Un protocole fondé sur un dilemme pour se prémunir des collusions dans les systèmes de réputation

Grégory Bonnet
gregory.bonnet@unicaen.fr

Université de Caen Basse-Normandie
UMR CNRS 6072 GREYC

Résumé

Dans les systèmes ouverts et décentralisés, un grand nombre d'agents interagissent afin de partager des ressources. Afin de se protéger de potentiels agents malveillants, des systèmes de réputation sont mis en place. Ils évaluent le comportement des agents a posteriori mais, comme ils se fondent sur une agrégation de confiances locales, ils sont eux-mêmes vulnérables aux manipulations, et en particulier aux coalitions malveillantes qui font de l'auto-promotion. Dans cet article, nous proposons une approche fondée sur la théorie des jeux pour prévenir de telles manipulations. Sa caractéristique principale est de permettre aux agents honnêtes d'utiliser à leur tour une manipulation, appelée attaque Sybil, pour tromper les agents malveillants et les placer face à un dilemme. Nous montrons théoriquement et empiriquement que la meilleure réponse à ce dilemme est une stratégie en équilibre mixte qui conduit les agents malveillants à abandonner la plupart de leurs manipulations.

Mots-clés : *auto-promotion, confiance et réputation, protocoles, théorie des jeux*

Abstract

In decentralized and open systems, a large number of agents interact in order to share resources. In order to prevent malicious behaviors, reputation systems are considered. Those systems evaluates the behavior of the agents afterwards. But, as reputation systems are based on the aggregation of local trust between the agents, they are vulnerable to manipulations, particularly to self-promotion among malicious coalitions. In this paper, we propose a game-theoretic approach to prevent such manipulations. Its main feature is that honest agents use in turn a manipulation, called Sybil attack, to fool malicious agents and to drive them into a dilemma. We show both theoretically and empirically that the best response to this dilemma is mixed strategy that leads the malicious agents to give up most of their manipulations.

Keywords: *self-promotion, trust and reputation, protocols, game theory*

1 Introduction

Les systèmes multi-agents sont composés d'un grand nombre d'agents qui interagissent entre eux et prennent des décisions, collectives ou non, dans le but de partager des ressources et d'assurer un service. Ces ressources peuvent être des compétences, de la puissance de calcul ou bien de la mémoire de masse. De même, les services peuvent prendre diverses formes comme des services Web, des calculs répartis ou bien des transactions financières. Afin d'assurer le fonctionnement nominal de tels systèmes lorsqu'ils sont décentralisés et ouverts, il est fait l'hypothèse que les agents, pouvant être altruistes, coopératifs, voire compétitifs, vont suivre certaines règles définies par un protocole ou des normes. Ceci revient à faire l'hypothèse que les agents agissent honnêtement vis-à-vis du système. Toutefois, comme ces systèmes sont ouverts, des agents malveillants peuvent y participer et détourner les règles du système afin d'en tirer parti, voire de provoquer une rupture du service lui-même. Afin de lutter contre de tels agents malveillants, il a été proposé d'utiliser des systèmes de réputation qui permettent aux agents de modéliser les interactions qu'ils observent et de décider s'il est *a priori* acceptable d'interagir avec un agent donné. Cette notion d'acceptation (ou confiance) signifie que l'agent tiers est considéré comme fiable. Cependant, si les systèmes de réputation sont efficaces pour détecter un unique agent non fiable, ils peuvent être mis en défaut par une coalition d'agents malveillants [14].

En effet, les systèmes de réputation sont fondés sur l'agrégation de valeurs de confiance locales et subjectives entre les agents en une valeur globale de réputation. En plus des problématiques liées à l'agrégation d'opinions, un ensemble d'agents malveillants peut rapporter

les uns pour les autres de hautes valeurs de confiance afin d'accroître artificiellement leurs valeurs de réputation. De telles manipulations sont appelées auto-promotions et peuvent être, par exemple, utilisées pour tromper le système de réputation d'eBay [12], l'algorithme PageRank de Google [7], voire même pour resquiller sur un réseau pair-à-pair [21], c'est-à-dire consommer des ressources sans en fournir en retour. À l'inverse, un ensemble d'agents malveillants peut rapporter collectivement une faible valeur de confiance envers un agent tiers afin de diminuer artificiellement sa valeur de réputation. Une telle manipulation est appelée diffamation et peut être utilisée conjointement à l'auto-promotion pour en accroître les effets. De plus, dans tout système ouvert, un agent peut se présenter sous de multiples fausses identités et ainsi construire une coalition virtuelle capable de manipuler le système de réputation. Cette manipulation est appelée une attaque Sybil [13]. Ainsi, se prémunir des coalitions malveillantes est critique pour les systèmes de réputation [16].

De nombreuses propositions ont été faites dans la litérature pour se protéger de ces coalitions, depuis l'utilisation de puzzles cryptographiques jusqu'à la détection de communautés dans le graphe d'accointance des agents, en passant par la définition de fonctions de réputation robustes. Des approches récentes suggèrent de considérer les agents malveillants comme rationnels et d'utiliser par-dessus le système de réputation des mécanismes fondés sur la théorie des jeux incitant les agents malveillants à se comporter honnêtement [10]. L'intérêt est alors d'éviter de faire des hypothèses sur les ressources des agents, la structure du graphe d'accointance ou la fonction de réputation elle-même.

Cet article est une version corrigée d'un article précédent [3]. Nous proposons un approche fondée sur la théorie des jeux pour empêcher les manipulations particulières que sont les auto-promotions de se produire dans un système de réputation. La spécificité de notre approche repose sur l'utilisation par les agents honnêtes d'attaques Sybil afin de tromper les agents malveillants et les prendre à leur propre jeu. Les agents malveillants sont alors placés face à un dilemme qui, s'ils sont rationnels, va les pousser à abandonner une grande partie de leurs manipulations. Cet article est organisé comme suit. Nous présenterons l'état de l'art en section 2. La section 3 sera consacrée à la description du protocole que nous analyserons ensuite formellement dans la section 4. Enfin, nous présenterons en section 5 les points forts et limites de l'approche à partir de résultats de simulations.

2 État de l'art

Dans la littérature concernant la confiance, les fonctions de réputation peuvent être symétriques, signifiant que chaque agent du système contribue au calcul de réputation, ou assymétrique, signifiant que les valeurs de confiance locales sont propagées au travers d'agents considérés comme *a priori* de confiance. [6] a montré que les fonctions de réputation symétriques ne pouvaient pas prémunir des manipulations tandis que les fonctions assymétriques le pouvaient à condition de satisfaire des propriétés restrictives qui les rendent faiblement informatives et délicates à définir. De plus, si des fonctions de réputation permettant de réduire la confiance dans des agents témoins non fiables existent, elles restent vulnérables au blanchiment, dans le sens où les agents malveillants peuvent toujours changer d'identité afin d'obtenir la valeur de confiance par défaut du système. C'est pourquoi des approches qui font abstraction du système de réputation ont été proposées [16].

Une première classe d'approches consiste à empêcher la mise en place des attaques Sybil qui facilitent les manipulations. De telles approches, comme celles de [5] ou [13], proposent d'utiliser une autorité de certification centralisée pour identifier chaque agent. Toutefois, cela réduit les propriétés de décentralisation et d'ouverture du système, et introduit un nouveau point de rupture. Une autre approche propose d'introduire des coûts récurrents pour participer au système, comme résoudre un puzzle cryptographique [4], payer un taxe monétaire ou utiliser des *captchas*. Ceci rend difficile la création d'un grand nombre de fausses identités mais ne tient pas compte du fait que les agents malveillants peuvent avoir à leur dispositions des ressources considérables comme un réseau de machines-zombies. De plus, ces méthodes sont très contraignantes pour les agents honnêtes.

Une deuxième classe d'approches consiste à détecter les coalitions malveillantes à l'intérieur du graphe de confiance en considérant que les interactions observées structurent un réseau social [22]. Il est alors fait l'hypothèse que les agents malveillants présentent soit un taux de *clusterisation* très élevé et peu de liens avec les agents honnêtes [9, 24], soit des identifiants similaires [8]. Ces approches utilisent alors des techniques de *clustering* provenant du domaine

de l'analyse de graphes ou de la découverte de liens pour séparer les agents honnêtes des agents malveillants. Toutefois, ces approches reposent sur des hypothèses fortes quant à la structure des interactions entre agents.

Afin de proposer un cadre général permettant de pallier ces limites, des travaux récents s'intéressent à des approches fondées sur la théorie des jeux [10, 17, 18, 19] en faisant l'hypothèse que les agents malveillants sont rationnels. Ces approches empruntent aux problèmes de manipulations sur les jeux de vote pondéré [2] et les enchères combinatoires [23] le concept d'imperméabilité aux fausses identités (*false-name-proofness* en anglais). Ce concept signifie qu'un agent ne doit jamais tirer un bénéfice à participer plus d'une fois à un même jeu. Par exemple, le protocole Informant [18] et son application aux réseaux Tor [19] font participer les agents à une enchère hollandaise qui récompense les agents malveillants se dénonçant eux-mêmes. Toutefois, cette récompense est générée par des coûts récurrents et incite de nouveaux agents malveillants à entrer dans le système.

Nous proposons alors un protocole pour prévenir les auto-promotions sans coût récurrent, ni calcul complexe pour les agents honnêtes. Ce protocole autorise les agents honnêtes à utiliser une attaque Sybil afin de placer les agents malveillants rationnels (et uniquement eux) face à un dilemme. Le protocole utilise ensuite la réponse à ce dilemme pour partitionner les agents du système entre honnêtes et malveillants.

3 Description du protocole

Dans cette section, nous présentons en premier lieu le système de réputation puis les manipulations que nous considérons. En second lieu, nous détaillons notre protocole.

3.1 Système de réputation

Dans ce travail, nous nous fondons sur la définition d'un système de réputation proposée par [6]. Cette définition permet de capturer la classe des systèmes de réputation personnalisés tels que Troika [1], Histos [25], BetaReputation [20] ou FlowTrust [6]. Personnalisé signifie ici que la réputation d'un agent dépend de celui qui l'évalue. Dans toute la suite, nous notons A l'agent évaluateur, B l'agent évalué et W_i chaque agent témoin.

Définition 1. *Soit $\mathcal{G} = (V, E)$ un graphe orienté où V est l'ensemble des agents et $E \subseteq V \times V$ une relation d'interaction étiquettée par une valeur de confiance $c : E \longmapsto [0, 1]$. La réputation d'un agent B selon A est donnée par une fonction $f_{\mathcal{G}} : V \times V \longmapsto [0, 1]$ où :*

$$f_{\mathcal{G}}(A, B) = \bigoplus_{P \in \mathcal{P}_{AB}} \odot(P)$$

$\mathcal{P}_{AB}$ est l'ensemble maximal par l'inclusion des chemins disjoints entre A et B dans $\mathcal{G}$, $\odot$ est un opérateur d'agrégation de c le long d'un unique chemin P entre A et B, et $\oplus$ est un opérateur d'agrégation de $\odot$ sur tous les chemins disjoints entre A et B.

Exemple 1. *Le système de réputation FlowTrust proposé par [6] est défini par $\odot = \prod$ et $\oplus = \max$. Ainsi, la réputation d'un agent B selon A est donnée par le valeur maximale des produits des valeurs de confiance parmi tous les chemins disjoints entre A et B.*

Après avoir calculé la réputation de B, l'agent A doit décider s'il fait confiance ou non à B.

Définition 2. *Soit $d_A : V \longmapsto \{0, 1\}$ une fonction de décision où 0 signifie que A ne fait pas confiance à B et 1 signifie que A fait confiance à B. d_A est une fonction de seuil $\theta_A \in [0, 1]$ sur $f_{\mathcal{G}}(A, B)$ telle que $d_A(B) = 0$ si $f_{\mathcal{G}}(A, B) \leq \theta_A$ et $d(B) = 1$ sinon.*

Exemple 2. *Reprenons l'exemple 1. Dans FlowTrust, $\theta_A = 0,5$. Ainsi, un agent A a confiance en un agent B si $f_{\mathcal{G}}(A, B) > 0,5$.*

Dans ce modèle, $f_{\mathcal{G}}$ représente le mécanisme d'agrégation des confiances, $\odot$ et $\oplus$ sont des connaissances communes des agents tandis que d_A est une fonction de décision privée et subjective. C'est pourquoi d_A peut être différente pour chaque agent. Quoi qu'il en soit, quelle que soit la définition de d_A, calculer la réputation d'un agent revient à construire et parcourir le graphe d'interaction $\mathcal{G}$ pour demander aux agents sur les chemins entre A et B leur valeur de confiance envers B. Ces agents peuvent alors collectivement manipuler le système en rapportant à A de fausses valeurs de confiance pour B.

3.2 Caractériser les manipulations

Dans cet article, nous considérons des manipulations d'auto-promotion où des agents malveillants se soutiennent mutuellement et désirent tromper tous les autres agents du système. Cette définition correspond à des manipulations

effectées sur des systèmes réels, comme les manipulations du PageRank, les empoisonnements de réseaux Tor ou le resquillage sur des réseaux pair-à-pair. En effet, dans tous ces cas, l'objectif des agents malveillants est de pousser n'importe quel agent honnête à interagir avec au moins l'un des membres de la coalition malveillante. Par conséquent, une auto-promotion est définie comme suit.

Définition 3. *$\forall\ A \in V$ qui interroge un agent W_i sur sa valeur de confiance $c(W_i, B)$ envers un agent B, si W_i est en collusion avec B alors $c(W_i, B) > \theta_{W_i}$.*

Cette définition implique deux symétries. La première est une symétrie de comportement entre les agents malveillants : chaque agent de la coalition malveillante va rapporter une valeur de confiance élevée envers les autres. Ils forment donc une grappe de confiance mutuelle. La seconde est une symétrie de comportement des agents malveillants envers les agents honnêtes : les agents malveillants agissent de la même manière quel que soit l'agent honnête qui les interroge car ils désirent tromper tous les autres agents du système.

Considérant ceci, nous pouvons alors caractériser une coalition malveillante. Comme cette coalition est définie par une confiance mutuelle élevée, si l'on considère deux agents témoins distincts qui font individuellement confiance à un agent tier, plus ces témoins ont confiance les uns dans les autres, plus il est possible qu'ils soient en collusion. C'est pourquoi, nous faisons l'hypothèse que chaque agent honnête dispose d'une fonction de suspicion qui calcule la probabilité (subjective) que deux agents soient en collusion.

Définition 4. *Soit $M_A : V \times V \longmapsto [0, 1]$ la* fonction de suspicion *de l'agent A telle que : $M_A(W_1, W_2) = c(W_1, W_2) \times c(W_2, W_1)$ si $c(W_1, W_2) > \theta_A$ et $c(W_2, W_1) > \theta_A$. Sinon $M_A(W_1, W_2) = 0$.*

Ainsi, plus $M_A(W_1, W_2)$ est élevée, plus il est possible que les agents W_1 et W_2 soient en collusion. Il convient de noter qu'une telle fonction n'est qu'une heuristique représentant une connaissance sur la caractérisation des coalitions malveillantes. Il est alors possible d'étendre cette définition afin de tenir compte d'informations supplémentaires que le concepteur du système de réputation pourrait avoir sur les agents malveillants. Par exemple, dans le cas d'un réseau pair-à-pair, la valeur de suspicion pourrait être pondérée proportionnellement à la distance d'édition entre les adresses IP des pairs. En effet, des pairs Sybil créés à partir d'une même machine ont tendance à partager une même plage d'adressage. Toutefois, dans le cadre de ce travail, nous ne considérons que les valeurs de confiance échangées entre les membres des coalitions malveillantes afin de minimiser le nombre d'hypothèses sur le système sous-jacent.

À partir de cette heuristique, un agent honnête peut demander à plusieurs agents témoins leurs valeurs de confiance mutuelles afin de déterminer s'il y a une coalition malveillante. Toutefois sachant cela, les agents malveillants peuvent cacher leur relation de confiance. Nous devons donc définir un protocole permettant d'obtenir cette information.

3.3 Faire parler les témoins

Le protocole que nous proposons doit détecter les agents en collusion parmi l'ensemble des agents témoins alors même que ces derniers, s'ils sont malveillants, peuvent cacher cette collusion. Afin d'inciter les agents malveillants à se révéler, les agents honnêtes vont utiliser une attaque Sybil pour dissimuler leur intention : les agents témoins seront contre-interrogés par plusieurs agents apparemment distincts, l'agent honnête et ses agents Sybil. Parmi les agents témoins, les agents maveillants caractérisés par la définition 3 sont alors incités à révéler leur confiance mutuelle afin de tromper tous les autres agents du système. L'algorithme 1 présente les principales étapes du protocole.

Algorithme 1 Protocole

```
1:  calculer f_G(A, B)
2:  si d_A(B) = 1 alors
3:    W' = {W_i ∈ W : c(W_i, B) > θ_A}
4:    pour tout W_i, W_j ∈ W' (i ≠ j) faire
5:      générer deux agents Sybil A' et A''
6:      A' demande à W_i sa valeur c(W_i, W_j)
7:      A'' demande à W_j sa valeur c(W_j, W_i)
8:      calculer M_A(W_i, W_j)
9:    fin pour
10:   retourne réviser f_G(A, B)
11: fin si
```

Tout d'abord, A utilise le système de réputation afin de calculer la réputation de B (ligne 1). Ainsi, A obtient un ensemble W de témoins

pour B et peut calculer $d_A(B)$. Comme remarqué dans [10], B peut être un agent honnête diffamé si $d_A(B) = 0$ ou B peut être un agent malveillant faisant de l'auto-promotion si $d_A(B) = 1$. En effet, seules les manipulations ayant été réussies doivent être considérées ; et une manipulation n'est réussie que si et seulement si elle conduit A à prendre la « bonne » décision, c'est-à-dire faire confiance s'il y a eu de l'auto-promotion. Dans ce cas, le protocole ne s'applique que lorsque $d_A(B) = 1$ (ligne 2).

Ensuite, A sélectionne le sous-ensemble de témoins qui ont confiance dans B (ligne 3). Pour chaque couple de témoins, A génère deux agents Sybil (ligne 4 et 5) qui vont interroger les potentiels agents malveillants. Ces agents Sybil demandent à W_i s'il a confiance dans W_j (ligne 6) et inversement (ligne 7). Si W_i (respectivement W_j) est honnête, il répondra honnêtement ; si W_i est malveillant, il sera incité à répondre selon la définition 3. Une fois que A a obtenu l'ensemble des valeurs de suspicion $M_A(W_i, W_j)$, il les utilise pour réviser $f_{\mathcal{G}}(A, B)$ (ligne 10).

3.4 Réviser la valeur de réputation

Plus la suspicion envers un agent témoin W_i est élevée, moins le témoignage $c(W_i, B)$ de ce dernier est digne de confiance. Nous proposons un mécanisme de révision qui retire stochastiquement certains témoignages de $f_{\mathcal{G}}(A, B)$. L'algorithme 2 présente les principales étapes du protocole.

Algorithme 2 Révision

1: $\mathcal{G}' \leftarrow \mathcal{G}$
2: **pour tout** $W_i \in W'$ **faire**
3: $\quad M_i \leftarrow \max_{W_j \in W'} M_A(W_i, W_j)$
4: $\quad V' \leftarrow V' \setminus \{W_i\}$ avec une probabilité M_i
5: **fin pour**
6: **retourne** $f_{\mathcal{G}'}(A, B)$

Le mécanisme procède comme suit : l'agent A retire du graphe d'interaction chaque témoin avec une probabilité égale à la plus haute valeur de suspicion qui lui a été calculée. Comme, selon la définition 1, les chemins considérés dans le graphe sont disjoints, retirer un témoin revient à retirer l'ensemble du chemin. De plus, s'il y a plusieurs témoignages suspects sur un chemin, chacun d'entre eux peut le retirer indépendamment.

3.5 Coût du protocole

Ce protocole est une couche implantée au-dessus d'un système de réputation existant. Il accroît donc le coût de communication global du système. Ce coût de communication correspond au nombre de messages échangés au cours d'un calcul de réputation pour un agent donné.

Proposition 1. *Le coût de communication du protocole donné par l'algorithme 1 est* $\mathcal{O}(4.|W-1|^2)$.

Démonstration. Chaque fois que la réputation d'un agent B est calculée par un agent A, chaque paire de témoins distincts dans W est interrogée, ce qui génère quatre messages (deux questions puis deux réponses). Ainsi trivialement, le protocole ajoute $4.|W-1|^2$ nouveaux messages. □

La nouvelle réputation ($f_{\mathcal{G}'}(A, B)$) calculée par ce protocole n'identifie pas formellement les agents malveillants mais atténue l'influence des témoignages des agents présentant une confiance mutuelle élevée. Si cela peut retirer le témoignage d'agents honnêtes, cela retire aussi ceux des agents malveillants. Si ces derniers désirent manipuler le système, ils vont devoir définir une stratégie pour répondre au contre-interrogatoire, sachant que cet interrogatoire se glisse au milieu de requêtes anodines.

4 Analyse du dilemme

Comme le protocole est connu de tous les agents, un agent malveillant doit se demander à chaque requête s'il est interrogé par un agent honnête ou par un agent Sybil. Cela revient à décider si l'agent maveillant révèle ou non sa valeur de confiance dans un autre agent malveillant lorsqu'un agent tiers la lui demande. Prendre une telle décision correspond à résoudre un dilemme que nous analysons du point de vue de la théorie de jeux. Pour cela, nous présentons tout d'abord le jeu sous sa forme stratégique, puis nous montrons certaines de ses propriétés.

4.1 Jeu sous forme stratégique

En supposant les agents malveillants rationnels, la table 1 représente, du point de vue de l'agent malveillant, le dilemme comme un jeu à somme nulle sous forme stratégique.

Agent	Révéler	Dissimuler
Honnête	$(1-\delta)g$	0
Sybil	$-\delta p$	δg

TABLE 1 – Jeu sous forme stratégique

Définition 5. *Soit $g \in \mathbb{R}$ et $p \in \mathbb{R}$ $(g > p)$ tels que g est le gain de l'agent malveillant à manipuler le système et p la pénalité à être identifié comme agent malveillant. Soit $\delta \in [0, 1]$ la probabilité qu'un agent donné soit un agent Sybil.*

Informellement, le dilemme est le suivant. D'un côté, si l'agent malveillant révèle sa valeur de confiance dans un autre à un agent honnête, il manipule l'agent honnête et est donc récompensé pour atteindre son objectif. Cette récompense correspond au gain à manipuler le système. De même, si l'agent malveillant dissimule sa valeur de confiance dans un autre à un agent Sybil alors il manipule l'agent honnête qui a généré la Sybil, et il obtient donc la même récompense car il est parvenu à manipuler le système. D'un autre côté, si l'agent malveillant dissimule sa valeur de confiance à un agent honnête, son gain est nul car il ne pourra pas manipuler le système. Dans le dernier cas, si l'agent malveillant révèle sa valeur de confiance dans un autre à un agent Sybil, il est pénalisé car non seulement sa manipulation échoue comme dans le cas précédent mais, de plus, son identité peut être compromise pour de futures interactions.

Un agent malveillant rationnel va donc chercher à partir de cette matrice une stratégie (se révéler ou se dissimuler) qui maximisera son gain. Or, cette matrice correspond à jeu d'appariement (*matching pennies game*) où δ est le paramètre de stratégie mixte de l'adversaire de l'agent malveillant, c'est-à-dire du protocole. Dans ce jeu, il n'existe pas de stratégie pure maximisant le gain d'un agent [11]. En conséquence, les agents malveillants, s'ils sont rationnels, sont obligés de jouer une stratégie mixte.

4.2 Équilibre de Nash en stratégie mixte

Notons par R l'action de révéler et D celle de dissimuler. Notons H le fait que l'agent évaluateur soit un agent honnête et S le fait qu'il soit un agent Sybil. Nous pouvons alors noter $\pi_B = \langle \sigma(R) = (1-m), \sigma(D) = m \rangle$ le profil de stratégie mixte de l'agent malveillant et $\pi_p = \langle \sigma(H) = (1-\delta), \sigma(S) = \delta \rangle$ le profil de stratégie mixte du protocole.

Proposition 2. *L'*équilibre de Nash en stratégie mixte *du jeu caractérisé par la table 1 est :*

$$m = \frac{g+p}{2g+p} \qquad \textit{et} \qquad \delta = \frac{g}{2g+p}$$

Démonstration. L'utilité espérée de π_B en fonction de m et δ est :

$$\begin{aligned} u_{\pi_B}(m, \delta) &= (1-m)((1-\delta)g - \delta p) + m\delta g \\ &= g - \delta g - \delta p + m(2\delta g + \delta p - g) \end{aligned}$$

Comme l'agent malveillant désire maximiser $u(\pi_B)$, les racines des dérivées partielles de $u_{\pi_B}(m, \delta)$ en fonction de δ sont :

$$\begin{aligned} -g - c + m(2g+p) &= 0 \\ m &= \frac{g+p}{2g+p} \end{aligned}$$

De même, comme le protocole désire minimiser $u(\pi_B)$, les racines des dérivées partielles de $-u_{\pi_B}(m, \delta)$ en fonction de m sont :

$$\begin{aligned} -2\delta g - \delta p + g &= 0 \\ \delta &= \frac{g}{2g+p} \end{aligned}$$

□

Remarquons que $m = 1 - \delta$, ce qui est caractéristique des jeux d'appariement. De manière intéressante, même si la pénalité p est nulle, un agent rationnel qui désire maximiser son gain doit jouer une stratégie telle que $m = \delta = \frac{1}{2}$.

4.3 Probabilité d'attaque réussie

Toutefois, même si l'agent malveillant joue une stratégie mixte qui maximise son gain, il lui faut manipuler le protocole deux fois de suite. En effet, un agent malveillant doit tout d'abord tromper l'agent honnête, puis tromper l'agent Sybil qui est ensuite généré. Dans tous les autres cas, la manipulation est un échec car dissimuler la valeur de confiance face à un agent honnête revient à abandonner la manipulation, et révèler la valeur de confiance face à l'agent Sybil revient à être sanctionné. En conséquence, une manipulation réussie est définie comme suit.

Définition 6. *Une manipulation est* réussie *si et seulement si l'agent malveillant joue R contre un agent honnête, puis D contre l'agent Sybil qui a été généré.*

Agent	Révéler	Dissimuler
Honnête	$m(1-m)$	m^2
Sybil	$(1-m)^2$	$m(1-m)$

TABLE 2 – Occurences des stratégies jointes

Selon la stratégie mixte déterminée précédemment (et $m = 1 - \delta$), la table 2 donne les probabilités d'occurence de ces stratégies jointes. Nous faisons l'hypothèse qu'en raison de l'entrelacement des requêtes dans l'ensemble du système et des latences induites par le temps de génération des agents Sybil, un agent malveillant ne peut pas déterminer si deux dilemmes sont corrélés.

Proposition 3. *La probabilité qu'une manipulation soit réussie est* $m^2 - 2m^3 + m^4$.

Comme nous connaissons la valeur optimale pour m grâce à la proposition 2, nous pouvons exprimer la probabilité de succès en fonction de g et p. De plus, nous pouvons exprimer p comme une fraction de g. Dans ce cas, même si $p = 0$, la probabilité qu'une manipulation soit réussie n'est que de $0,0625$. Toutefois, il est important de noter que cela ne concerne que deux agents malveillants. Si la coalition est plus grande alors cette probabilité augmente.

Pour conclure cette analyse, le protocole que nous avons défini incite les agents malveillants à révéler leurs valeurs de confiance mutuelles, ce qui peut les conduire à être identifiés comme suspects. Ceci place les agents malveillants face à un dilemme qui ne peut être résolu de manière optimale qu'en jouant une stratégie mixte. Cette stratégie mixte les conduit alors à abandonner volontairement des manipulations. De plus, comme une manipulation réussie correspond à un jeu itéré dont chaque instance est décorrélée, la probabilité de réussite d'une manipulation par un unique agent est faible.

5 Résultats de simulation

Afin d'évaluer notre approche, nous l'avons implantée au-dessus du système FlowTrust [6] et nous avons comparé ses performances avec et sans dilemme. Notons que ces résultats dépendent d'un grand nombre de paramètres, de la topologie du graphe à la distribution des valeurs de confiance en passant par le type de système de réputation même. Toutefois, cela nous procure un éclairage sur l'efficacité du protocole.

5.1 Paramètres expérimentaux

Nous considérons un graphe d'interaction construit sur un graphe binomial d'Erdös-Rényi de paramètre $0,15$. Les valeurs de confiance sont fixées selon une distribution uniforme. Lors d'une expérience, nous fixons le nombre d'agents à 50 puis 100 sans faire d'hypothèse sur leur position au sein du graphe. Les agents malveillants sont ensuite tirés aléatoirement de manière uniforme parmi les agents du système.

Pour chaque expérience, nous lançons 10000 simulations où un agent honnête tiré aléatoirement évalue un ensemble d'agents aléatoires. En effet, un système de réputation totalement décentralisé ne peut pas évaluer l'ensemble du réseau sans problèmes de passage à l'échelle. L'agent honnête calcule alors la réputation de chaque agent évalué sans notre protocole, puis avec notre protocole où les agents malveillants jouent la stratégie pure de toujours révéler (non optimale), puis une stratégie mixte (optimale). Dans chaque cas, si un agent malveillant obtient la plus haute valeur de réputation, nous considérons que la manipulation est réussie. Notre critère de performance est la proportion de telles manipulations sur les 10000 simulations.

Nous avons choisi de nous comparer au système FlowTrust présenté dans l'exemple 1 car ce dernier, bien que peu informatif, est robuste aux manipulations lorsque les agents ont une vision globale du système. De plus, pour nous placer dans le cadre le moins favorable, nous considérons que les agents malveillants n'ont pas de pénalité de collusion ($p = 0$). Tous les résultats sont donnés dans les figures 1, 2 et 3. Sur chacune de ces figures, les courbes rouges (diamant) représentent les manipulations réussies sans notre protocole, les courbes vertes (triangle pointant vers le haut) représentent les manipulations réussies sous stratégie mixte tandis que les courbes jaunes (triangle pointant vers le bas) illustrent la stratégie pure.

5.2 Du nombre d'agents malveillants

Afin d'étudier l'influence du réseau en termes de taille et du nombre d'agents malveillants, nous avons fait varier la proportion d'agents malveillants entre 10 et 50% tandis qu'un agent honnête évalue 5 agents sélectionnés aléatoirement. Les résultats sont donnés en figure 1.

Sans surprise, dans tous les cas, le nombre de manipulations réussies augmente lorsque le

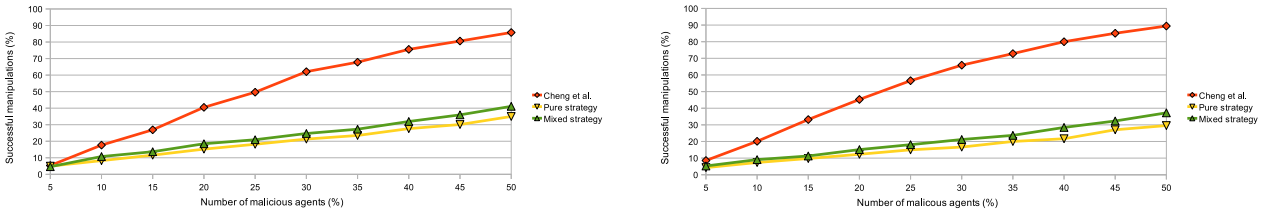

FIGURE 1 – Manipulations réussies selon la taille de la coalition (réseau de taille 50 puis 100)

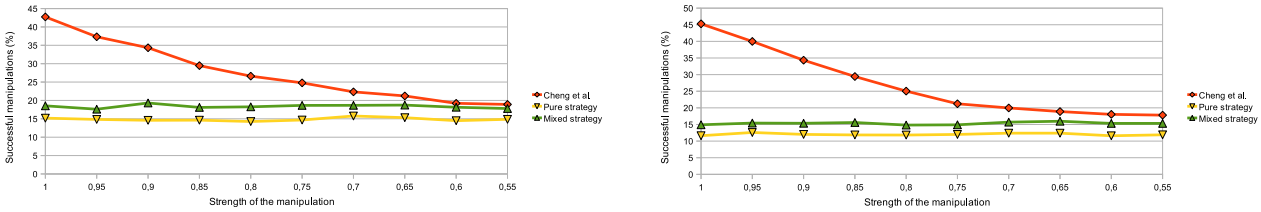

FIGURE 2 – Manipulations réussies selon la force de la manipulation (réseau de taille 50 puis 100)

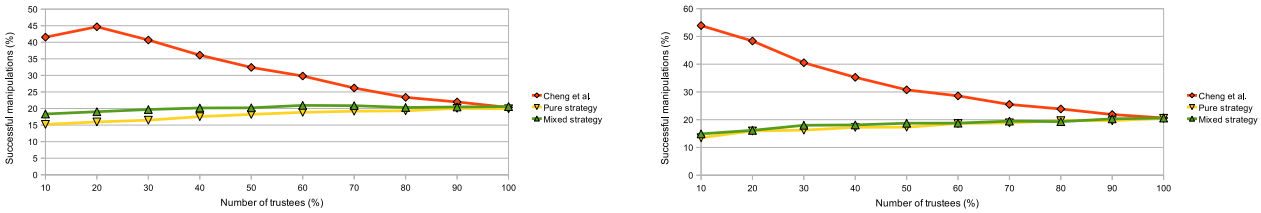

FIGURE 3 – Manipulations réussies selon le nombre d'agents évalués (réseau de taille 50 puis 100)

nombre d'agents malveillants augmente. Toutefois, la stratégie mixte réduit en moyenne de 55% le nombre de manipulations réussies tandis que la stratégie pure les réduit de 62%. Le protocole est un peu plus efficace lorsque la taille du réseau augmente puisqu'il y a un gain de 11% en moyenne sur les réseaux de taille 100 par rapport aux réseaux de taille 50. En conséquence, plus il y a d'agents dans le réseau, plus notre protocole est efficace. Les résultats théoriques sont aussi confirmés : les agents malveillants doivent jouer une stratégie mixte pour maximiser le nombre de manipulations réussies. Toutefois, même si les agents malveillants jouent une stratégie mixte optimale et qu'ils ne subissent pas de pénalité, les manipulations sont fortement réduites par notre protocole.

5.3 De la force de la manipulation

Un agent malveillant rationnel peut désirer réduire la force de sa manipulation afin d'éviter d'être suspecté par les agents honnêtes. Pour cela, la coalition malveillante va présenter une valeur de confiance mutuelle plus faible afin de réduire les valeurs calculées par la fonction de suspicion des agents honnêtes. Afin de mettre en évidence l'effet d'un tel comportement, nous avons considéré des réseaux contenant 20% d'agents malveillants. Dans chaque expérience, un agent honnête évalue 5 agents sélectionnés aléatoirement, et nous avons fait varier les valeurs de confiance que les agents malveillants rapportent dans l'intervalle $[0.5, 1]$. Les résultats sont donnés en figure 2.

Sans surprise encore, le nombre de manipulations réussies sans notre protocole diminue au fur et à mesure que les agents malveillants réduisent la force de leur manipulation jusqu'à converger vers 20%. Toutefois, nous pouvons remarquer que notre protocole est très efficace car, quelle que soit la force de la manipulation, le nombre de manipulations réussies reste autour de 20% pour la stratégie mixte et 15% pour la stratégie pure. Lorsque le réseau croît en taille, ces nombres sont réduits à 15% et 10% respectivement, confortant le fait que plus il y a d'agents dans le réseau, plus notre protocole est efficace. De plus, et de manière surprenante, l'efficacité du protocole reste stable quelle que soit la valeur exprimée par les agents malveillants. Bien que l'algorithme 2 retire les suspects stochastiquement, la performance n'est

pas dégradée lorsque les agents malveillants adoptent un comportement imitant les agents honnêtes. Ainsi, notre protocole peut prévenir non seulement les manipulations mais il ne semble pas pouvoir être manipulé en retour.

5.4 De la quantité d'information

Jusqu'à présent, nous avons uniquement considéré des cas décentralisés où les agents honnêtes n'évaluent qu'un sous-ensemble du réseau selon un compromis entre exploitation et exploration. Afin de mettre en évidence l'effet de ce compromis, nous fait varier la quantité d'information qu'un agent possède sur l'ensemble du réseau. Pour cela, nous avons considéré des réseaux contenant 20% d'agents malveillants et nous avons fait varier le nombre d'agents évalués par agent honnête entre 10% du réseau et 100%. Les résultats sont donnés en figure 3.

Cette fois, les résultats sont similaires sur les deux tailles de réseau. Nous pouvons remarquer que le nombre de manipulations réussies sans notre protocle décroît au fur et à mesure que le nombre d'agents évalués augmente. Ainsi, l'efficacité de notre approche sous stratégie mixte passe d'une réduction de 64% des manipulations jusqu'à un gain nul. Nous pouvons aussi remarquer que la stratégie mixte est aussi efficace que la stratégie pure : les agents malveillants ne peuvent pas augmenter le nombre de manipulations réussies en jouant sur leur stratégie. Dans les deux cas, les performances de notre protocole convergent pour atteindre celles du système de réputation originel. Toutefois, ce système de réputation est robuste aux manipulations lorsqu'il est centralisé. Or, dans un système réel, il n'est pas toujours possible de faire cette hypothèse. Aussi, notre protocole reste très efficace dans le cas général.

6 Conclusion

Afin d'assurer le fonctionnement nominal des systèmes ouverts et décentralisés, la présence d'agents malveillants doit être considérée. Les systèmes de réputation répondent à cette problématique mais, s'ils sont efficaces pour détecter des agents malveillants isolés, ils sont vulnérables aux attaques Sybil, et plus généralement aux coalitions d'agents malveillants. Ainsi, prévenir les manipulations provenant de coalitions est une problématique critique pour les systèmes de réputation. De nombreuses propositions ont été faites, depuis l'utilisation des puzzles cryptographiques à la définition de fonctions de réputation robustes en passant par la détection de communautés. Toutefois, ces approches impliquent de recentraliser partiellement le système ou d'introduire un coût récurrent sur les agents. Cependant, une approche plus récente se fonde sur la théorie des jeux dans le but de définir des mécanismes incitant les agents malveillants à ne pas commettre de manipulations.

Dans ce contexte, nous proposons un protocole mettant en œuvre un dilemme pour détecter et prévenir les auto-promotions dans un système de réputation. Ce protocole se fonde sur le fait que les agents honnêtes utilisent à leur tour une attaque Sybil pour manipuler les agents malveillants. Ce protocole conduit les agents malveillants à révéler les liens de confiance mutuelle qui les unissent, liens utilisés comme heuristique pour la détection de collusions. Notre analyse théorique du protocole montre que les agents malveillants, s'ils sont rationnels, doivent jouer une stratégie mixte et, en conséquence, abandonner certaines manipulations pour maximiser leur efficacité. Nos résultats expérimentaux montrent que notre protocole réduit en moyenne de moitié les manipulations réussies. De plus, son efficacité n'est pas diminuée lorsque les agents malveillants tentent de se faire passer pour des agents honnêtes.

Toutefois, ce travail ouvre plusieurs perspectives. Tout d'abord, de nouvelles expériences doivent être considérées pour mettre en évidence les limites de notre approche. Nous pouvons nous demander par exemple comment se comporte le protocole lorsque les valeurs de confiances entre agents honnêtes sont correlées. De plus, notre protocole doit être comparé à d'autres systèmes de réputation, notamment à EigenTrust [15] qui est un système non-personnalisé n'entrant pas dans le cadre du modèle de Cheng et Friedman [6]. Dans un second temps, nous devons nous pencher sur la fonction de suspicion. En effet, cette fonction est une heuristique qui caractérise ce qu'est un comportement de coalition malveillante, et elle peut ainsi être définie de plusieurs manières. Par exemple, un agent qui produit de trop nombreux témoignages pourrait être considéré comme un agent suspect. De plus, comme nous n'avons pas fait d'hypothèse sur la structure du graphe d'interaction, nous pourrions combiner notre heuristique avec d'autres fondées, cette fois, sur la topologie du réseau comme celle utilisée dans SybilLimit [24] afin d'améliorer la performance du protocole. Une autre voie consiste à consi-

dérer la dynamique du système. En effet, notre protocole ne considère qu'une vue du système à un instant donné. Mais, si un agent malveillant joue une stratégie mixte, un agent honnête joue, quant à lui, une stratégie pure. En conséquence, si le protocole considérait plusieurs réponses successives au dilemme, ceci permettrait de détecter quel type de stratégie joue un agent, et donc de déduire s'il est malveillant ou non. Raisonner sur la stratégie, et non sur la réponse au dilemme elle-même, est une voie intéressante pour pallier les limites actuelles de ce protocole.

Références

[1] A. Abdul-Rahman and S. Hailes. Using recommendations for managing trust in distributed systems. In *3rd MICC*, 1997.

[2] Y. Bachrach and E. Elkind. Divide and conquer : false-name manipulations in weighted voting games. In *7th AAMAS*, pages 975–982, 2008.

[3] G. Bonnet. A protocol based on a game-theoretic dilemma to prevent malicious coalitions in reputation systems. In *28th ECAI*, pages 187–192, 2012.

[4] N. Borisov. Computational puzzles as Sybil defenses. In *6th P2P*, pages 171–176, 2006.

[5] M. Castro, P. Drusche, A. Ganesh, A. Rowstron, and D.-S. Wallach. Secure routing for structured peer-to-peer overlay networks. In *5th OSDI Symposium*, 2002.

[6] A. Cheng and E. Friedman. Sybilproof reputation mechanisms. In *3rd P2PEcon*, pages 128–132, 2005.

[7] A. Cheng and E. Friedman. Manipulability of PageRank under Sybil strategies. In *1st NetEcon*, 2006.

[8] T. Cholez, I. Chrisment, and O. Festor. Efficient DHT attack mitigation through peers' ID distribution. In *24th IPDPS*, pages 1–8, 2010.

[9] V. Conitzer, N. Immorlica, J. Letchford, K. Munagala, and L. Wagman. False-name-proofness in social networks. In *6th WINE*, pages 1–17, 2010.

[10] V. Conitzer and M. Yokoo. Using mechanism design to prevent false-name manipulations. *AI Magazine*, Vol. 31(4) :65–77, 2010.

[11] T. Dang. Gaming or guessing : mixing and best-responding in matching pennies. Technical report, University of Arizona, 2009.

[12] F. Dini and G. Spagnolo. Buying reputation on eBay : do recent changes help ? *IJEB*, Vol. 7(6) :581–598, 2009.

[13] J.-R. Douceur. The Sybil attack. In *1st IPTPS*, 2002.

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

[15] S.-D. Kamvar, M.-T. Schlosser, and H. Garcia-Molina. The EigenTrust algorithm for reputation management in P2P networks. In *12th WWW*, pages 640–651, 2003.

[16] B.-N. Levine, C. Shields, and N.-B. Margolin. A survey of solutions to the sybil attack. Technical report, University of Massachussets Amherst, 2006.

[17] X. Liao, D. Hao, and K. Sakurai. A taxonomy of game theoretic approaches agains attacks in wireless ad hoc networks. In *28th SCIS*, pages 1–8, 2011.

[18] N.-B. Margolin and B.-N. Levine. Informant : detecting Sybils using incentives. In *11th FC*, pages 192–207, 2007.

[19] A.-K. Pal, D. Nath, and S. Chakreborty. A discriminatory rewarding mechanism for Sybil detection with applications to Tor. In *8th ICCCIS*, pages 84–91, 2010.

[20] A. Jøsang and R. Ismail. The beta reputation system. In *15th Bled EC Conference*, 2002.

[21] M. Sirivianos, J.-H. Park, R. Cheng, and X. Yang. Free-riding in BitTorrent networks with the large view exploit. Technical report, California Irvine, 2001.

[22] B. Viswanath, A. Post, K.-P. Gummadi, and A. Mislove. An analysis of social network-based Sybil defenses. In *16th SIGCOMM*, 2010.

[23] 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.

[24] H. Yu, P.-B. Gibbons, M. Kaminsky, and X. Feng. SybilLimit : a near-optimal social network defense against Sybil attacks. *IEEE/ACM Transactions on Networking*, Vol. 18(3) :885–898, 2010.

[25] G. Zacharia and P. Maes. Trust management through reputation mechanisms. *Applied Artificial Intelligence*, Vol. 14 :881–907, 2000.