $\Pi \notin CS_{HG}$ telles que :

$$C_{a_m,i} \sim_{a_m} C^{\Pi}_{a_m} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$$
$$\text{et } \forall N_2 \subseteq N : \forall a_i \in N_2, N_2 \succ_{a_i} C^{\Pi}_{a_i}, a_m \in N_2$$

Si $CS_{HG} = \emptyset$, la manipulation constructive mise en œuvre par a_m est k-rationnelle si et seulement si $\forall i \in [1,k[$, $(2)_i = 0$ et $(2)_k > 0$. Si $CS_{HG} \neq \emptyset$, la manipulation constructive mise en œuvre par a_m est k-rationnelle si et seulement si :

$$\forall i \in [1,k[, \frac{|\{\Pi \in CS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,i}\}|}{|CS_{HG}|} = \frac{(1)_i + (2)_i}{card_{M_C}(\mathcal{P}_{HG}|HG)}$$
$$\text{et } \frac{|\{\Pi \in CS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\}|}{|CS_{HG}|} < \frac{(1)_k + (2)_k}{card_{M_C}(\mathcal{P}_{HG}|HG)}$$

Démonstration : Fixons un jeu hédonique $HG = \langle N, \succeq, \mathbb{P} \rangle$ telle que $\mathbb{P}$ retourne la solution du jeu HG aléatoirement uniformément parmi CS_{HG} et un agent malhonnête $a_m \in N$ ayant le profil de préférence $C_{a_m,1} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,k} \succeq_{a_m} \ldots \succeq_{a_m} C_{a_m,2^{n-1}}$ Fixons HG^{M_C} le jeu résultant de la mise en œvre de la manipulation constructive par a_m sur HG. Rappelons que par la définition 3.2.5, la manipulation M_C est k-rationnelle si $\forall i \in [1,k[$ nous avons $\mathrm{P}^*(a_m,i|HG) = \mathrm{P}^*(a_m,i|HG^M)$ et que $\mathrm{P}^*(a_m,k|HG) < \mathrm{P}^*(a_m,k|HG^M)$.

Montrons dans un premier temps les conditions nécessaires lorsque $CS_{HG} = \emptyset$. Comme $CS_{HG} = \emptyset$, pour tout k tel que $C_{a_m,k} \succ_{a_m} \{a_m\}$, nous avons $\mathrm{P}^*(a_m,k|HG) = 0$. Par conséquent, la manipulation M_C est k-rationnelle si :

– pour tout $i \in [1,k[$, il n'existe pas de structure de coalitions $\Pi' \in CS_{HG^{M_C}}$ telle que $C^{\Pi'}_{a_m} \sim_{a_m} C_{a_m,i}$ ou que $(C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\} \sim_{a_m} C_{a_m,i}$;

– il existe une structure de coalitions $\Pi' \in CS_{HG^{M_C}}$ telle que $C^{\Pi'}_{a_m} \sim_{a_m} C_{a_m,k}$ ou que $(C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\} \sim_{a_m} C_{a_m,k}$.

Fixons une structure de coalitions $\Pi' \in CS_{HG^{M_C}}$. Soit $\Pi \in \mathcal{P}_N$ la structure de coalitions telle que $\Pi = f^{-1}(\Pi')$ et $C_0 = C^{\Pi'}_s$. Par la propriété 5.2.7, comme Π' est stable au sens du cœur, nous avons : $C_0 \cup \{a_m\} \succeq \{a_m\}$, $C_0 \neq C^{\Pi}_{a_m}$ et $\forall N_2 \subseteq N : \forall a_i \in N_2, N_2 \succ_{a_i} C^{\Pi}_i$, $a_m \in N_2$. Par conséquent, il existe une de coalition $\Pi' \in CS_{HG^{M_C}}$ tel que $C^{\Pi'}_{a_m} \sim_{a_m} C_{a_m,i}$ ou que $(C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\} \sim_{a_m} C_{a_m,i}$ s'il existe une structure de coalitions $\Pi \notin CS_{HG}$ telle que

$$C_{a_m,i} \sim_{a_m} C^{\Pi}_{a_m} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$$
$$\text{et } \forall N_2 \subseteq N : \forall a_i \in N_2, N_2 \succ_{a_i} C^{\Pi}_{a_i}, a_m \in N_2$$

Notons par $(2)_i$ le nombre de telles structures de coalitions. Les conditions nécessaire à la k-rationnalité de la manipulation constructive lorsque $CS_{HG} = \emptyset$ peuvent être réécrite par : $\forall i \in [1,k[$, $(2)_i = 0$ et $(2)_k > 0$.

Montrons maintenant les conditions nécessaire à la k-rationnalité de la manipulation constructive lorsque $CS_{HG} \neq \emptyset$. Par définition, la manipulation constructive est k-rationnelle si :

$$\forall i \in [1,k[, \frac{|\{\Pi \in CS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,i}\}|}{|CS_{HG}|} = \frac{|\{\Pi \in CS_{HG^{M_C}} | Cond_1(i) \vee Cond_2(i)\}|}{card_{M_C}(\mathcal{P}_{HG}|HG)}$$
$$\text{et } \frac{|\{\Pi \in CS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\}|}{|CS_{HG}|} < \frac{|\{\Pi \in CS_{HG^{M_C}} | Cond_1(k) \vee Cond_2(k)\}|}{|CS_{HG^{M_C}}|}$$

où $Cond_1(i)$ et $Cond_2(i)$ désignent respectivement les conditions $C^{\Pi'}_{a_m} \sim_{a_m} C_{a_m,i}$ et $(C^{\Pi'}_s \setminus \{s\}) \cup \{a_m\} \sim_{a_m} C_{a_m,i}$. Fixons une structure de coalitions $\Pi' \in CS_{HG^{M_C}}$. Soit $\Pi \in \mathcal{P}_N$ la structure de coalitions telle que $\Pi = f^{-1}(\Pi')$ et $C_0 = C^{\Pi'}_s$. Par la propriété 5.2.6, si $\Pi \in CS_{HG}$ alors $Cond_1(i) \vee Cond_2(i)$ est satisfait si :

$$C_{a_m,i} \sim_{a_m} C^{\Pi}_{a_m} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$$

Notons par $(1)_i$ le nombre de structure de coalitions $\Pi' \in CS_{HG^{M_C}}$ telles que $f^{-1}(\Pi') \in CS_{HG}$ satisfait $Cond_1(i) \vee Cond_2(i)$. Comme nous l'avons montré précédement par la propriété 5.2.7, $(2)_i$ disgne le nombre de structures de coalitions $\Pi' \in CS_{HG^{M_C}}$ telles que $f^{-1}(\Pi') \notin CS_{HG}$ satisfaisant $Cond_1(i) \vee Cond_2(i)$. Ainsi, nous avons $|\{\Pi \in CS_{HG^{M_C}} | Cond_1(i) \vee Cond_2(i)\}| = (1)_i + (2)_i$ avec :

$(1)_i$ le nombre de structures de coalitions $\Pi \in CS_{HG}$ telles que :

$$C_{a_m,i} \sim_{a_m} C^{\Pi}_{a_m} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$$

$(2)_i$ le nombre de structures de coalitions $\Pi \notin CS_{HG}$ telles que :

$$C_{a_m,i} \sim_{a_m} C^{\Pi}_{a_m} \vee \exists C \in \Pi : C \cup \{a_m\} \sim_{a_m} C_{a_m,i}$$
$$\text{et } \forall N_2 \subseteq N : \forall a_i \in N_2, N_2 \succ_{a_i} C^{\Pi}_{a_i}, a_m \in N_2$$

Rappelons que $|CS_{HG^{M_C}}|$ est définie par $card_{M_C}(\mathcal{P}_{HG}|HG)$. Par conséquent, si $CS_{HG} \neq \emptyset$, la manipulation constructive est k-rationnelle si et seulement si :

$$\forall i \in [1,k[, \frac{|\{\Pi \in CS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,i}\}|}{|CS_{HG}|} = \frac{(1)_i + (2)_i}{card_{M_C}(\mathcal{P}_{HG}|HG)}$$
$$\text{et } \frac{|\{\Pi \in CS_{HG} | C^{\Pi}_{a_m} \sim_{a_m} C_{a_m,k}\}|}{|CS_{HG}|} < \frac{(1)_k + (2)_k}{card_{M_C}(\mathcal{P}_{HG}|HG)}$$

□

Bibliographie

[Abdul-Rahman et Hailes, 2000] Abdul-Rahman, A. et Hailes, S. (2000). Supporting trust in virtual communities. *In Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, HICSS 2000*, page 10.

[Adar et Huberman, 2000] Adar, E. et Huberman, B. A. (2000). Free riding on gnutella. *First Monday*, 5(10):1–22.

[Agrawal, 1995] Agrawal, R. (1995). Sample mean based index policies with o (log n) regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078.

[Aknine *et al.*, 2004] Aknine, S., Pinson, S. et Shakun, M. F. (2004). A multi-agent coalition formation method based on preference models. *Group Decision and Negotiation*, 13(6):513–538.

[Alpcan et Başar, 2010] Alpcan, T. et Başar, T. (2010). *Network security : A decision and game-theoretic approach*. Cambridge University Press.

[Altman et Tennenholtz, 2005] Altman, A. et Tennenholtz, M. (2005). Ranking systems : the PageRank axioms. *In Proceedings of the 6th Conference on Economics and Computation*, pages 1–8.

[Altman et Tennenholtz, 2007a] Altman, A. et Tennenholtz, M. (2007a). An axiomatic approach to personalized ranking systems. *In Proceedings of the 20th International Joint Conference on Artificial Intelligence, IJCAI 2007*, pages 1187–1192.

[Altman et Tennenholtz, 2007b] Altman, A. et Tennenholtz, M. (2007b). Incentive compatible ranking systems. *In Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems, AAMAS 2007*, page 84.

[Anantharam *et al.*, 1987] Anantharam, V., Varaiya, P. et Walrand, J. (1987). Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays. *IEEE Automatic Control*, 32(11):968–976.

[Angin *et al.*, 2010] Angin, P., Bhargava, B., Ranchal, R., Singh, N., Linderman, M., Othmane, L. B. et Lilien, L. (2010). An entity-centric approach for privacy and identity management in cloud computing. *In Proceedings of the 29th Symposium on Reliable Distributed Systems*, pages 177–183.

[Arrow, 1963] Arrow, K. J. (1963). *Social Choice and Individual Values*. Numéro 12. Yale University Press.

[Artz et Gil, 2007] Artz, D. et Gil, Y. (2007). A survey of trust in computer science and the semantic web. *Web Semantics : Science, Services and Agents on the World Wide Web*, 5(2):58–71.

[Audibert *et al.*, 2009] Audibert, J.-Y., Munos, R. et Szepesvári, C. (2009). Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

[Auer *et al.*, 2002] Auer, P., Cesa-Bianchi, N. et Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

[Auer *et al.*, 1995] Auer, P., Cesa-Bianchi, N., Freund, Y. et Schapire, R. (1995). Gambling in a rigged casino : the adversarial multi-armed bandit problem. *In 36th Annual Symposium on Foundations of Computer Science*, pages 322–331.

[Auer et Ortner, 2010] Auer, P. et Ortner, R. (2010). UCB revisited : Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65.

[Aumann, 1985] Aumann, R. J. (1985). On the non-transferable utility value : A comment on the roth-shafer examples. *Econometrica : Journal of the Econometric Society*, 53:667–677.

[Aumann et Dreze, 1974] Aumann, R. J. et Dreze, J. H. (1974). Cooperative games with coalition structures. *International Journal of game theory*, 3(4):217–237.

[Aziz *et al.*, 2014] Aziz, H., Brandt, F. et Harrenstein, P. (2014). Fractional hedonic games. *In Proceedings of the 13th international conference on Autonomous agents and multi-agent systems, AAMAS 2014*, pages 5–12.

[Aziz *et al.*, 2011] Aziz, H., Brandt, F. et Seedig, H. G. (2011). Stable partitions in additively separable hedonic games. *In The 10th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2011*, pages 183–190.

[Aziz et Paterson, 2009] Aziz, H. et Paterson, M. (2009). False name manipulations in weighted voting games : splitting, merging and annexation. *In Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2009*, pages 409–416.

[Bachrach et Elkind, 2008] Bachrach, Y. et Elkind, E. (2008). Divide and conquer : False-name manipulations in weighted voting games. *In Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems, AAMAS 2008*, pages 975–982.

[Balabanović et Shoham, 1997] Balabanović, M. et Shoham, Y. (1997). Fab : content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72.

[Ballester, 2004] Ballester, C. (2004). Np-completeness in hedonic games. *Games and Economic Behavior*, 49(1):1–30.

[Barbera, 2001] Barbera, S. (2001). An introduction to strategy-proof social choice functions. *Social Choice and Welfare*, 18(4):619–653.

[Barsalou, 2009] Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 364(1521):1281–1289.

[Bartholdi III et Orlin, 1991] Bartholdi III, J. J. et Orlin, J. B. (1991). Single transferable vote resists strategic voting. *Social Choice and Welfare*, 8(4):341–354.

[Bartholdi III *et al.*, 1989] Bartholdi III, J. J., Tovey, C. A. et Trick, M. A. (1989). The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3):227–241.

[Basu *et al.*, 1998] Basu, C., Hirsh, H., Cohen, W. *et al.* (1998). Recommendation as classification : Using social and content-based information in recommendation. *In Proceedings of the 15th National Conference on Artificial intelligence, AAAI 1998*, pages 714–720.

[Bernstein *et al.*, 2000] Bernstein, D. S., Zilberstein, S. et Immerman, N. (2000). The complexity of decentralized control of markov decision processes. *In Proceedings of the 16th conference on Uncertainty in artificial intelligence*, pages 32–37.

[Bertsimas *et al.*, 2011] Bertsimas, D., Farias, V. F. et Trichakis, N. (2011). The price of fairness. *Operations research*, 59(1):17–31.

[Bilge *et al.*, 2009] Bilge, L., Strufe, T., Balzarotti, D. et Kirda, E. (2009). All your contacts are belong to us : automated identity theft attacks on social networks. *In Proceedings of the 18th international conference on World wide web*, pages 551–560.

[Binmore *et al.*, 1998] Binmore, K., Castelfranchi, C., Doran, J. et Wooldridge, M. (1998). Rationality in multi-agent systems. *The Knowledge Engineering Review*, 13(03):309–314.

[Bloch, 1997] Bloch, F. (1997). *New Directions in the Economic Theory of the Environment*, volume 25, chapitre 10 Non-cooperative models of coalition formation in games with spillovers, page 311. Cambridge University Press (Cambridge, UK).

[Bogomolnaia et Jackson, 2002] Bogomolnaia, A. et Jackson, M. O. (2002). The stability of hedonic coalition structures. *Games and Economic Behavior*, 38(2):201–230.

[Bonnet, 2012] Bonnet, G. (2012). A protocol based on a game-theoretic dilemma to prevent malicious coalitions in reputation systems. *In Proceedings of the 20th European Conference on Artificial Intelligence, ECAI 2012*, pages 187–192.

[Borisov, 2006] Borisov, N. (2006). Computational puzzles as sybil defenses. *In Proceedings of the 6th IEEE International Conference on Peer-to-Peer Computing*, pages 171–176.

[Bradshaw *et al.*, 2003] Bradshaw, J. M., Sierhuis, M., Acquisti, A., Feltovich, P., Hoffman, R., Jeffers, R., Prescott, D., Suri, N., Uszok, A. et Van Hoof, R. (2003). Adjustable autonomy and human-agent teamwork in practice : An interim report on space applications. *In Agent autonomy*, pages 243–280. Springer.

[Bramer *et al.*, 2007] Bramer, M., Bramer, M. et Bramer, M. (2007). *Principles of data mining*, volume 131. Springer.

[Brams et Fishburn, 2002] Brams, S. J. et Fishburn, P. C. (2002). Voting procedures. *Handbook of social choice and welfare*, 1:173–236.

[Breese *et al.*, 1998] Breese, J. S., Heckerman, D. et Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. *In Proceedings of the 14th conference on Uncertainty in artificial intelligence*, pages 43–52.

[Bubeck et Cesa-Bianchi, 2012] Bubeck, S. et Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122.

[Burnetas et Katehakis, 1997] Burnetas, A. N. et Katehakis, M. N. (1997). Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22(1):222–255.

[Carbo *et al.*, 2002] Carbo, J., Molina, J. M. et Davila, J. (2002). Comparing predictions of sporas vs. a fuzzy reputation system. *In Proceedings of the 3rd International Conference on Fuzzy Sets and Fuzzy Systems*, volume 200, pages 147–153.

[Chan *et al.*, 2003] Chan, H., Perrig, A. et Song, D. (2003). Random key predistribution schemes for sensor networks. *In Proceedings of the Symposium on Security and Privacy*, pages 197–213.

[Chang, 2002] Chang, R. K. (2002). Defending against flooding-based distributed denial-of-service attacks : a tutorial. *Communications Magazine, IEEE*, 40(10):42–51.

[Cheng et Friedman, 2005] Cheng, A. et Friedman, E. (2005). Sybilproof reputation mechanisms. *In Proceedings of the 3rd Workshop on Economics of Peer-to-Peer Systems*, pages 128–132.

[Cheng et Friedman, 2006] Cheng, A. et Friedman, E. (2006). Manipulability of pagerank under sybil strategies.

[Chevaleyre *et al.*, 2006] Chevaleyre, Y., Dunne, P. E., Endriss, U., Lang, J., Lemaitre, M., Maudet, N., Padget, J., Phelps, S., Rodriguez-Aguilar, J. A. et Sousa, P. (2006). Issues in multiagent resource allocation. *Informatica (Slovenia)*, 30(1):3–31.

[Chevaleyre *et al.*, 2007] Chevaleyre, Y., Endriss, U., Lang, J. et Maudet, N. (2007). *A short introduction to computational social choice*. Springer.

[Clippel et Serrano, 2005] Clippel, G. D. et Serrano, R. (2005). Marginal contributions and externalities in the value. Working Paper, Brown University, Department of Economics 2005-11.

[Coleman, 1986] Coleman, J. S. (1986). Social theory, social research, and a theory of action. *American journal of Sociology*, 91(6):1309–1335.

[Conitzer *et al.*, 2003] Conitzer, V., Lang, J. et Sandholm, T. (2003). How many candidates are needed to make elections hard to manipulate ? *In Proceedings of the 9th conference on Theoretical aspects of rationality and knowledge*, pages 201–214.

[Conitzer et Sandholm, 2002] Conitzer, V. et Sandholm, T. (2002). Complexity of manipulating elections with few candidates. *In Proceedings of the 18th National Conference on Artificial intelligence, AAAI 2002*, pages 314–319.

[Conitzer et Sandholm, 2004] Conitzer, V. et Sandholm, T. (2004). Computing shapley values, manipulating value division schemes, and checking core membership in multi-issue domains. *In Proceedings of the 19th National Conference on Artificial intelligence, AAAI 2004*, volume 4, pages 219–225.

[Conitzer et Sandholm, 2006] Conitzer, V. et Sandholm, T. (2006). Nonexistence of voting rules that are usually hard to manipulate. *In Proceedings of the 21st National Conference on Artificial intelligence, AAAI 2006*, volume 6, pages 627–634.

[Conte *et al.*, 1999] Conte, R., Castelfranchi, C. et Dignum, F. (1999). *Autonomous norm acceptance*. Springer.

[Coughlin, 1982] Coughlin, P. (1982). Pareto optimality of policy proposals with probabilistic voting. *Public Choice*, 39(3):427–433.

[Danezis et Mittal, 2009] Danezis, G. et Mittal, P. (2009). Sybilinfer : Detecting sybil nodes using social networks. *In Proceedings of the Network and Distributed System Security, NDSS2009*.

[Dang et Jennings, 2004] Dang, V. D. et Jennings, N. R. (2004). Generating coalition structures with finite bound from the optimal guarantees. *In Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS 2004*, pages 564–571.

[Dasgupta *et al.*, 1999] Dasgupta, P., Narasimhan, N., Moser, L. E. et Melliar-Smith, P. (1999). Magnet : mobile agents for networked electronic trading. *Knowledge and Data Engineering, IEEE Transactions on*, 11(4):509–525.

[de Borda, 1781] de Borda, J. C. (1781). *Mémoire sur les élections au scrutin*. Histoire de l'Academie Royale des Sciences.

[de Condorcet, 1785] de Condorcet, N. (1785). *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. L'imprimerie royale.

[Delannoi et Dowlen, 2010] Delannoi, G. et Dowlen, O. (2010). *Sortition, Theory and Practice*. Academic UK and USA.

[Dellarocas, 2000] Dellarocas, C. (2000). Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. *In Proceedings of the 2nd ACM conference on Electronic commerce*, pages 150–157.

[Dimitrov *et al.*, 2006] Dimitrov, D., Borm, P., Hendrickx, R. et Sung, S. C. (2006). Simple priorities and core stability in hedonic games. *Social Choice and Welfare*, 26(2):421–433.

[Douceur, 2002] Douceur, J. R. (2002). The sybil attack. *In Peer-to-peer Systems*, pages 251–260. Springer.

[Drèze et Greenberg, 1980] Drèze, J. H. et Greenberg, J. (1980). Hedonic coalitions : Optimality and stability. *Econometrica*, 48(4):987–1003.

[Drogoul et Ferber, 1994] Drogoul, A. et Ferber, J. (1994). Multi-agent simulation as a tool for modeling societies : Application to social differentiation in ant colonies. *In Artificial Social Systems*, pages 2–23. Springer.

[Duersch *et al.*, 2012] Duersch, P., Oechssler, J. et Schipper, B. C. (2012). Pure strategy equilibria in symmetric two-player zero-sum games. *International Journal of Game Theory*, 41(3):553–564.

[Dumitriu *et al.*, 2005] Dumitriu, D., Knightly, E., Kuzmanovic, A., Stoica, I. et Zwaenepoel, W. (2005). Denial-of-service resilience in peer-to-peer file sharing systems. *In Proceedings of the ACM SIGMETRICS Performance Evaluation Review*, volume 33, pages 38–49.

[Elkind et Lipmaa, 2005] Elkind, E. et Lipmaa, H. (2005). Hybrid voting protocols and hardness of manipulation. *In Algorithms and Computation*, pages 206–215. Springer.

[Elkind et Wooldridge, 2009] Elkind, E. et Wooldridge, M. (2009). Hedonic coalition nets. *In Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2009*, pages 417–424.

[Ellison *et al.*, 1997] Ellison, R. J., Fisher, D. A., Linger, R. C., Lipson, H. F. et Longstaff, T. (1997). Survivable network systems : An emerging discipline. Rapport technique, DTIC Document.

[Endriss *et al.*, 2006] Endriss, U., Maudet, N., Sadri, F. et Toni, F. (2006). Negotiating socially optimal allocations of resources. *J. Artif. Intell. Res.(JAIR)*, 25:315–348.

[Faratin *et al.*, 1998] Faratin, P., Sierra, C. et Jennings, N. R. (1998). Negotiation decision functions for autonomous agents. *Robotics and Autonomous Systems*, 24(3):159–182.

[Feldman *et al.*, 2006] Feldman, M., Papadimitriou, C., Chuang, J. et Stoica, I. (2006). Free-riding and whitewashing in peer-to-peer systems. *Selected Areas in Communications, IEEE Journal on*, 24(5):1010–1019.

[Ferber, 1999] Ferber, J. (1999). *Multi-agent systems - an introduction to distributed artificial intelligence*. Addison-Wesley-Longman.

[Foster et Young, 2003] Foster, D. P. et Young, H. P. (2003). Learning, hypothesis testing, and nash equilibrium. *Games and Economic Behavior*, 45(1):73–96.

[Franklin et Graesser, 1997] Franklin, S. et Graesser, A. (1997). Is it an agent, or just a program ? : A taxonomy for autonomous agents. *In Intelligent agents III agent theories, architectures, and languages*, pages 21–35. Springer.

[Friedman *et al.*, 2007] Friedman, E., Resnick, P. et Sami, R. (2007). *Algorithmic Game Theory*, chapitre Manipulation-resistant reputation systems, pages 677–697. Cambridge University Press Cambridge, UK.

[Gamson, 1961] Gamson, W. A. (1961). A theory of coalition formation. *American Sociological Review*, 26(3):373–382.

[Gärdenfors, 1976] Gärdenfors, P. (1976). Manipulation of social choice functions. *Journal of Economic Theory*, 13(2):217–228.

[Génin, 2010] Génin, T. (2010). *Stratégies de formation de coalitions dans les systèmes multi-agents*. Thèse de doctorat, Paris 6.

[Génin et Aknine, 2011] Génin, T. et Aknine, S. (2011). Étude de protocoles et de stratégies de négociation pour l'obtention de structures de coalitions pareto optimales dans un problème de formation de coalitions. *In Sixièmes Journées Francophones Modèles formels de l'interaction, MFI 2011*, pages 187–196.

[Gibbard, 1973] Gibbard, A. (1973). Manipulation of voting schemes : a general result. *Econometrica : journal of the Econometric Society*, 41(4):587–601.

[Gilbert, 1995] Gilbert, N. (1995). Simulation : an emergent perspective. *In Proceedings of the conference on New Technologies in the Social Sciences*, pages 27–29.

[Gittins, 1979] Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):148–177.

[Golbeck, 2006] Golbeck, J. (2006). Computing with trust : Definition, properties, and algorithms. *In Proceedings of the Securecomm and Workshops, 2006*, pages 1–7.

[Grabisch et Funaki, 2012] Grabisch, M. et Funaki, Y. (2012). A coalition formation value for games in partition function form. *European Journal of Operational Research*, 221(1):175–185.

[Grandison et Sloman, 2000] Grandison, T. et Sloman, M. (2000). A survey of trust in internet applications. *Communications Surveys & Tutorials, IEEE*, 3(4):2–16.

[Gu *et al.*, 2008] Gu, G., Zhang, J. et Lee, W. (2008). Botsniffer : Detecting botnet command and control channels in network traffic. *In Proceedings of the Network and Distributed System Security Symposium, NDSS 2008*.

[Guo et Conitzer, 2010] Guo, M. et Conitzer, V. (2010). Strategy-proof allocation of multiple items between two agents without payments or priors. *In Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2010*, pages 881–888.

[Guttman et Maes, 1998] Guttman, R. H. et Maes, P. (1998). Cooperative vs. competitive multi-agent negotiations in retail electronic commerce. *In Cooperative Information Agents II Learning, Mobility and Electronic Commerce for Information Discovery on the Internet*, pages 135–147. Springer.

[Hajduková *et al.*, 2003] Hajduková, J. *et al.* (2003). Computational complexity of stable partitions with b-preferences. *International Journal of Game Theory*, 31(3):353–364.

[Hajduková *et al.*, 2004] Hajduková, J. *et al.* (2004). Stable partitions with w-preferences. *Discrete Applied Mathematics*, 138(3):333–347.

[Harsanyi, 1963] Harsanyi, J. C. (1963). A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220.

[Harsanyi et Selten, 1988] Harsanyi, J. C. et Selten, R. (1988). *A General Theory of Equilibrium Selection in Games*, volume 1. The MIT Press.

[Hart et Kurz, 1983] Hart, S. et Kurz, M. (1983). Endogenous formation of coalitions. *Econometrica : Journal of the Econometric Society*, 51:1047–1064.

[Hershey *et al.*, 2007] Hershey, J. R., Olsen, P. *et al.* (2007). Approximating the kullback leibler divergence between gaussian mixture models. *In Proceedings of the Acoustics, Speech and Signal*, volume 4.

[Hildrum *et al.*, 2004] Hildrum, K., Kubiatowicz, J. D., Rao, S. et Zhao, B. Y. (2004). Distributed object location in a dynamic network. *Theory of Computing Systems*, 37(3):405–440.

[Hoefer *et al.*, 2014] Hoefer, M., Váz, D. et Wagner, L. (2014). Hedonic coalition formation in networks. *In Proceedings of the 28th Conference on Artificial intelligence, AAAI 2014*.

[Hoffman *et al.*, 2009] Hoffman, K., Zage, D. et Nita-Rotaru, C. (2009). A survey of attack and defense techniques for reputation systems. *ACM Computer Survey*, 42(1):1–31.

[Huynh *et al.*, 2006] Huynh, T. D., Jennings, N. R. et Shadbolt, N. R. (2006). An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154.

[Ishikida et Varaiya, 1994] Ishikida, T. et Varaiya, P. (1994). Multi-armed bandit problem revisited. *Journal of Optimization Theory and Applications*, 83(1):113–154.

[Janssen et Jager, 2003] Janssen, M. A. et Jager, W. (2003). Simulating market dynamics : Interactions between consumer psychology and social networks. *Artificial Life*, 9(4):343–356.

[Jennings et Wooldridge, 1996] Jennings, N. et Wooldridge, M. (1996). Software agents. *IEE review*, 42(1):17–20.

[Jensen et Meckling, 1994] Jensen, M. C. et Meckling, W. H. (1994). Self-interest, altruism, incentives, and agency theory. *Journal of applied corporate finance*, 7(2):40–45.

[Jin *et al.*, 2007] Jin, Y., Gu, Z. et Ban, Z. (2007). Restraining false feedbacks in peer-to-peer reputation systems. *In Proceedings of the Semantic International Conference on Computing, ICSC 2007*, pages 304–312.

[Johansson *et al.*, 2008] Johansson, B., Speranzon, A., Johansson, M. et Johansson, K. H. (2008). On decentralized negotiation of optimal consensus. *Automatica*, 44(4):1175–1179.

[Jøsang et Golbeck, 2009] Jøsang, A. et Golbeck, J. (2009). Challenges for robust trust and reputation systems. *In Proceedings of the 5th International Workshop on Security and Trust Management (SMT 2009), Saint Malo, France.*

[Jøsang et Haller, 2007] Jøsang, A. et Haller, J. (2007). Dirichlet reputation systems. *In Proceedings of the The 2nd International Conference on Availability, Reliability and Security, ARES 2007*, pages 112–119.

[Jøsang et Ismail, 2002] Jøsang, A. et Ismail, R. (2002). The Beta reputation system. *In Proceedings of the 15th Bled Electronic Commerce Conference*, pages 41–55.

[Jøsang *et al.*, 2007] Jøsang, A., Ismail, R. et Boyd, C. (2007). A survey of trust and reputation systems for online service provision. *Decision support systems*, 43(2):618–644.

[Kamvar *et al.*, 2003] Kamvar, S. D., Schlosser, M. T. et Garcia-Molina, H. (2003). The EigenTrust algorithm for reputation management in P2P networks. *In Proceedings of the 12th International World Wide Web Conference*, pages 640–651.

[Katehakis et Veinott Jr, 1987] Katehakis, M. N. et Veinott Jr, A. F. (1987). The multi-armed bandit problem : decomposition and computation. *Mathematics of Operations Research*, 12(2):262–268.

[Keinänen, 2010] Keinänen, H. (2010). An algorithm for generating nash stable coalition structures in hedonic games. *In Foundations of Information and Knowledge Systems*, pages 25–39. Springer.

[Kelso Jr et Crawford, 1982] Kelso Jr, A. S. et Crawford, V. P. (1982). Job matching, coalition formation, and gross substitutes. *Econometrica : Journal of the Econometric Society*, 50(05):1483–1504.

[Kent et Atkinson, 1998] Kent, S. et Atkinson, R. (1998). Security architecture for the internet protocol.

[Kitano *et al.*, 1999] Kitano, H., Tadokoro, S., Noda, I., Matsubara, H., Takahashi, T., Shinjou, A. et Shimada, S. (1999). Robocup rescue : Search and rescue in large-scale disasters as a domain for autonomous agents research. *In Proceedings of the International Conference on Systems, Man, and Cybernetics, SMC 1999*, volume 6, pages 739–743.

[Kleijnen, 1995] Kleijnen, J. P. (1995). Verification and validation of simulation models. *European Journal of Operational Research*, 82(1):145–162.

[Koops et Leenes, 2006] Koops, B.-J. et Leenes, R. (2006). Identity theft, identity fraud and/or identity-related crime. *Datenschutz und Datensicherheit-DuD*, 30(9):553–556.

[Koulouriotis et Xanthopoulos, 2008] Koulouriotis, D. E. et Xanthopoulos, A. (2008). Reinforcement learning and evolutionary algorithms for non-stationary multi-armed bandit problems. *Applied Mathematics and Computation*, 196(2):913–922.

[Koutrouli et Tsalgatidou, 2011] Koutrouli, E. et Tsalgatidou, A. (2011). Credibility enhanced reputation mechanism for distributed e-communities. *In Proceedings of the 19th Euromicro International Conference on Parallel, Distributed and Network-Based Processing, PDP 2011*, pages 627–634.

[Kuipers et Byun, 1991] Kuipers, B. et Byun, Y.-T. (1991). A robot exploration and mapping strategy based on a semantic hierarchy of spatial representations. *Robotics and autonomous systems*, 8(1):47–63.

[Kuleshov et Precup, 2014] Kuleshov, V. et Precup, D. (2014). Algorithms for multi-armed bandit problems. *CoRR*.

[Kullback, 1968] Kullback, S. (1968). *Information theory and statistics*. Courier Corporation.

[Larson et Sandholm, 2000] Larson, K. S. et Sandholm, T. W. (2000). Anytime coalition structure generation : an average case study. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(1):23–42.

[Lau *et al.*, 2000] Lau, F., Rubin, S. H., Smith, M. H. et Trajkovic, L. (2000). Distributed denial of service attacks. *In Proceedings of the International Conference on Systems, Man, and Cybernetics, SMC 2000*, volume 3, pages 2275–2280.

[Leskovec *et al.*, 2008] Leskovec, J., Lang, K., Dasgupta, A. et Mahoney, M. (2008). Statistical properties of community structure in large social and information networks. *In 17th International World Wide Web Conference*, pages 695–704.

[Levien et Aiken, 1998] Levien, R. et Aiken, A. (1998). Attack-resistant trust metrics for public key certification. *In Proceedings of 7th Usenix Security Symposium*.

[Levine *et al.*, 2006] Levine, B. N., Shields, C. et Margolin, N. B. (2006). A survey of solutions to the sybil attack. *University of Massachusetts Amherst, Amherst, MA*.

[Liu et Zhao, 2010] Liu, K. et Zhao, Q. (2010). Distributed learning in multi-armed bandit with multiple players. *Signal Processing, IEEE Transactions on*, 58(11):5667–5681.

[Luck et d'Inverno, 2001] Luck, M. et d'Inverno, M. (2001). Autonomy : A nice idea in theory. *In Intelligent Agents VII Agent Theories Architectures and Languages*, pages 351–353. Springer.

[Malik et Bouguettaya, 2009] Malik, Z. et Bouguettaya, A. (2009). Rateweb : Reputation assessment for trust establishment among web services. *The VLDB Journal—The International Journal on Very Large Data Bases*, 18(4):885–911.

[Margolin et Levine, 2008] Margolin, N. B. et Levine, B. N. (2008). Quantifying resistance to the sybil attack. *In Proceedings of the 12th International Conference on Financial Cryptography and Data Security, FC 2008*, pages 1–15.

[Marti et Garcia-Molina, 2006] Marti, S. et Garcia-Molina, H. (2006). Taxonomy of trust : Categorizing p2p reputation systems. *Computer Networks*, 50(4):472–484.

[Maskin, 1999] Maskin, E. (1999). Nash equilibrium and welfare optimality. *The Review of Economic Studies*, 66(1):23–38.

[Matsui *et al.*, 2008] Matsui, T., Matsuo, H., Silaghi, M., Hirayama, K. et Yokoo, M. (2008). Resource constrained distributed constraint optimization with virtual variables. *In Proceedings of the 23rd Conference on Artificial intelligence, AAAI 2008*, pages 120–125.

[Mazouzi *et al.*, 2002] Mazouzi, H., Seghrouchni, A. E. F. et Haddad, S. (2002). Open protocol design for complex interactions in multi-agent systems. *In Proceedings of the 1th international joint conference on Autonomous agents and multiagent systems, AAMAS 2002*, pages 517–526.

[McKelvey *et al.*, 1978] McKelvey, R. D., Ordeshook, P. C. et Winer, M. D. (1978). The competitive solution for n-person games without transferable utility, with an application to committee games. *American Political Science Review*, 72(02):599–615.

[Meyer et Wetzel, 2004] Meyer, U. et Wetzel, S. (2004). A man-in-the-middle attack on umts. *In Proceedings of the 3rd ACM workshop on Wireless security*, pages 90–97.

[Michalak *et al.*, 2010] Michalak, T., Sroka, J., Rahwan, T., Wooldridge, M., McBurney, P. et Jennings, N. R. (2010). A distributed algorithm for anytime coalition structure generation. *In Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2010*, pages 1007–1014.

[Milgrom et Weber, 1985] Milgrom, P. R. et Weber, R. J. (1985). Distributional strategies for games with incomplete information. *Mathematics of Operations Research*, 10(4):619–632.

[Miller *et al.*, 2005] Miller, N., Resnick, P. et Zeckhauser, R. (2005). Eliciting informative feedback : The peer-prediction method. *Management Science*, 51(9):1359–1373.

[Morgenstern et Von Neumann, 1953] Morgenstern, O. et Von Neumann, J. (1953). *Theory of games and economic behavior*. Princeton University Press.

[Mui *et al.*, 2002] Mui, L., Mohtashemi, M. et Halberstadt, A. (2002). A computational model of trust and reputation. *In Proceedings of the 35th Annual Hawaii International Conference on the System Sciences, HICSS 2002*, pages 2431–2439.

[Muller et Vercouter, 2004] Muller, G. et Vercouter, L. (2004). Détection décentralisée d'agents menteurs (article court). *In Douxièmes journées francophones sur les systèmes multi-agents, JFSMA 04*, pages 243–248.

[Muller et Vercouter, 2005] Muller, G. et Vercouter, L. (2005). Decentralized monitoring of agent communications with a reputation model. *In Trusting Agents for Trusting Electronic Societies*, pages 144–161. Springer.

[Myerson, 2013] Myerson, R. B. (2013). *Game theory*. Harvard university press.

[Nash, 1951] Nash, J. (1951). Non-cooperative games. *Annals of mathematics*, 54:286–295.

[Nash, 1950] Nash, J. F. (1950). Equilibrium points in n-person games. *National Academy of Sciences of the United States of America*, 36(1):48–49.

[Nehring et Puppe, 2007] Nehring, K. et Puppe, C. (2007). The structure of strategy-proof social choice - part i : General characterization and possibility results on median spaces. *Journal of Economic Theory*, 135(1):269–305.

[Newsome *et al.*, 2004] Newsome, J., Shi, E., Song, D. et Perrig, A. (2004). The sybil attack in sensor networks : analysis & defenses. *In Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 259–268.

[Okada, 1996] Okada, A. (1996). A noncooperative coalitional bargaining game with random proposers. *Games and Economic Behavior*, 16(1):97–108.

[Olfati-Saber *et al.*, 2007] Olfati-Saber, R., Fax, J. A. et Murray, R. M. (2007). Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233.

[Oo et Aung, 2014] Oo, H. N. et Aung, A. M. (2014). A survey of different electronic voting systems. *International Journal of Scientific Engineering and Technology Research*, 03(16): 3460–3464.

[Osborne et Rubinstein, 1994] Osborne, M. J. et Rubinstein, A. (1994). *A course in game theory*. MIT press.

[Page *et al.*, 1999] Page, L., Brin, S., Motwani, R. et Winograd, T. (1999). The PageRank citation ranking : bringing order to the Web. Rapport technique, Stanford InfoLab.

[Panait et Luke, 2005] Panait, L. et Luke, S. (2005). Cooperative multi-agent learning : The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11(3):387–434.

[Pardalos *et al.*, 2008] Pardalos, P. M., Migdalas, A. et Pitsoulis, L. (2008). *Pareto optimality, game theory and equilibria*, volume 17. Springer Science & Business Media.

[Parunak, 1997] Parunak, H. V. D. (1997). " go to the ant" : Engineering principles from natural multi-agent systems. *Annals of Operations Research*, 75:69–101.

[Pazzani et Billsus, 2007] Pazzani, M. J. et Billsus, D. (2007). Content-based recommendation systems. *In The adaptive web*, pages 325–341. Springer.

[Peters et Elkind, 2015] Peters, D. et Elkind, E. (2015). Simple causes of complexity in hedonic games. *In Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 617–623.

[Pinson et Moraitis, 1997] Pinson, S. et Moraitis, P. (1997). An intelligent distributed system for strategic decision making. *Group Decision and Negotiation*, 6(1):77–108.

[Pinyol et Sabater-Mir, 2013] Pinyol, I. et Sabater-Mir, J. (2013). Computational trust and reputation models for open multi-agent systems : a review. *Artificial Intelligence Review*, 40(1):1–25.

[Poundstone *et al.*, 1993] Poundstone, W., By-Neumann, B. O. W. et Von, J. (1993). *Prisoner's Dilemma*. Doubleday.

[Rahwan, 2007] Rahwan, T. (2007). *Algorithms for coalition formation in multi-agent systems*. Thèse de doctorat, University of Southampton.

[Rahwan et Jennings, 2007] Rahwan, T. et Jennings, N. R. (2007). An algorithm for distributing coalitional value calculations among cooperating agents. *Artificial Intelligence*, 171(8):535–567.

[Rahwan et Jennings, 2008] Rahwan, T. et Jennings, N. R. (2008). Coalition structure generation : Dynamic programming meets anytime optimization. *In Proceedings of the 23rd Conference on Artificial intelligence, AAAI 2008*, volume 8, pages 156–161.

[Rahwan *et al.*, 2007] Rahwan, T., Ramchurn, S. D., Dang, V. D., Giovannucci, A. et Jennings, N. R. (2007). Anytime optimal coalition structure generation. *In Proceedings of the 22nd Conference on Artificial intelligence AAAI 2007*, volume 7, pages 1184–1190.

[Railsback *et al.*, 2006] Railsback, S. F., Lytinen, S. L. et Jackson, S. K. (2006). Agent-based simulation platforms : Review and development recommendations. *Simulation*, 82(9): 609–623.

[Ray et Vohra, 1999] Ray, D. et Vohra, R. (1999). A theory of endogenous coalition structures. *Games and Economic Behavior*, 26(2):286–336.

[Renoux *et al.*, 2014] Renoux, J., Mouaddib, A.-I. et Le Gloannec, S. (2014). Un modèle de décision distribué pour la collecte d'information active multiagents. *In Reconnaissance de Formes et Intelligence Artificielle, RFIA 2014*.

[Resnick, 1994] Resnick, M. (1994). *Turtles, termites, and traffic jams : Explorations in massively parallel microworlds*. Mit Press.

[Resnick *et al.*, 2001] Resnick, P. *et al.* (2001). The social cost of cheap pseudonyms. *Journal of Economics & Management Strategy*, 10(2):173–199.

[Resnick *et al.*, 2000] Resnick, P., Kuwabara, K., Zeckhauser, R. et Friedman, E. (2000). Reputation systems. *ACM Communications*, 43(12):45–48.

[Resnick *et al.*, 2006] Resnick, P., Zeckhauser, R., Swanson, J. et Lockwood, K. (2006). The value of reputation on ebay : A controlled experiment. *Experimental Economics*, 9(2):79–101.

[Robbins, 1952] Robbins, H. (1952). Some aspects of the sequential design of experiments. *Journal of the AMS*, 58(5):527–535.

[Robinson, 1985] Robinson, M. S. (1985). Collusion and the choice of auction. *The RAND Journal of Economics*, pages 141–145.

[Rothkopf *et al.*, 1998] Rothkopf, M. H., Pekeč, A. et Harstad, R. M. (1998). Computationally manageable combinational auctions. *Management science*, 44(8):1131–1147.

[Russell et Norvig, 1995] Russell, S. J. et Norvig, P. (1995). *Artificial intelligence - a modern approach : the intelligent agent book*. Prentice Hall.

[Sabater *et al.*, 2006] Sabater, J., Paolucci, M. et Conte, R. (2006). Repage : Reputation and image among limited autonomous partners. *Journal of artificial societies and social simulation*, 9(2).

[Sabater et Sierra, 2001] Sabater, J. et Sierra, C. (2001). Regret : A reputation model for gregarious societies. *In Proceedings of the 4th workshop on deception fraud and trust in agent societies*, volume 70.

[Sabater et Sierra, 2005] Sabater, J. et Sierra, C. (2005). Review on computational trust and reputation models. *Artificial intelligence review*, 24(1):33–60.

[Sandhlom et Lesser, 1997] Sandhlom, T. W. et Lesser, V. R. (1997). Coalitions among computationally bounded agents. *Artificial intelligence*, 94(1):99–137.

[Sandholm, 1993] Sandholm, T. (1993). An implementation of the contract net protocol based on marginal cost calculations. *In Proceedings of the 11th National Conference on Artificial intelligence, AAAI 1993*, volume 93, pages 256–262.

[Sandholm, 2002] Sandholm, T. (2002). Algorithm for optimal winner determination in combinatorial auctions. *Artificial intelligence*, 135(1):1–54.

[Sandholm, 1999] Sandholm, T. W. (1999). *Multiagent systems : a modern approach to distributed artificial intelligence*, chapitre Distributed rational decision making, pages 201–258. MIT press.

[Satterthwaite, 1975] Satterthwaite, M. A. (1975). Strategy-proofness and arrow's conditions : Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of economic theory*, 10(2):187–217.

[Schafer *et al.*, 1999] Schafer, J. B., Konstan, J. et Riedl, J. (1999). Recommender systems in e-commerce. *In Proceedings of the 1st ACM conference on Electronic commerce*, pages 158–166.

[Schmeidler, 1969] Schmeidler, D. (1969). The nucleolus of a characteristic function game. *SIAM Journal on applied mathematics*, 17(6):1163–1170.

[Sen, 1986] Sen, A. (1986). Social choice theory. *Handbook of mathematical economics*, 3:1073–1181.

[Serenko et Detlor, 2004] Serenko, A. et Detlor, B. (2004). Intelligent agents as innovations. *Ai & Society*, 18(4):364–381.

[Seuken et Zilberstein, 2008] Seuken, S. et Zilberstein, S. (2008). Formal models and algorithms for decentralized decision making under uncertainty. *Autonomous Agents and Multi-Agent Systems*, 17(2):190–250.

[Shannon, 1951] Shannon, C. E. (1951). Prediction and entropy of printed english. *Bell system technical journal*, 30(1):50–64.

[Shapley, 1952] Shapley, L. S. (1952). A value for n-person games. Rapport technique.

[Shehory et Kraus, 1998] Shehory, O. et Kraus, S. (1998). Methods for task allocation via agent coalition formation. *Artificial Intelligence*, 101(1):165–200.

[Shen *et al.*, 2006] Shen, W., Wang, L. et Hao, Q. (2006). Agent-based distributed manufacturing process planning and scheduling : a state-of-the-art survey. *Systems, Man, and Cybernetics, Part C : Applications and Reviews, IEEE Transactions on*, 36(4):563–577.

[Singh *et al.*, 2006] Singh, A. *et al.* (2006). Eclipse attacks on overlay networks : Threats and defenses. *In Proceedings of the INFOCOM*.

[Singh et Liu, 2003] Singh, A. et Liu, L. (2003). Trustme : Anonymous management of trust relationships in decentralized p2p systems. *In Proceedings of the International Conference on Peer-to-Peer Computing*, page 1.

[Smith, 1980] Smith, R. (1980). Communication and control in problem solver. *IEEE Transactions on computers*, 29(12):1104–1113.

[Specht et Lee, 2004] Specht, S. M. et Lee, R. B. (2004). Distributed denial of service : Taxonomies of attacks, tools, and countermeasures. *In Proceedings of the International Conference on Parallel and Distributed Computing (and Communications) Systems*, pages 543–550.

[Srivatsa *et al.*, 2005] Srivatsa, M., Xiong, L. et Liu, L. (2005). TrustGuard : countering vulnerabilities in reputation management for decentralized overlay networks. *In Proceedings of the 14th International World Wide Web Conference*, pages 422–431.

[Staab et Engel, 2009] Staab, E. et Engel, T. (2009). Collusion detection for grid computing. *In Proceedings of the 9th International Symposium on Cluster Computing and the Grid, CCGrid 2009*, pages 412–419.

[Stephen, 1994] Stephen, M. (1994). *Formalising trust as a computational concept.* Thèse de doctorat, University of Stirling, scotland.

[Stone et Veloso, 2000] Stone, P. et Veloso, M. (2000). Multiagent systems : A survey from a machine learning perspective. *Autonomous Robots*, 8(3):345–383.

[Su *et al.*, 2013] Su, X., Zhang, M., Mu, Y. et Bai, Q. (2013). A robust trust model for service-oriented systems. *Journal of Computer and System Sciences*, 79(5):596–608.

[Sung et Dimitrov, 2010] Sung, S.-C. et Dimitrov, D. (2010). Computational complexity in additive hedonic games. *European Journal of Operational Research*, 203(3):635–639.

[Suzuki *et al.*, 2015] Suzuki, T. *et al.* (2015). Solutions for cooperative games with and without transferable utility. Rapport technique, School of Economics and Management.

[Tavakolifard et Almeroth, 2012] Tavakolifard, M. et Almeroth, K. C. (2012). A taxonomy to express open challenges in trust and reputation systems. *Journal of Communications*, 7(7):538–551.

[Tennenholtz, 2004] Tennenholtz, M. (2004). Reputation systems : An axiomatic approach. *In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 544–551.

[Theodorakopoulos et Baras, 2006] Theodorakopoulos, G. et Baras, J. S. (2006). On trust models and trust evaluation metrics for ad hoc networks. *Selected Areas in Communications, IEEE Journal on*, 24(2):318–328.

[Tideman, 2006] Tideman, N. (2006). *Collective decisions and voting : the potential for public choice.* Ashgate Publishing, Ltd.

[Tran-Thanh *et al.*, 2010] Tran-Thanh, L., Chapman, A., Munoz De Cote Flores Luna, J. E., Rogers, A. et Jennings, N. R. (2010). Epsilon–first policies for budget–limited multi-armed bandits. *In Proceedings of the 24th Conference on Artificial intelligence, AAAI 2010.*

[Vallée et Bonnet, 2015] Vallée, T. et Bonnet, G. (2015). Using kl divergence for credibility assessment. *In Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2015*, pages 1797–1798.

[Vallée *et al.*, 2014b] Vallée, T., Bonnet, G. et Bourdon, F. (2014b). Multi-armed bandit policies for reputation systems. *In Advances in Practical Applications of Heterogeneous Multi-Agent Systems. The PAAMS Collection*, pages 279–290. Springer.

[Vallée *et al.*, 2015] Vallée, T., Bonnet, G. et Bourdon, F. (2015). Politiques de bandits manchots et crédibilité dans les systèmes de réputation. *Revue d'intelligence artificielle*, 29(3-4):369–398.

[Vallée *et al.*, 2014a] Vallée, T., Bonnet, G., Bourdon, F. *et al.* (2014a). De l'utilisation des politiques de bandits manchots dans les systèmes de réputation. *In Reconnaissance de Formes et Intelligence Artificielle, RFIA 2014.*

[Vallée *et al.*, 2013] Vallée, T., Bonnet, G., Zanuttini, B. et Bourdon, F. (2013). Étude des attaques sybil sur les jeux hédoniques. *In 7e journées francophones Modèles Formels de l'Interaction, MFI 2013*, page 12 p.

[Vallée *et al.*, 2014c] Vallée, T., Bonnet, G., Zanuttini, B. et Bourdon, F. (2014c). A study of sybil manipulations in hedonic games. *In Proceedings of the 13th international conference on Autonomous agents and multi-agent systems, AAMAS 2014*, pages 21–28.

[Vauvert et El Fallah-Seghrouchni, 2001] Vauvert, G. et El Fallah-Seghrouchni, A. (2001). Coalition formation among strong autonomous and weak rational agents. *In Proceedings. 10th European Workshop Modelling Autonomous Agents in a Multi-agent World.*

[Vermorel et Mohri, 2005] Vermorel, J. et Mohri, M. (2005). Multi-armed bandit algorithms and empirical evaluation. *In Proceedings of the 16th European Conference on Machine Learning, Porto, ECML 2005*, pages 437–448.

[Von Ahn *et al.*, 2003] Von Ahn, L., Blum, M., Hopper, N. J. et Langford, J. (2003). Captcha : Using hard ai problems for security. *In Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques, EUROCRYPT2003*, pages 294–311. Springer.

[von Neumann et Morgenstern, 1944] von Neumann, J. et Morgenstern, O. (1944). Theory of games and economic behavior. *Nature*, 246:15–18.

[Waggoner *et al.*, 2012] Waggoner, B., Xia, L. et Conitzer, V. (2012). Evaluating resistance to false-name manipulations in elections. *In Proceedings of the 26th Conference on Artificial intelligence, AAAI 2015*.

[Walsh, 2009] Walsh, T. (2009). Where are the really hard manipulation problems ? the phase transition in manipulating the veto rule. *In Proceedings of the 21st International Joint Conference on Artificial Intelligence, IJCAI 2009*, pages 324–329.

[Walter *et al.*, 2008] Walter, F. E., Battiston, S. et Schweitzer, F. (2008). A model of a trust-based recommendation system on a social network. *Autonomous Agents and Multi-Agent Systems*, 16(1):57–74.

[Wang et Vassileva, 2003] Wang, Y. et Vassileva, J. (2003). Trust and reputation model in peer-to-peer networks. *In Proceedings of the 3rd International Conference on the Peer-to-Peer Computing, P2P 2003*, pages 150–157.

[Weyns *et al.*, 2007] Weyns, D., Omicini, A. et Odell, J. (2007). Environment as a first class abstraction in multiagent systems. *Autonomous agents and multi-agent systems*, 14(1):5–30.

[Whitby *et al.*, 2004] Whitby, A., Jøsang, A. et Indulska, J. (2004). Filtering out unfair ratings in bayesian reputation systems. *In Proceedings of the 7th Int. Workshop on Trust in Agent Societies*, volume 6, pages 106–117.

[Whittle, 1980] Whittle, P. (1980). Multi-armed bandits and the gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)*, 42(2):143–149.

[Wieder, 2008] Wieder, T. (2008). The number of certain k-combinations of an n-set. *Applied Mathematics E-Notes*, 8:45–52.

[Winter, 1989] Winter, E. (1989). A value for cooperative games with levels structure of cooperation. *International Journal of Game Theory*, 18(2):227–240.

[Winter, 1991] Winter, E. (1991). On non-transferable utility games with coalition structure. *International Journal of Game Theory*, 20(1):53–63.

[Wood et Stankovic, 2002] Wood, A. et Stankovic, J. A. (2002). Denial of service in sensor networks. *Computer*, 35(10):54–62.

[Wooldridge, 2009] Wooldridge, M. (2009). *An introduction to multiagent systems*. John Wiley & Sons.

[Xia *et al.*, 2009] Xia, L., Zuckerman, M., Procaccia, A. D., Conitzer, V. et Rosenschein, J. S. (2009). Complexity of unweighted coalitional manipulation under some common voting rules. *In Proceedings of the 21st International Joint Conference on Artificial Intelligence, IJCAI 2009*, volume 9, pages 348–352.

[Xiong et Liu, 2004] Xiong, L. et Liu, L. (2004). Peertrust : Supporting reputation-based trust for peer-to-peer electronic communities. *Knowledge and Data Engineering, IEEE Transactions on*, 16(7):843–857.

[Xiu et Liu, 2005] Xiu, D. et Liu, Z. (2005). A formal definition for trust in distributed systems. *In Information Security*, pages 482–489. Springer.

[Young, 1995] Young, P. (1995). Optimal voting rules. *The Journal of Economic Perspectives*, 9(1):51–64.

[Yu *et al.*, 2006] Yu, H., Kaminsky, M., Gibbons, P. B. et Flaxman, A. (2006). SybilGuard : defending against Sybil attacks via social networks. *SIGCOMM Computer Communication Review*, 36(4):267–278.

[Yun Yeh, 1986] Yun Yeh, D. (1986). A dynamic programming approach to the complete set partitioning problem. *BIT Numerical Mathematics*, 26(4):467–474.

[Zhao et Li, 2009] Zhao, H. et Li, X. (2009). H-trust : A group trust management system for peer-to-peer desktop grid. *Journal of Computer Science and Technology*, 24(5):833–843.

[Zhou et Hwang, 2007] Zhou, R. et Hwang, K. (2007). Powertrust : A robust and scalable reputation system for trusted peer-to-peer computing. *Parallel and Distributed Systems, IEEE Transactions on*, 18(4):460–473.

De la manipulation dans les systèmes multi-agents : une étude sur les jeux hédoniques et les systèmes de réputation

Cette thèse porte sur la robustesse des systèmes multi-agents aux comportements stratégiques, génériquement appelés manipulations. Nous considérons dans ce manuscrit deux familles de systèmes, les jeux de coalitions hédoniques et les systèmes de réputation, dont les propriétés complémentaires permettent d'aborder une large gamme de questions. Dans le domaine des jeux de coalitions hédoniques, nous proposons une méthode d'analyse consistant à étudier les conditions minimales nécessaires à la mise en œuvre de manipulations efficaces. Pour cela, nous identifions trois manipulations minimales mêlant faux profils de préférences et fausses identités et étudions leur efficacité sur trois concepts de solution individuellement rationnels (stabilité au sens de Nash, au sens du cœur et stabilité individuelle). Par une étude tant théorique qu'empirique, nous montrons que la stabilité au sens de Nash est robuste aux manipulations contrairement à la stabilité individuelle et la stabilité au sens du cœur. Dans le domaine des systèmes de réputation, nous proposons de modéliser un système de réputation et le problème de décision associé par un problème de bandits manchots. Par une étude empirique, nous montrons dans un premier temps qu'utiliser des politiques de sélection ayant une facteur d'exploration adapté au système de réputation réduit les interactions avec des agents manipulateurs et augmentent le coût des manipulations. Dans un second temps, nous proposons une nouvelle mesure de crédibilité fondée sur la divergence de Kullback-Leibler et sur l'erreur d'estimation des agents pour détecter puis filtrer les faux témoignages.

Manipulation on multi-agents systems : a study on hedonic games and reputation systems

In this thesis, we study the robustness of multi-agent systems to strategic behaviors, namely manipulations. We consider in this manuscript two families of systems, hedonic games and reputation systems, which have complementary properties that allow to adress a broad range of questions. In the domain of hedonic games, we propose an analysis methodology which consists in studying the necessary minimal conditions to implement manipulations in an efficient way. To this end, we identify three minimal manipulations based on false preference profiles and false identities, and we study their efficiency on three canonical solution concepts that satisfy individual rationality (Nash stability, individual stability and core stability). By both theoretical and empirical results, we proved that Nash stability is robust to manipulations, unlike individual and core stability. In the domain of reputation systems, we propose to model reputation systems and the associated decision problem with a multiarmed bandit. Firstly, we show by an empirical study that using multiarmed bandit policies with an exploration faction tuned with respect to the reputation system reduces interactions with malicious agents and increases the cost of manipulations. Secondly, we propose a new credibility assessment based on the Kullback-Leibler divergence and the estimation error of the agents that allow to detect and filter false feedbacks.

Mots-clés : Intelligence artificielle ; Système multi-agents ; Thèorie de jeux ; Jeux de coalition hédoniques ; Systèmes de réputation ; Manipulations ; Processus de décision

---

Discipline : informatique et applications

Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen, CNRS UMR 6072, Université de Caen Normandie, 14032 CAEN cedex, FRANCE