Etude des attaques Sybil sur les jeux hédoniques

Thibaut Vallée, Grégory Bonnet, Bruno Zanuttini, François Bourdon

Étude des attaques Sybil sur les jeux hédoniques

Thibaut Vallée Grégory Bonnet Bruno Zanuttini François Bourdon

GREYC (UMR6072), ENSICAEN, Université de Caen Basse-Normandie,
Campus Côte de Nacre, Boulevard du Maréchal Juin
14032 Caen Cedex 5, France
Emails: firstname.name@unicaen.fr

Résumé :
Un jeu hédonique modélise la manière dont un ensemble d'agents décide collectivement qui va coopérer avec qui, et ceci à partir de préférences exprimées par les agents vis-à-vis des autres. Dans cet article, nous étudions une manipulation particulière sur de tels jeux, appelée attaque Sybil. Il s'agit d'une manipulation où un agent malveillant s'introduit dans le jeu sous plusieurs fausses identitées, appelées Sybil, afin d'en modifier favorablement le résultat. Nous présentons deux manipulations portant sur des jeux hédoniques ayant des solutions stables selon le concept de solution de Nash et étudions leurs propriétés. Nous montrons que ce sont en essence les deux seules attaques possibles, et qu'elles sont difficiles à calculer. Enfin, nous montrons empiriquement que de telles manipulations sont rarement possibles sur une collection de jeux aléatoires.

Mots-clés : Jeux hédoniques, Attaque Sybil, Manipulation, Systèmes multi-agents, Nash-stabilité

Abstract:
Hedonic games model agents which decide with whom they will momentarily collaborate, given some preferences on other agents. We study manipulations on such games, by a malicious agent which may introduce multiple false identities (a.k.a. Sybil agents), so that the outcome of the game is more interesting for it. Taking Nash stability as the solution concept, we exhibit two manipulations and study their properties. Then we show the rather unexpected result that these are essentially the only possible Sybil manipulations, and moreover that they are computationally hard to achieve. With small experiments, we also show that they are seldom possible in random games.

Keywords: Hedonic Games, Sybil Attack, Manipulation, Multi-Agent Systems, Nash-stability

1 Introduction

Une des problématique des systèmes multi-agents est de déterminer, à un instant donné, quels agents vont coopérer avec quels autres. De tels problèmes peuvent être modélisés par un jeu de coalition. Canoniquement, les jeux de coalition sont fondés sur des fonctions d'utilité dites transférables, mais les jeux hédoniques diffèrent de ce modèle en permettant à chaque agent d'exprimer des préférences sur les autres agents avec lesquels il souhaite coopérer. Le problème revient alors à trouver un partitionnement des agents qui est satisfaisant pour chacun d'entre eux. Parmi les partitions possibles, une partition en équilibre de Nash est une partition où aucun agent ne souhaite individuellement changer de coalition. Cependant, cette notion d'équilibre ne garantit pas l'optimalité du résultat pour tous les agents. Ainsi, un agent égoïste peut mentir sur ces préférences individuelles afin que le solution du jeu lui soit plus favorable, quand bien même cela serait au détriment des autres agents. Un tel mensonge est appelé une manipulation.

Dans cette étude, nous nous intéressons à une catégorie particulière de manipulations appelée attaque *Sybil*. Une telle manipulation consiste, pour un agent malveillant, à intégrer le système sous de multiples fausses identités, de manière à faire croire aux autres agents que ces identités constituent des agents distincts. L'agent malveillant peut alors mentir sur ces propres préférences et faire exprimer par ses fausses identités d'autres préférences définies sur mesure. Ces manipulations sont donc une généralisation du vote stratégique abondament étudié dans le cadre des systèmes de votes. Toutefois, à notre connaissance, ces manipulations n'ont pas encore été étudiées dans le cadre des jeux hédoniques.

Dans cet article, après avoir présenté un état de l'art en section 2, nous proposons en section 3 un modèle d'attaque Sybil sur les jeux hédoniques. Dans les sections 4 et 5, nous présentons alors deux attaques Sybil qui ont la particularité de n'utiliser qu'une seule fausse identité, qui ne nécessitent pas de connaître les préférences des autres agents et ne reposent que sur quelques hypothèses de bon sens. Nous présentons alors les conditions nécessaires à leur efficacité et montrons qu'il est difficile de calculer si elles le seront. Nous montrons ensuite dans la section 6 qu'il s'agit en essence des deux seules attaques possibles. Ce résultat est plutôt inattendu puique nous ne limitons pas dans notre modéle d'attaque le nombre de Sybil possibles, ni la connaissance des agents malveillants sur

les préférences des autres agents. De plus, dans la section 7, nous montrons empiriquement que les conditions nécessaires à la mise en place de ces manipulations sont rares, ce qui nous permet de dire que les jeux hédoniques ayant des solutions stables selon le concept de solution de Nash sont robustes aux attaques Sybils.

2 État de l'art

La problématique de partitionement d'un ensemble d'agents tel que tous soient satisfaits de leur coalition été largement étudié dans la littérature. De nombreux modèles permettant aux agents de déterminer quelles coalitions ils souhaitent former ont été proposés [4, 7, 12, 14, 11, 16]. Ces derniers considèrent différentes propriétés permettant de garantir un équilibre comme l'optimalité au sens de Pareto ou encore la stabilité au sens de Nash. Toutefois, décider s'il existe une partition stable au sens de Nash est un problème NP-complet [3].

Comme la stabilité d'une solution dépend des préférences de chaque agent, il est naturel de se demander si un agent malveillant peut manipuler le système en mentant sur ces propres préférences. Dans la littérature, les manipulations ont été étudiées sur de nombreux systèmes tels que les réseaux pair-à-pair [19], les systèmes de votes [5], les jeux de votes pondérés [2], les enchères combinatoires [9], les problèmes d'appariement [20], les réseaux sociaux [8] ou encore les sytèmes de réputation [17]. Dans le contexte des systèmes de votes, Bartholdi, Tovey et Trick [5] ont présentés deux familles de manipulations : les manipulations contructives et les manipulations destructives qui ont respectivement pour objectif de faire gagner ou perdre un candidat donné. Ces auteurs ont montré par ailleurs que même si certaines règles de votes, tels que la règle de Copeland, sont résistantes aux manipulations, la majorité ne le sont pas. Walsh [21] a montré empiriquement que même s'il est NP-difficile de manipuler un vote dans le pire des cas, cela reste facile en pratique pour STV et les votes par veto.

Une attaque Sybil [13] (aussi appelé *false-name manipulation* [22]) consiste à introduire de fausses identités dans un système. Différentes approches ont été proposées afin de rendre un système robuste à de telles manipulations. Nous pouvons citer l'utilisation d'une autorité centrale de certification [13] ou encore la recherche de cliques d'agents dans le graphe d'interaction du système [6, 10, 23]. Cependant, ces approches portent sur les systèmes de réputation ou les enchères combinatoires. Ainsi, á notre connaissance, il n'existe pas d'études portant sur les attaques Sybil dans le cadre des jeux hédoniques. Nous pouvons toutefois mentionner les travaux portant sur le clonage stratégique de candidats dans les systèmes de votes [15]. En effet, de tels problèmes peuvent rappeler les jeux hédoniques où dupliquer un agent revient à dupliquer les coalitions possibles sur lesquelles les agents expriment des préférences.

3 Modèle

Dans cette section, nous présentons dans un premier temps notre modèle de jeu hédonique ayant pour solutions stables le concept de solution de Nash. Dans un second temps, nous présentons notre modèle de manipulation pour les attaques Sybil sur de tels jeux. Tout au long de cette section, un exemple applicatif illustrera et justifiera nos hypothèses de travail.

3.1 Jeux hédoniques

Afin de motiver notre modèle, considérons une plateforme de jeu vidéo en ligne où les joueurs peuvent créer avec d'autre des parties, plusieurs d'entre elles pouvant être lancées indépendamment par un sous-ensemble de joueurs (possiblement tous). Un joueur peut alors préférer jouer avec certains plutôt qu'avec d'autres. Ces préférences peuvent se fonder sur des expériences de jeu précédentes ou sur des liens sociaux préalables entre joueurs. La question est alors de déterminer quel joueur rejoindra quelle partie, en sachant qu'il n'est pas possible pour un joueur de rejoindre une partie déjà commencée. Notons qu'un joueur peut toujours jouer seul, même si cela est peut ne pas être satisfaisant.

Un tel problème de décision collective peut être modélisé par un jeu hédonique, chaque agent (joueur) exprimant des préférences sur les coalitions du système (parties) où une *coalition* est un ensemble d'agents avec qui il souhaite jouer.

Définition 1 *Un* jeu hédonique *(en abrégé* jeu*) est un couple $G = \langle N, \succeq \rangle$ où $N = (a_1, \ldots, a_n)$ est un ensemble d'*agents *et $\succeq$ est un* profil de préférence *$(\succeq_1, \ldots, \succeq_n)$ qui associe à chaque agent a_i une relation de préférence $\succeq_i$ sur les sous-ensembles de N contenant a_i (coalitions).*

Comme précisé dans notre exemple applicatif, la relation de préférence d'un agent peut être construite à partir d'une notion de confiance ou de réputation vis-à-vis des autres agents, ou encore à partir des résultats de jeux précédents. Dans notre cas, nous ne nous intéressons uniquement au jeu à un instant donnée et supposons donc que le profil de préférence existe, sans faire d'hypothèse sur son origine.

Définition 2 *Soit N un ensemble d'agents. Une* relation de préférence $\succeq$ *sur N est une relation réflexive et transitive sur un sous-ensemble de N. $\succ$ (resp. $\sim$) est une préférence stricte (resp. symétrique). Pour deux coalitions $C, C' \subseteq N$, $C \succ C'$ se lit "C est* préférée *à C'", et $C \sim C'$ est lu "l'agent est* indifférent *entre C et C'".*

Exemple 1 *Dans la suite de cet article, nous utiliserons l'exemple de la Figure 1 où, par simplification, nous omettons volontairement les indices sur les relations de préférence et écrivons $13m$ pour la coalition $\{h_1, h_3, m\}$. La notation h correspond à un agent "honnête", m à l'agent "malveillant" et s à un agent "Sybil" (une fausse identité de m). Ici, du point de vue de l'agent h_1, la coalition 12 est préférée à la coalition $13m$ qui est à son tour préférée à 13 et $12m$, h_1 étant indifférent à ces deux dernières coalitions.*

Résoudre un jeu hédonique consiste à trouver un ensemble de coalitions qui *satisfont* les préférences de tous les agents. Dans notre modèle, nous interdisons les coalitions recouvrantes, c'est-à-dire nous interdisons à un agent d'appartenir à plusieurs coalitions simultanément. Nous notons par $\Pi = \{C_1, \ldots, C_m\}$ une partition de N et par C_i^Π l'unique coalition dans Π contenant l'agent a_i. Enfin, nous adoptons le concept de solution fondé sur la stabilité selon Nash. Une partition Π est stable au sens de Nash si aucun agent ne désire (unilatéralement) rejoindre une autre coalition de cette partition.

Définition 3 *Soit $G = \langle N, \succeq \rangle$ un jeu hédonique. Une partition Π de N est dite* stable au sens de Nash *si :*

$$\forall a_i \in N, \nexists C \in \Pi \cup \{\emptyset\}, C \cup \{a_i\} \succ_i C_i^\Pi$$

Exemple 2 *La ligne "NS_G" de la Figure 1 indique les deux seules partitions stables au sens de Nash du jeu.*

En général, un jeu G peut avoir zéro, une ou plusieurs partitions stables au sens de Nash. Même si ce concept de solution peut sembler restrictif, il nous permet de modéliser les cas où aucun agent ne donne son accord à une solution. Dans notre exemple applicatif, cela arrive lorsque chaque joueur désire jouer avec un autre qui, lui, ne le désire pas.

Dans la suite, nous désignons par *partition stable* une partition stable au sens de Nash et écrivons NS_G pour noter l'ensemble des partitions stables dans un jeu G. Notons que dans un jeu hédonique, tout agent peut décider d'être dans la coalition singleton, ce qui signifie simplement qu'il ne collabore pas avec les autres agents. Ainsi, pour un agent a_i, les coalitions qui ne sont pas préférées à la coalition singleton $\{a_i\}$ ne peuvent pas appartenir à une partition stable. Afin de simplifier la lecture de cet article, nous ne représentons dans $\succeq_i$ que les coalitions préférées (ou également préférées) à la coalition singleton.

3.2 Attaque Sybil

Un agent effectue une attaque Sybil [13] en entrant dans le système sous de multiples fausses identités. Par exemple, dans notre exemple applicatif, un joueur peut pour cela créer différents comptes. Dans cet article, nous supposons qu'il n'y a qu'un unique agent malveillant m qui essaye de manipuler le jeu hédonique en créant un nombre arbitraire de fausses identités. Nous ne faisons alors aucune hypothèse sur la connaissance que m a du jeu.

Intuitivement, l'agent malveillant manipule un jeu par une attaque Sybil en rapportant, d'une part, de fausses préférences pour lui-même et, d'autre part, en introduisant de fausses identités ayant leurs propres relations de préférence dans le jeu.

Définition 4 *Soit $G = \langle N, \succeq \rangle$ un jeu hédonique et $m \in N$ un agent. Une* attaque Sybil *sur G effectuée par m est définie par un ensemble de nouveaux agents $\{s_1, \ldots, s_k\}$ ($k \geq 0$) appelés* agents Sybil*, une relation de préférence $\succeq'_m$ pour m et une relation de préférence $\succeq'_{s_i}$ pour chaque agent Sybil s_i.*

Remarquons que cette définition est générale dans le sens où aucune hypothèse n'est faite sur le nombre d'agents Sybil, ni sur les connaissances de m vis-à-vis du jeu. En effet, l'agent

h_1	$12 \succ 13m \succ 13 \sim 12m \succ 123m \sim 1 \succ 1m \sim 123$
h_2	$12 \sim 23m \succ 123 \sim 2m \succ 12m \succ 23 \succ 123m \sim 2$
h_3	$13 \sim 23m \succ 3m \succ 123 \succ 23 \succ 123m \sim 3 \succ 13m$
m	$1m \succ 2m \succ 3m \succ m \succ 12m \succ 13m \sim 23m \succ 123m$

NS_G	$\Pi_1 = \{12, 3m\}, \Pi_2 = \{13, 2m\}$
UR_G	$\Pi_3 = \{1, 23m\}, \Pi_4 = \{12m, 3\}, \Pi_5 = \{123m\}$

FIGURE 1 – Exemple de jeu hédonique avec 4 agents.

malveillant peut ne pas connaître le nombre d'agents du jeu ou, à l'inverse, avoir une connaissance parfaite des préférences de ces derniers. De même, nous pouvons remarquer que si l'agent malveillant m n'utilise aucune fausse identité, l'attaque consiste alors pour m à simplement mentir sur ses préférences.

Comme de nouveaux agents sont introduits par la manipulation, nous devons définir la relation de préférence $\succeq'_i$ de chaque autre agent h_i vis-à-vis des coalitions contenant ces nouveaux agents. Pour cela, nous posons deux hypothèses. La première est l'indépendence aux alternatives non pertinentes [1] couramment utilisée dans les systèmes de votes. Dans notre contexte, cela signifie que si un agent préfère une coalition C_1 à une coalition C_2, alors l'introduction d'un nouvel agent dans le jeu ne changera pas cet ordre. La seconde hypothèse, que nous appelons le bénéfice du doute, modélise le fait que les agents sont *a priori* indifférents à former une coalition avec un agent qu'ils ne connaissent pas.

Hypothèse 1 (Indépendance aux alternatives)

$$\forall C_1, C_2 \subseteq N, \forall a_i \in C_1 \cap C_2 : \\ C_1 \succeq_i C_2 \Leftrightarrow C_1 \succeq'_i C_2$$

Hypothèse 2 (Bénéfice du doute)

$$\forall C \subseteq N, \forall a_i \in C, \forall u \notin N, \quad C \sim'_i C \cup \{u\}$$

Exemple 3 *Si s intègre le jeu de la Figure 1 alors $\succeq'_1$ satisfait* $12 \sim'_1 12s \succ'_1 13m \sim'_1 13ms \succ'_1 13 \sim'_1 13s \sim'_1 12m \ldots$

L'hypothèse 1 est une hypothèse de bon sens. L'hypothèse 2 est, quant à elle, nécessaire pour modéliser des systèmes ouverts en permettant aux nouveaux agents de coopérer avec des agents déjà présents. En effet, [18] énonce cette propriété désirable dans le cadre des systèmes de réputation : *les nouveaux entrants ne doivent pas être pénalisés en ayant par défaut une faible valeur de réputation*. Dans notre exemple applicatif, il est raisonable de supposer que les joueurs sont indifférents à la présence de nouveaux joueurs car cela peut permettre de lier de nouvelles amitiés ou affronter de nouveaux adversaires.

Ainsi, les hypothèses 1 et 2 permettent de définir clairement les nouvelles relations de préférence des agents honnêtes lors de l'arrivée d'un nouvel agent. Lorsque plusieurs identités intègrent le jeu, les nouvelles relations de préférence sont déterminées par induction : le premier nouvel agent est introduit, puis le second sur les préférences qui en résultent, et ainsi de suite.

Définition 5 *Soit $G = \langle N, \succeq \rangle$ un jeu hédonique avec $N = \{h_1, \ldots, h_n, m\}$. Le* jeu résultant *d'une attaque Sybil $(\{s_1, \ldots, s_k\}, \succeq'_m, (\succeq'_{s_1}, \ldots, \succeq'_{s_k}))$ sur G par m est le jeu $G' = \langle N \cup \{s_1, \ldots s_k\}, (\succeq'_1, \ldots, \succeq'_n, \succeq'_m, \succeq'_{s_1}, \ldots, \succeq'_{s_k}) \rangle$, où $\succeq'_i$ est définie à partir de $\succeq_i$ et des hypothèses 1 et 2.*

3.3 Rationalité de l'attaquant

Nous considérons les agents malveillants comme rationnels dans le sens où ils n'effectueront une attaque que si et seulement si ils préfèrent la solution du jeu résultant de la manipulation à la solution du jeu initial. La définition formelle de cette notion de rationalité est donnée dans cette section.

Toutefois, nous supposons que les agents malveillants ne connaissent pas la manière dont la solution est déterminée si ce n'est que cette solution sera nécessairement stable au sens de Nash. Elle peut par exemple être déterminée par un protocole de négociation ou un tirage aléatoire. Nous représentons alors cette ignorance par l'hypothèse ci-dessous.

Hypothèse 3 *Le solution d'un jeu G est selectionnée aléatoirement uniformément parmi NS_G.*

Cette hypothèse nous permet de définir de manière générale l'objectif d'un agent malveillant. Intuitivement, une manipulation est *efficace* si elle augmente la proportion de partitions stables *satisfaisantes*. La satisfaction de l'agent malveillant se définit alors relativement à une *coalition seuil* C_θ qui la coalition minimalement préférée dont m souhaite être membre. C_θ est alors choisie par m et lui permet de comparer deux manipulations distinctes.

Définition 6 *Soit $G = \langle N, \succeq \rangle$ un jeu hédonique. Une partition Π de N est dite* satisfaisante *pour m relativement à la coalition seuil C_θ si et seulement si $\Pi \in NS_G$ et $C_m^\Pi \succeq_m C_\theta$.*

Exemple 4 *Sur la Figure 1, il n'y a pas de partition satisfaisante pour m si $C_\theta = 1m$. En revanche, si m choisit $C_\theta = 3m$ alors les deux partitions stables (ligne "NS_G") sont satisfaisantes.*

Dans un jeu résultant d'une manipulation, m est présent à la fois sous sa véritable identité et sous celles de ces fausses identités $s_1, \ldots, s_k$. Intuitivement, si m désire rejoindre une coalition C, il sera également satisfait si une de ses fausses identités la rejoint à sa place. Ainsi, nous pouvons redéfinir la notion de partition satisfaisante pour un jeu résultant d'une manipulation.

Définition 7 *Soit $G' = \langle N \cup \{s_1, \ldots, s_k\}, \succeq' \rangle$ un jeu hédonique résultant d'une manipulation de m. Une partition Π' est dite* satisfaisante *relativement à la coalition seuil C_θ si $\Pi' \in NS_{G'}$ et que $C_m^{\Pi'} \succeq_m C_\theta$ ou que $\exists s_i \in \{s_1, \ldots, s_k\}, C_{s_i}^{\Pi'} \cup \{m\} \setminus \{s_i\} \succeq_m C_\theta$.*

Nous insistons ici sur le fait que la satisfaction de m dans le jeu G' résultant de la manipulation est définie à partir de ses *préférences initiales* $\succeq_m$ (données dans G et donc portant sur N). En particulier, une coalition contenant plusieurs des identités de m ne peut pas être satisfaisante. Ceci modélise le fait que l'agent malveillant ne peut pas concrétement agir simultanément sous plusieurs fausses identités. Par conséquent, une fois la solution du jeu déterminée, une seule des identités de m peut participer aux coalitions, les autres devant être abandonnées. Cet abandon se fait après la formation effective des coalitions et donc n'affecte pas la solution elle-même. Dans notre exemple applicatif, cette hypothèse est réaliste car un joueur peut à tout moment quitter une partie mais ne peut pas en rejoindre une autre déjà commencée. Plus généralement, une identité peut toujours quitter une coalition en simulant une défaillance critique.

Comme la solution d'un jeu G est choisie uniformément parmi l'ensemble des partitions stables, nous définissons l'efficacité d'une manipulation en fonction de la proportion de partitions satisfaisantes pour m qu'elle génère.

Définition 8 *Soit G un jeu hédonique, m un agent malveillant et G' le jeu résultant d'une manipulation de m sur G. Soit C_θ la coalition seuil choisie par m. Une manipulation est* efficace *relativement à C_θ si $r_\theta^{G'} > r_\theta^G$ où r_θ^G est le ratio de partitions satisfaisantes pour m dans G :*

$$r_\theta^G = \frac{|\{\Pi | \Pi \textit{ est satisfaisante pour } m\}|}{|NS_G|}$$

Par convention, $r_\theta^G = 0$ lorsque $NS_G = \emptyset$. Remarquons que si C_θ est fixée comme étant la coalition singleton $\{m\}$ alors toutes les partitions stables sont satisfaisantes pour m. Ainsi, si le jeu initial possède au moins une partition stable alors r_θ^G vaut 1. Dans ce cas particulier, aucune manipulation ne peut donner un ratio strictement supérieur (et donc être efficace).

4 Attaque Sybil constructive

Nous présentons maintenant une première attaque Sybil. Cette manipulation est dite *constructive* dans le sens où l'agent malveillant manipule le jeu afin de rendre stables des partitions instables désirées.

4.1 Définitions

Pour toute partition instable Π d'un jeu G, les agents peuvent être séparés en deux groupes : ceux qui ne veulent pas changer de coalition et ceux qui le souhaitent. Ces derniers sont dit *responsables* de l'instabilité de Π.

Définition 9 *Soit G un jeu hédonique, a_i un agent et Π une partition instable de G. L'agent a_i est dit* responsable *de l'instabilité de Π si $\exists C \in \Pi$ telle que $C \cup \{a_i\} \succ_i C_i^{\Pi}$. Une telle coalition est dite* attractive *(pour a_i).*

Exemple 5 *La ligne "UR_G" de la Figure 1 montre l'ensemble des partitions instables dont l'agent m est l'unique responsable.*

Intuitivement, l'attaque constructive est efficace si l'agent malveillant m est l'unique responsable de l'instabilité d'une partition Π et que la coalition attractive pour m dans Π est également satisfaisante si m la rejoint. Informellement, m manipule alors le jeu en prétendant être indifférent à toutes les coalitions possibles et en introduisant une seule fausse identité qui rapportera ses véritables préférences afin de profiter du bénéfice du doute (Hypothèse 2).

De manière à simplifier la lecture, nous définissons par l'*indifférence* de m (noté $\sim'_m$) la relation de préférence où m est indifférent à toute les coalitions qu'il peut rejoindre ($C_1 \sim_m C_2$ pour tout $C_1, C_2 \ni m$). Nous notons aussi $\succeq_m [m/s]$ la relation de préférence obtenue en remplacant m par s dans toutes les coalitions de $\succeq_m$.

Définition 10 *Soit $G = \langle \{h_1, \ldots, h_n, m\}, \succeq \rangle$ un jeu hédonique. L'*attaque constructive *de G par m est la manipulation utilisant un agent Sybil s où m et s rapportent respectivement les relations de préférence suivantes : $\succeq'_m := \sim_m$ et $\succeq'_s := \succeq_m [m/s]$.*

Notons que l'agent Sybil indique qu'il ne *désire pas* rejoindre m (car m est remplacé par s in $\succeq'_s$).

Exemple 6 *La manipulation constructive du jeu de la Figure 1 consiste à ajouter s avec pour relation de préférence $1s \succ'_s 2s \succ'_s 3s \succ'_s s$.*

4.2 Efficacité

Nous présentons ici sous quelles conditions l'attaque constructive est efficace. Dans les lemmes suivants, nous caractérisons les conditions nécessaires pour qu'un agent honnête ou un agent Sybil souhaite changer de coalition dans une partition du jeu résultant de la manipulation. Trivialement, l'agent malveillant ne changera jamais de coalition car il est indifférent à toutes les coalitions.

Fixons tout d'abord un jeu $G = \langle N, \succeq \rangle$, un agent malveillant $m \in N$ et une partition Π de G. Notons G' le jeu résultant de l'attaque constructive de G par m. Notons aussi $C_0 \in \Pi \cup \{\emptyset\}$ une coalition quelconque de Π et Π' la partition de G' obtenue par l'ajout de l'agent Sybil s dans C_0, c'est-à-dire $\Pi' = \Pi \setminus \{C_0\} \cup \{C_0 \cup \{s\}\}$. Afin de simplifier l'écriture, nous notons $\Pi' = \Pi[s \to C_0]$.

Lemme 1 *Un agent honnête h ne désire changer de coalition dans Π' que si et seulement si il le désire dans Π.*

Démonstration Par définition, h désire changer de coalition dans Π' si et seulement si $\exists C' \in \Pi'$ telle que $C' \cup \{h\} \succ'_h C_h^{\Pi'}$. En raison de l'Hypothèse 2, nous avons $C' \cup \{h\} \sim'_h C' \cup \{h\} \setminus \{s\}$ et $C_h^{\Pi'} \sim'_h C_h^{\Pi'} \setminus \{s\}$. Ainsi, $C' \cup \{h\} \succ'_h C_h^{\Pi'}$ est équivalent à $C' \cup \{h\} \setminus \{s\} \succ'_h C_h^{\Pi'} \setminus \{s\}$. Cependant, comme $C_h^{\Pi'} \setminus \{s\} = C_h^{\Pi}$ et que $C' \setminus \{s\}$ est définie dans Π, nous pouvons dire que h ne souhaite rejoindre C' dans Π' que si il souhaite rejoindre $C' \setminus \{s\}$ dans Π. □

Lemme 2 *L'agent Sybil s ne désire changer de coalition dans Π' que si et seulement si $m \in C_0$ ou que $\exists C \in \Pi$ telle que $m \notin C$ et $C \cup \{m\} \succ_m C_0 \cup \{m\}$.*

Démonstration Supposons tout d'abord $m \in C_0$. Dans ce cas, s souhaite nécessairement changer de partition dans Π' (au moins pour former la coalition $\{s\}$).
Supposons maintenant que $m \notin C_0$. Si s préfère une coalition $C' \in \Pi'$ à la coalition $C_0 \cup \{s\}$ alors il existe C telle que $m \notin C$ et $C \cup \{m\} \succ_m C_0 \cup \{m\}$. Comme s souhaite changer de coalition, nous avons $C' \cup \{s\} \succ'_s C_0 \cup \{s\}$. Or par définition de $\succeq'_s$, nous avons nécessairement $m \notin C'$ et $C' \cup \{m\} \succ_m C_0 \cup \{m\}$.
Inversement, si nous supposons qu'il existe une telle coalition C, par définition de $\succeq'_s$, s souhaite rejoindre la coalition $C \cup \{s\}$ dans Π'. □

Le corollaire suivant se déduit à partir des deux lemmes ci-dessus.

Corollaire 1 *Une partition Π' est stable dans G' si et seulement si, pour la partition Π telle*

$$NS_G^\theta = \{\Pi \in NS_G \mid \Pi \text{ est satisfaisante pour } m\}$$
$$UR_G = \{\Pi \notin NS_G \mid m \text{ est l'unique responsable de l'instabilité de } \Pi\}$$
$$UR_G^\theta = \{\Pi \in UR_G | \exists C \in \Pi, C \cup \{m\} \succeq_m C_\theta\}$$

FIGURE 2 – Partitions remarquables d'un jeu hédonique

que $\Pi' = \Pi[s \to C_0]$*, soit (1)* Π *est stable dans* G*, soit (2)* m *est l'unique responsable de l'instabilité de* Π *dans* G *et* $C_0 \cup \{m\}$ *est la coalition que* m *préfère à toutes les autres dans* Π*.*

Exemple 7 *Sur le jeu de la Figure 1,* $\Pi'_3 = \Pi_3[s \to 1] = \{1s, 23m\}$ *est une partition satisfaisante dans le jeu résultant de la manipulation constructive si* $C_\theta = 1m$ *tandis que la partition* $\Pi'_1 = \Pi_1[s \to 12] = \{12s, 3m\}$ *est stable mais non satisfaisante.*

Nous pouvons maintenant caractériser les conditions nécessaires à l'efficacité de l'attaque constructive sur un jeu hédonique G. Pour une coalition seuil C_θ donnée, consiérons les partitions NS_G^θ, UR_G et UR_G^θ dont les définitions sont données en Figure 2.

Exemple 8 *Sur la Figure 1, pour* $C_\theta = 1m$ *ou* $C_\theta = 2m$*,* UR_G^θ *est* $\{\Pi_3\}$ *et, pour* $C_\theta = 3m$*,* UR_G^θ *est* $\{\Pi_3, \Pi_4\}$*.*

Par souci de lisibilité, nous supposons ici que l'agent malveillant possède des préférences linéaires. Toutefois, le résultat suivant peut être généralisé en comptant le nombre de coalitions satisfaisantes présentes dans les partitions de $\Pi \in NS_G^\theta \cup UR_G^\theta$ au lieu de ne compter que 1 par partition.

Proposition 1 *Supposons que* m *rapporte une relation de préférence linéaire (asymétrique et totale). L'attaque constructive est efficace sur un jeu* G *si et seulement si soit (1)* $NS_G^\theta = \emptyset$ *et* $UR_G^\theta \neq \emptyset$*, soit (2)* $NS_G^\theta \neq \emptyset$ *et* $\frac{|UR_G^\theta|}{|UR_G|} > \frac{|NS_G^\theta|}{|NS_G|}$*.*

Démonstration Supposons tout d'abord $NS_G^\theta = \emptyset$. La manipulation est alors efficace si et seulement si G' possède au moins une partition satisfaisante Π'. Notons $\Pi' = \Pi[s \to C_0]$. Selon le Corollaire 1, ceci n'est possible que si $\Pi \in NS_G$ ou que $\Pi \in UR_G$. Comme Π' est satisfaisante, soit $C_0 \cup \{m\} \succeq_m C_\theta$, soit $C_m^\Pi \succeq_m C_\theta$. Dans les deux cas, comme $NS_G^\theta = \emptyset$, nous avons $\Pi \notin NS_G$, ce qui implique que $\Pi \in UR_G$ et donc que $\Pi \in UR_G^\theta$.

Supposons maintenant $NS_G^\theta \neq \emptyset$. Comparons les ratios de partitions satisfaisantes dans G et dans G'. Ainsi, $r_\theta^G = |NS_G^\theta|/|NS_G|$. Selon le Corollaire 1 et par linéarité de $\succeq_m$, nous avons $|NS_{G'}| = |NS_G| + |UR_G|$.

Comptons désormais combien de partitions sont satisfaisantes dans $NS_{G'}$. Considérons dans un premier temps les partitions stables $\Pi \in NS_G$. Pour chaque partition $\Pi \in NS_G$, il y a exactement une partition stable Π' (dans G') de la forme $\Pi[s \to C_0]$. Comme Π est stable, $C_0 \cup \{m\} \not\succeq_m C_m^\Pi$. Par conséquent, Π' est satisfaisante dans G' si et seulement si Π est satisfaisante dans G. Considérons maintenant les partitions instables. Selon le Corollaire 1, pour chaque partition $\Pi \notin NS_G$, la partition $\Pi' = \Pi[s \to C_0]$ n'est stable dans G' que si m est l'unique responsable de l'instabilité de Π dans G, et que $C_0 \cup \{m\}$ est la coalition que m préfère à toutes les autres dans Π. Ainsi, Π' est satisfaisante dans G' si et seulement si $C_0 \cup \{m\} \succeq C_\theta$, *i.e.*, $\Pi \in UR_G^\theta$.

Enfin, le nombre de partitions satisfaisantes dans G' est $|NS_G^\theta| + |UR_G^\theta|$. Ainsi, la manipulation n'est efficace que si et seulement si $\frac{|NS_G^\theta|+|UR_G^\theta|}{|NS_G|+|UR_G|} > \frac{|NS_G^\theta|}{|NS_G|}$, i.e. $\frac{|UR_G^\theta|}{|UR_G|} > \frac{|NS_G^\theta|}{|NS_G|}$. □

Remarque 1 *La condition (1) ne nécessite pas la linéarité de* $\succeq_m$*. Comme nous le montrerons dans la section 6 la condition (2) n'a pas de réelle importance en pratique. En effet, lorsque* $0 < |NS_G^\theta| < |NS_G|$*, l'attaque destructive (présentée dans la section 5) est aussi efficace.*

Exemple 9 *Sur la Figure 1, l'attaque constructive est efficace (*$NS_G^\theta = \emptyset$ *et* $UR_G^\theta = \{\Pi_3\}$*) si* $C_\theta = 1m$*. Cependant, si* $C_\theta = 2m$*, la solution est strictement pire pour* m *(*$NS_G^\theta = \{\Pi_2\}$ *et* $UR_G^\theta = \{\Pi_3\}$*). En effet, dans ce cas,* $r_\theta^G = 1/2$ *et* $r_\theta^{G'} = 2/5$*.*

Il est intéressant de remarquer que décider si l'attaque constructive est efficace requière plus de connaissances sur le jeu initial que pour simplement construire cette attaque. En effet, il est nécessaire de connaître les préférences des autres agents pour calculer le nombre de partitions stables dans G. Nous concluons donc cette section en montrant qu'il s'agit d'un problème de décision difficile que m doit résoudre s'il ne souhaite pas risquer d'empirer sa situation en manipulant le jeu (comme illustré par l'Exemple 9).

Proposition 2 *Il est NP-difficile pour m de décider si l'attaque constructive est efficace sur G.*

Démonstration Nous allons ici donner une réduction au problème de décider si, pour un jeu donné G_0, il existe au moins une partition stable au sens de Nash. Ce problème a été prouvé NP-complet [3].

À partir de G_0, construisons un jeu G où $NS^{\theta}_{G} = \emptyset$ mais où $UR^{\theta}_{G} \neq \emptyset$ uniquement si G_0 possède une partition stable. Selon la Proposition 1, l'attaque constructive n'est efficace sur G que si G_0 possède une partition stable au sens de Nash.

Notons $G_0 = \langle N_0, \succeq_0 \rangle$ avec $N_0 = \{h_1, \ldots, h_n\}$. Le jeu G est défini à partir de G_0 en y ajoutant deux nouveaux agents, h et m, ayant les relations de préférences suivantes : $\{h, m\} \succ_m \{m\}$, $\{h\} \succ_h C$ pour toute coalition $C \neq \{h\}$. Pour tout agent $h_i \in N_0$, $\succeq_i$ est calculée à partir de $(\succeq_0)$ selon les Hypothèses 1 et 2.

Intuitivement, h souhaite être seul et m veut rejoindre h. Les autres agents sont indifférents à eux et conservent leurs préférences de G_0. Trivialement, G peut être construit en un temps polynomial en la taille de G_0. Fixons C_θ comme étant la coalition $\{h, m\}$.

Il n'existe alors pas de partition stable dans G puisque h souhaite être dans une coalition singleton et que m veut le rejoindre. Supposons qu'il existe une partition stable Π_0 dans G_0 et considérons la partition $\Pi = \Pi_0 \cup \{\{h\}, \{m\}\}$ dans G. Alors, m est l'unique responsable de l'instabilité de Π. Par ailleurs, la coalition attractive pour m est satisfaisante dans Π. Ainsi, $\Pi \in UR^{\theta}_{G}$. Par dualité, si toutes les partitions Π_0 de G_0 sont instables, comme $h_1, \ldots, h_n$ sont indifférents envers h, m, toute partition de G intégrant h et m est alors instable.

Ainsi, G n'a pas de partition stable et $UR^{\theta}_{G} \neq \emptyset$ si et seulement si G_0 a une partition stable. □

5 Attaque Sybil destructive

Nous présentons dans cette section une attaque Sybil destructive dans le sens où l'agent malveillant manipule le jeu afin de rendre instable une partition stable non désirée.

5.1 Définitions

Dans la définition de la stabilité au sens de Nash, un unique agent "veto" peut refuser une coalition et ainsi rendre une partition donnée instable. L'attaque destructive repose sur ce constat et utilise une unique fausse identité qui posera un veto sur toutes les partitions qui ne satisfont pas m.

Définition 11 *Soit $G = \langle N, \succeq \rangle$ un jeu hédonique. L'*attaque destructive *de m sur G est la manipulation utilisant un agent Sybil s où m et s rapportent respectivement les relations de préférence suivantes : $\succeq'_m := \succeq_m$ et $\succeq'_s$ définie pour toute coalition $C \subseteq N$ par (1) $C \cup \{s\} \succ'_s \{s\}$ si $m \in C$ et $C \not\succeq_m C_\theta$, et par (2) $\{s\} \succ'_s C \cup \{s\}$ dans les autres cas.*

Notons que s rapporte $\{s\} \succ'_s C_\theta \cup \{s\}$ et que m refuse d'être dans la même coalition que s.

Exemple 10 *Sur la Figure 1, si $C_\theta = 2m$, la relation de préférence de s lors d'une attaque destructive est $3ms$, ms, $12ms$, $13ms$, $23ms$, $123ms \succ'_s s$.*

5.2 Efficacité

Nous montrons maintenant que l'attaque destructive ne peut être efficace sur un jeu hédonique que s'il existe au moins une partition satisfaisante dans le jeu initial. Pour cela, fixons un jeu G où un agent malveillant m choisit une coalition seuil C_θ. Notons G' le jeu résultant d'une attaque destructive de m sur G.

Lemme 3 *Il existe une partition stable dans G' si et seulement si il existe au moins une partition satisfaisante pour m dans G. De plus, toute partition stable de G' est satisfaisante pour m.*

Démonstration Supposons tout d'abord l'existence d'une partition Π satisfaisante dans G. Dans ce cas, $\Pi \cup \{\{s\}\}$ est nécessairement satisfaisante dans G'. Supposons maintenant qu'une partition Π' soit satisfaisante pour m dans G'. Prenons la partition Π de G telle que $\Pi' = \Pi[s \to C_0]$. Si m est satisfait par Π' dans G', il est nécessairement satisfait par Π dans G. Dans le cas contraire, Π' ne pourrait être satisfaisante que par la présence de $C_0 \cup \{s\}$. Or, selon la définition de $\succeq'_s$, s souhaiterait changer de coalition pour rejoindre la coalition de m, ce qui contredit la stabilité de Π'.

Montrons maintenant que toutes les partitions stables de G' sont satisfaisantes pour m. Supposons qu'il existe une partition stable Π' dans G' mais non satisfaisante pour m. Selon la définition de $\succeq'_s$, s doit être dans la même coalition que m. Cependant, m préférerait alors être dans sa coalition singleton, ce qui contredit la stabilité de Π'. □

Il est intéressant de noter que lorsque l'attaque Sybil destructive est efficace, elle l'est *totalement* dans le sens où toutes les partitions stables du jeu résultant de la manipulation sont satisfaisantes. Les conditions nécessaires à son efficacité sont alors les suivantes :

Proposition 3 *L'attaque destructive est efficace sur un jeu hédonique G si et seulement si il existe dans G au moins une partition satisfaisante et au moins une partition stable mais non satisfaisante.*

Dans cette proposition, la présence d'une partition stable mais non satisfaisante dans G est nécessaire par le fait que si toutes les partitions d'un jeu G sont satisfaisantes, alors il ne peut y avoir de manipulation donnant un ratio de partitions satisfaisantes — strictement — supérieur.

Exemple 11 *Sur la Figure 1, l'attaque destructive est efficace pour la coalition seuil $C_\theta = 2m$ (Π_2 est satisfaisante, Π_1 ne l'est pas et donc seule $\Pi_2[s \to \emptyset] = \{13, 2m, s\}$ est stable et satisfaisante dans G'). En revanche, pour $C_\theta = 3m$, l'attaque destructive n'est pas efficace (m est déjà totalement satisfait dans G), tout comme pour $C_\theta = 1m$ ($r_\theta^G = r_\theta^{G'} = 0$).*

Comme pour l'attaque constructive, il est, d'une part, difficile de décider si cette manipulation est efficace et, d'autre part, cela nécessite de connaître les préférences des autres agents. Cela est toutefois moins problématique pour m car cette attaque, au pire cas, ne change pas sa situation.

Proposition 4 *Il est NP-difficile pour m de décider si une attaque destructive est efficace sur G.*

Démonstration La démonstration est similaire à celle de la Proposition 2. Soit $G_0 = \langle N_0, \succeq_0 \rangle$ un jeu hédonique donné. Nous pouvons construire à partir de G_0 un jeu G possédant à la fois une partition stable et satisfaisante et une partition stable mais non satisfaisante uniquement si G_0 possède au moins une partition stable.

Le jeu G est défini à partir de G_0 en y intégrant trois nouveaux agents h, h', et m dont les relations de préférence sont $\{h, h', m\} \succ_a \{a\}$ pour $a \in \{h, h', m\}$. Les préférences des agents de N_0 sont définies dans G à partir de $\succeq_0$ et des Hypothèses 1 et 2. Intuitivement, h, h' et m souhaitent tous être membres $\{h, h', m\}$ ou être dans la coalition singleton. Les autres agents restent indifférents vis-à-vis d'eux. Fixons alors la coalition seuil C_θ comme étant $\{h, h', m\}$.

Soit une partition Π de G. Si au moins un des trois nouveaux agents est en coalition avec un agent de N_0 alors Π n'est pas stable puisque cet agent préférera rejoindre sa coalition singleton. Π ne peut pas être stable non plus si exactement deux des nouveaux agents sont en coalition. Les deux cas restants sont ceux où h, h', m sont respectivement dans leur coalition singleton ou tous ensemble. Π ne peut alors être stable que si $\Pi \setminus \{\{h\}, \{h'\}, \{m\}, \{h, h', m\}\}$ est stable dans G_0. Notons par ailleurs que les deux partitions sont alors stables mais que seule celle incluant la coalition $\{h, h', m\}$ est satisfaisante pour m, comme désiré. □

6 Jeux manipulables

Dans cette section, nous montrons enfin que les jeux hédoniques ne sont manipulables dans le cas général que si les conditions caractérisées précédemment sont satisfaites, dans le sens où si une attaque Sybil quelconque est efficace sur un jeu G, alors l'attaque constructive ou l'attaque destructive l'est également. Ce résultat est inattendu puisque ces deux attaques ne nécessitent

qu'une unique fausse identité et peuvent être construites sans connaissance sur le jeu (y compris le nombre d'agents présents). Ainsi, utiliser un large nombre de fausses identités et/ou avoir des connaissances sur le jeu n'aide nullement un agent malveillant dans la construction d'une attaque Sybil.

Proposition 5 *Soit G un jeu hédonique, m un agent malveillant dans G et C_θ une coalition seuil choisie par m. Si une attaque Sybil quelconque M de m sur G est efficace, alors l'attaque constructive ou l'attaque destructive de m sur G l'est aussi.*

Démonstration Supposons qu'il existe une manipulation M efficace sur un jeu G. Supposons que l'attaque destructive ne soit pas efficace sur G et montrons alors que l'attaque constructive est efficace.

Comme l'attaque destructive n'est pas efficace, la Proposition 3 indique que soit toutes les partitions stables dans G sont satisfaisantes, soit aucune ne l'est. Dans le premier cas, M ne peut pas être efficace, ce qui contredit l'hypothèse. Ainsi, G ne possède pas de partition satisfaisante pour m.

Notons G' le jeu résultant de la manipulation M sur G et S l'ensemble des agents Sybil utilisés par m afin d'effectuer M. Comme M est efficace, il existe une partition Π' dans G' qui est satisfaisante pour m. Notons Π la partition de G telle que $\Pi = \{C' \setminus S \mid C' \in \Pi'\}$. Montrons alors que soit Π est satisfaisante dans G (ce qui serait en contradiction avec le fait qu'il n'en existe pas), soit $\Pi \in UR^\theta_G$.

Remarquons tout d'abord qu'aucun agent honnête h_i ne souhaite changer de coalition dans Π. En effet, dans le cas contraire, selon l'hypothèse de bénéfice du doute (Hypothèse 2), h_i désirerait effectuer le même changement dans Π', ce qui est en contradiction avec la stabilité de Π'. Du point de vue de m, nous distinguons alors deux cas possibles.

Supposons qu'il n'existe pas dans Π' de coalition préférée par m à celle dont il est membre : $\forall C \in \Pi', C \setminus S \cup \{m\} \not\succeq_m C^{\Pi'}_m \setminus S$. Dans ce cas, la partition Π est stable puisque m ne souhaite pas changer de coalition. Par ailleurs, comme m est dans la coalition qu'il préfère dans Π' et que Π' est satisfaisante, Π est également satisfaisante. Ceci est en contradiction avec l'absence de partition satisfaisante dans G.

Ainsi, nous avons nécessairement une coalition $C \in \Pi'$ telle que $C \cup \{m\} \setminus S \succeq_m C^{\Pi'}_m \setminus S$. m souhaite alors rejoindre cette coalition C dans Π. De plus, commme Π' est satisfaisante pour m dans G', une telle coalition C doit être satisfaisante pour m. Ainsi, Π doit nécessairement être dans l'ensemble UR^θ_G. Comme G ne possède pas de partition satisfaisante ($NS^\theta_G = \emptyset$), l'attaque constructive est alors efficace sur G (Proposition 1 (1), voire également la Remarque 1). □

Remarque 2 *La démonstration montre que c'est seulement dans le cas où un jeu G ne possède aucune partition satisfaisante que l'attaque constructive peut être efficace sans que l'attaque destructive ne le soit.*

7 Étude empirique

Nous présentons dans cette section des résultats de simulations qui suggèrent que, même si certains jeux sont manipulables (c'est-à-dire qu'il est possible de les manipuler efficacement), la majorité des jeux hédoniques ne le sont pas en pratique. Nous considérons ici les attaques constructives, les attaques destructives et aussi qu'une troisième manipulation que nous nommons l'*attaque hybride*. Une attaque hybride consiste à appliquer successivement une attaque constructive puis une attaque destructive. Trivialement, une telle attaque ne peut être efficace que si l'attaque constructive l'est et que le jeu résultant ne possède pas un ratio de satisfaction de 1.

Afin d'obtenir une estimation de la probabilité qu'un jeu soit manipulable, nous avons effectué un ensemble d'expériences en faisant varier le nombre d'agents entre 3 et 10. Pour chaque expérience, nous avons effectué 10000 simulations, chacune consistant à manipuler un jeu G dont des relations de préférence des agents sont tirées aléatoirement de manière uniforme. La Figure 3) indique la proportion de ces jeux qui sont manipulables et la Figure 4 indique la proportion de ces mêmes jeux selon leur nombre de partitions stables initiales.

La Figure 3 suggère que plus il y a d'agents, plus la probabilité qu'un jeu quelconque soit manipulable est faible. Ce résultat est conforme à l'intuition puisque plus il y a d'agents, moins l'agent malveillant a de chances d'être l'unique responsable de l'instabilité d'une partition. La

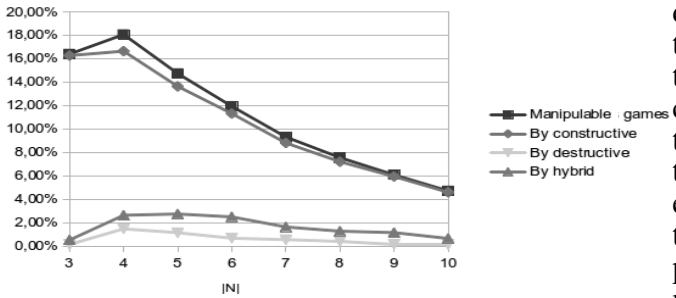

FIGURE 3 – Proportion de jeux manipulables en fonction du nombre d'agents

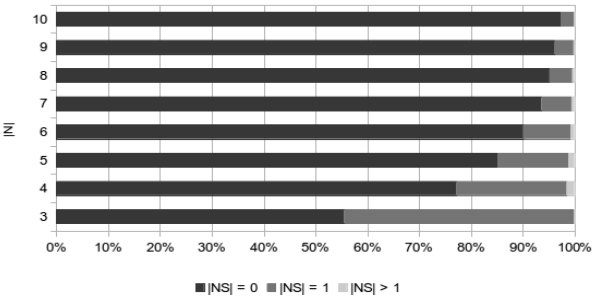

FIGURE 4 – Proportion de partitions stables en fonction du nombre d'agents

Figure 4, quant à elle, suggère que plus il y a d'agents, moins un jeu hédoniste possède de partitions stables selon le concept de solution de Nash. Ici aussi, ce résultat est conforme à l'intuition car augmenter le nombre d'agents augmente de fait le nombre d'agents "veto".

Sur la Figure 3, nous constatons que la proportion de jeux manipulables décroit rapidement. En effet, pour 6 agents, seulement 67 des 10000 jeux sont manipulables par l'attaque destructive. De plus, au-delà de 7 agents, moins de 10% des jeux sont manipulables quelle que soit la manipulation.

En raison, d'une part, de la difficulté pour un agent malveillant de décider si une manipulation donnée est efficace et, d'autre, du fait que les jeux sont en pratique rarement manipulables, nous concluons que les jeux hédoniques ayant des solutions stables selon le concept de solution de Nash sont robustes aux attaques Sybil.

8 Conclusion

Dans cet article, nous nous sommes intéressés à la robustesse des jeux hédoniques ayant des solutions stables selon le concept de solution de Nash face à un type général de manipulation, appelée attaque Sybil. Nous avons montré qu'un jeu n'est manipulable que s'il satisfait certaines conditions, et avons présenté deux attaques Sybil qui couvrent ces conditions. Ces attaques sont simples à construire dans le sens où elles ne nécessitent qu'une unique fausse identité et aucune connaissance sur le jeu (y compris le nombre d'agents présents). En revanche, nous avons montré qu'il est difficile pour l'agent malveillant de décider si elles sont efficaces. De plus, nous avons montré empiriquement que les conditions nécessaires à l'efficacité des attaques Sybil quelconques sont rarement présentes dans les jeux hédoniques. C'est pourquoi nous concluons que les jeux hédoniques ayant des solutions stables selon le concept de solution de Nash sont robustes aux attaques Sybil.

Nos résultats reposent sur deux hypothèses portant sur l'attitude des agents honnêtes vis-à-vis des agents qu'ils ne connaissent pas. Ces deux hypothèses, l'indépendance aux alternatives non pertinentes et le bénéfice du doute, peuvent sembler être à l'avantage des agents malveillants. Cependant, il est important de noter que relâcher l'une de ces hypothèses rendrait encore plus difficile toute manipulation. Ceci vient alors renforcer notre conclusion. Toutefois, il pourait être intéressant de considérer de nouvelles hypothèses comme celles où les agents honnêtes préfèrent collaborer avec les nouveaux entrants, ou au contraire les refusent. Intuitivement, mettre en œuvre une manipulation serait plus simple et efficace dans le premier cas, et plus difficile et moins efficace dans le second. Plus fondamentalement, nous envisagons d'étendre notre étude à d'autres concepts de solution tels que la stabilité individuelle ou la stabilité contractuelle [14]. Ces concepts de solution sont moins restictifs que la stabilité au sens de Nash et, par conséquent, les conditions nécessaires à la mise en œuvre de manipulations sur les jeux les utilisant pourraient être moins restrictives elles-aussi.

Références

[1] K. J. Arrow. *Social Choice and Individual Values*. Yale University Press, 1970.

[2] H. Aziz, Y. Bachrach, E. Elkind, and M. Paterson. False-name manipulations in weighted voting games. *Journal of Artificial Intelligence Research*, 40 :57–93, 2011.

[3] C. Ballester. NP-completeness in hedonic games. *Games and Economic Behavior*, 49(1) :1–30, 2004.

[4] S. Banerjee, H. Konishi, and T. Sönmez. Core in a simple coalition formation game. *Social Choice and Welfare*, 18 :135–153, 2001.

[5] J. J. Bartholdi, C. A. Tovey, and M. A. Trick. The computational difficulty of manipulating an election. *Social Choice and Welfare*, 6(3) :227–241, 1989.

[6] G. Bonnet. 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*, pages 187–191, 2012.

[7] P. Caillou, S. Aknine, and S. Pinson. A multi-agent method for forming and dynamic restructuring of pareto-optimalcoalitions. In *Proceedings of the 1st International Joint Conference on Autonomous Agents and Multi-Agent Systems*, pages 1074–1081, 2002.

[8] V. Conitzer, N. Immorlica, J. Letchford, K. Munagala, and L. Wagman. False-name-proofness in social networks. *Lecture Notes in Computer Science*, 6484 :209–221, 2010.

[9] V. Conitzer and M. Yokoo. Using mechanism design to prevent false-name manipulations. *Artificial Intelligence Magazine*, 31(4) :65–78, 2010.

[10] G. Danezis and P. Mittal. SybilInfer : detecting Sybil nodes using social networks. In *Proceedings of the 16th Annual Network and Distributed System Security Conference*, 2009.

[11] F. Delaplace and P. Lescanne. HedN game, a relational framework for network based cooperation. In *Proceedings of the 7th European Conference on Complex Systems*, 2010.

[12] D. Dimitrov and S. C. Sung. Enemies and friends in hedonic games : individual deviations, stability and manipulation. Technical report, Tilburg University, 2004.

[13] J. R. Douceur. The Sybil attack. In *Proceedings of the 1st International Workshop on Peer-to-Peer Systems*, pages 251–260, 2002.

[14] E. Elkind and M. Wooldridge. Hedonic coalition nets. In *Proceedings of the 8th International Conference on Autonomous Agents and Multi-Agent Systems*, pages 417–424, 2009.

[15] Edith Elkind, Piotr Faliszewski, and Arkadii Slinko. Cloning in elections : Finding the possible winners. *Journal of Artificial Intelligence Research*, 42(1) :529–573, 2011.

[16] T. Génin and S. Aknine. Coalition formation strategies for self-interested agents in hedonic games. In *Proceedings of the 19th European Conference on Artificial Intelligence*, pages 1015–1016, 2010.

[17] K. Hoffman, D. Zage, and C. Nita-Rotaru. A survey of attack and defense techniques for reputation systems. *ACM Computer Survey*, 42(1) :1–31, 2009.

[18] S.D. Ramchurn, T.D. Huynh, and N. R. Jennings. Trust in multiagent systems. *The Knowledge Engineering Review*, 19 :1–25, 2004.

[19] H. Rowaihy, W. Enck, P. D. McDaniel, and T. La Porta. Limiting Sybil attacks in structured P2P networks. In *Proceedings of the 26th IEEE International Conference on Computer Communications*, pages 2596–2600, 2007.

[20] Taiki Todo and Vincent Conitzer. False-name-proof matching. In *Proceedings of the 12th International Joint Conference on Autonomous Agents and Multi Agent Systems*, 2013.

[21] T. Walsh. Where are the hard manipulation problems ? *Journal of Artificial Intelligence Research*, 42 :1–29, 2011.

[22] M. Yokoo, Y. Sakurai, and S. Matsubara. The effect of false-name bids in combinatorial auctions : new fraud in Internet auctions. *Game and Economic Behavior*, Vol. 46 :174–188, 2004.

[23] H. Yu, M. Kaminsky, P. B. Gibbons, and A. Flaxman. SybilGuard : defending against Sybil attacks via social networks. *SIGCOMM Computer Communication Review*, 36(4) :267–278, 2006.