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

Grégory Bonnet

Normandie Université, France
UNICAEN, GREYC, F-14032 Caen, France
CNRS, UMR 6273, F-14032 Caen, France
gregory.bonnet@unicaen.fr

RÉSUMÉ. *Dans les systèmes ouverts et décentralisés, un grand nombre d'agents interagissent pour partager des ressources ou des tâches. 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 ou de la diffamation. 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.*

ABSTRACT. *In decentralized and open systems, a large number of agents interact in order to share resources or tasks. 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 and slandering through 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.*

MOTS-CLÉS : *confiance et réputation, manipulations, raisonnement stratégique, théorie des jeux.*

KEYWORDS: *trust and reputation, manipulations, strategic reasoning, game theory.*

DOI:10.3166/RIA.28.411-431 © 2014 Lavoisier

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 en détourner les règles afin d'en tirer parti, voire de provoquer une rupture du service lui-même. Pour 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 (Hoffman *et al.*, 2009).

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 (Dini, Spagnolo, 2009), l'algorithme PageRank de Google (Cheng, Friedman, 2006), voire même pour resquiller sur un réseau pair-à-pair (Sirivianos *et al.*, 2001), c'est-à-dire consommer des ressources sans en fournir en retour. À l'inverse, des agents malveillants peuvent 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 (Douceur, 2002). C'est pourquoi se prémunir des coalitions malveillantes est une problématique fondamentale pour les systèmes de réputation (Levine *et al.*, 2006).

De nombreuses propositions ont été faites dans la littérature 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 (Conitzer, Yokoo, 2010). 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.

Nous proposons dans cet article une approche fondée sur la théorie des jeux pour empêcher les manipulations particulières que sont les auto-promotions et la diffamation de se produire dans un système de réputation. Cet article vient étendre des résultats précédents qui ne concernaient que l'auto-promotion (Bonnet, 2012 ; 2013) en considérant également la diffamation. 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 présentons l'état de l'art en section 2. La section 3 est consacrée à la description du protocole que nous analysons ensuite formellement dans la section 4. Enfin, nous présentons 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 asymétrique, signifiant que les valeurs de confiance locales sont propagées au travers d'agents considérés a priori comme de confiance. (Cheng, Friedman, 2005) ont montré que les fonctions de réputation symétriques ne pouvaient pas prémunir des manipulations tandis que les fonctions asymétriques le pouvaient à condition de satisfaire des propriétés restrictives qui les rendent faiblement informatives et délicates à définir. Ces travaux sont à rapprocher de ceux de (Altman, Tennenholtz, 2005 ; 2010) qui, s'intéressant aux systèmes de recommandation, ont montré que des axiomes jugés intéressants (comme un axiome de majorité, d'indépendance aux alternatives non pertinentes et d'incitation à la sincérité) ne peuvent pas être simultanément satisfaits. 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 (Levine *et al.*, 2006).

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 (Castro *et al.*, 2002) ou (Douceur, 2002), 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 (Borisov, 2006), payer une taxe monétaire ou utiliser des *captchas* (Von Ahn *et al.*, 2003). 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 disposition 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 (Viswanath *et al.*, 2010). 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 (Conitzer *et al.*, 2010 ; Yu *et al.*, 2010), soit des identifiants similaires (Cholez *et al.*, 2010). 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 (Conitzer, Yokoo, 2010 ; Liao *et al.*, 2011 ; Margolin, Levine, 2007 ; Pal *et al.*, 2010) en faisant l'hypothèse que les agents malveillants sont rationnels. Ces approches empruntent le concept d'imperméabilité aux fausses identités (*false-name-proofness* en anglais) aux problèmes de manipulations sur les jeux de vote pondéré (Bachrach, Elkind, 2008), les enchères combinatoires (Yokoo *et al.*, 2004) et, plus généralement, le choix social computationel (Brandt *et al.*, 2012). 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 (Margolin, Levine, 2007) et son instanciation sur les réseaux Tor (Pal *et al.*, 2010) 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. Une autre solution consiste alors à définir des mécanismes, certes manipulables, mais pour lesquels décider si une manipulation est efficace est algorithmiquement complexe (Brandt *et al.*, 2012 ; Barberà, 2010 ; Faliszweki *et al.*, s. d.). Cependant, comme remarqué par (Walsh, 2011), il s'agit uniquement d'une complexité au pire cas pour les agents malveillants et il existe des manipulations faciles à mettre en œuvre en pratique – comme par exemple sur les jeux hédoniques (Vallée *et al.*, 2014).

Dans ce contexte, nous proposons un protocole pour prévenir les auto-promotions et les diffamations 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 agents 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 considérons la définition d'un système de réputation proposée par (Cheng, Friedman, 2005). Cette définition permet de capturer la classe des systèmes de réputation dits personnalisés tels que Troika (Abdul-Rahman, Hailes, 1997), Histos (Zacharia, Maes, 2000), BetaReputation (Josang, Ismail, 2002) ou FlowTrust (Cheng, Friedman, 2005). Personnalisé signifie ici que la réputation d'un agent dépend de celui qui l'évalue.

DÉFINITION 1. — *Soit $\mathcal{G} = (V, E)$ un graphe orienté où V est l'ensemble des agents $\{a_1 \dots a_n\}$ et $E \subseteq V \times V$ une relation d'interaction étiquetée par une valeur de confiance $c : E \longmapsto [0, 1]$. La réputation d'un agent a_j selon un agent a_i est donnée par une fonction $f_{\mathcal{G}} : V \times V \longmapsto [0, 1]$ où :*

$$f_{\mathcal{G}}(a_i, a_j) = \bigoplus_{P \in \mathcal{P}_{ij}} \odot(P)$$

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

EXEMPLE 2. — Le système de réputation FlowTrust proposé par (Cheng, Friedman, 2005) est défini par $\odot = \prod$ et $\oplus = \max$. Ainsi, la réputation d'un agent a_j selon un agent a_i est donnée par la valeur maximale des produits des valeurs de confiance parmi tous les chemins disjoints entre a_i et a_j. □

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

DÉFINITION 3. — *Soit $d_i : V \longmapsto \{0, 1\}$ une fonction de décision où 0 signifie que a_i ne fait pas confiance à a_j et 1 signifie que a_i fait confiance à a_i. d_i est une fonction de seuil $\theta_i \in [0, 1]$ sur $f_{\mathcal{G}}(a_i, a_j)$ telle que $d_i(a_j) = 0$ si $f_{\mathcal{G}}(a_i, a_j) \leq \theta_i$ et $d_i(a_j) = 1$ sinon.*

EXEMPLE 4. — Reprenons l'exemple 2. Dans FlowTrust, $\theta_i = 0{,}5$. Ainsi, un agent a_i a confiance en un agent a_j si $f_{\mathcal{G}}(a_i, a_j) > 0{,}5$. □

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_i est une fonction de décision privée et subjective. C'est pourquoi d_i peut être différente pour chaque agent. Quoi qu'il en soit, quelle que soit la définition de d_i, 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_i et a_j leur valeur de confiance envers a_j. Ces agents, s'ils sont

malveillants, peuvent alors collectivement manipuler le système en rapportant à a_i de fausses valeurs de confiance pour a_j.

3.2. Caractériser les manipulations

Dans cet article, nous considérons deux types de manipulations : les auto-promotions où des agents malveillants se soutiennent mutuellement et les diffamations où des agents malveillants rapportent de faibles valeurs de confiance envers des agents honnêtes (Hoffman *et al.*, 2009). Dans les deux cas, nous considérons que les agents malveillants désirent tromper tous les autres agents du système. Cette hypothèse correspond à des manipulations effectuées sur des systèmes réels, comme les manipulations du PageRank (Sheldon, 2010) ou du système de réputation d'eBay (Dini, Spagnolo, 2009) ou encore le resquillage sur des réseaux pair-à-pair (Sirivianos *et al.*, 2001). 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.

3.2.1. Auto-promotion

Une auto-promotion est définie comme suit.

DÉFINITION 5. — *$\forall\ a_i \in V$ qui interroge un agent a_k sur sa valeur de confiance $c(a_k, a_j)$ envers un agent a_j, si a_k promeut a_j alors $c(a_k, a_j) > \theta_k$.*

Cette définition implique deux homogénéités. La première est une homogénéité de comportement des agents malveillants envers les autres 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 homogénéité 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 6. — *Soit a_i un agent désirant calculer la réputation d'un agent a_j. La* fonction de suspicion d'auto-promotion *de a_i est une fonction de la forme $M_i^{\uparrow} : V \times V \longmapsto [0,1]$ telle que $M_i^{\uparrow}(a_k, a_l) = c(a_k, a_l) \times c(a_l, a_k)$ si $c(a_k, a_j) > \theta_i$ et $c(a_l, a_j) > \theta_i$. Sinon $M_i^{\uparrow}(a_k, a_l) = 0$.*

Ainsi, plus $M_i^{\uparrow}(a_k, a_l)$ est élevée, plus il est possible que les agents a_k et a_l soient en collusion. Il convient de noter qu'une telle fonction n'est qu'une heuristique repré-

sentant une connaissance sur la caractérisation des auto-promotions. 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.

3.2.2. Diffamation

Suivant la même méthodologie, nous définissons une diffamation comme suit :

DÉFINITION 7. — *$\forall\ a_i \in V$ qui interroge un agent a_k sur sa valeur de confiance $c(a_k, a_j)$ envers un agent a_j, si a_k diffame a_j alors $c(a_k, a_j) < \theta_k$.*

La diffamation présente deux homogénéités de comportement 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 et ils rapportent tous une faible confiance envers tous les agents honnêtes du système.

DÉFINITION 8. — *Soit a_i un agent désirant calculer la réputation d'un agent a_j et a_s un agent connu de a_i comme étant inconnu de a_k. La* fonction de suspicion de diffamation *de l'agent a_i est une fonction de la forme $M_i^{\downarrow} : V \longmapsto [0, 1]$ telle que $M_i^{\downarrow}(a_k) = 1 - c(a_k, a_s)$ si $c(a_k, a_j) < \theta_i$ et $c(a_k, a_s) < \theta_i$. Sinon $M_i^{\downarrow}(a_k) = 0$.*

Plus $M_i^{\downarrow}(a_k)$ est élevée, plus il est possible que l'agent a_k diffame les agents honnêtes. Remarquons que cette fonction nécessite a priori plus de connaissance de la part de a_i sur le système car il doit connaître des agents inconnus de a_k et le savoir. Nous verrons dans la section suivante comment résoudre ce problème.

À partir de ces heuristiques, un agent honnête peut demander à plusieurs agents témoins leurs valeurs de confiance mutuelles afin de déterminer s'il y a une auto-promotion ou demander à un agent témoin sa valeur de confiance en plusieurs agents distincts dans le cas des diffamations. Toutefois sachant cela, les agents malveillants peuvent mentir sur leurs valeurs 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, si ce sont des agents 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 soit, dans le cas d'une auto-promotion, par plusieurs agents apparemment distincts (l'agent honnête et ses agents

Sybil), soit, dans le cas d'une diffamation, à propos d'un agent connu seulement de l'agent honnête (un agent Sybil). Parmi les agents témoins, les agents malveillants caractérisés par les définitions 5 et 7 sont alors incités à révéler leur confiance mutuelle ou à diffamer afin de tromper tous les autres agents du système. L'algorithme 1 présente les principales étapes du protocole.

Algorithme 1 Protocole

1: calculer $f_{\mathcal{G}}(a_i, a_j)$
2: **si** $d_i(a_j) = 1$ **alors**
3: $W' = \{a_k \in W : c(a_k, a_j) > \theta_i\}$
4: **pour tout** $a_k, a_l \in W'$ $(k \neq l)$ **faire**
5: générer deux agents Sybil a'_i et a''_i
6: a'_i demande à a_k sa valeur $c(a_k, a_l)$
7: a''_i demande à a_l sa valeur $c(a_l, a_k)$
8: calculer $M_i^{\uparrow}(a_k, a_l)$
9: **fin pour**
10: **sinon**
11: $W' = \{a_k \in W : c(a_k, a_j) < \theta_i\}$
12: **pour tout** $a_k \in W'$ **faire**
13: générer deux agents Sybil a'_i et a''_i
14: a'_i demande à a_k sa valeur $c(a_k, a''_i)$
15: calculer $M_i^{\downarrow}(a_k)$
16: **fin pour**
17: **fin si**
18: **retourne** réviser $f_{\mathcal{G}}(a_i, a_j)$

Tout d'abord, a_i utilise le système de réputation afin de calculer la réputation de a_j (ligne 1). Ainsi, a_i obtient un ensemble W de témoins pour a_j et peut calculer $d_i(a_j)$. Comme remarqué dans (Conitzer, Yokoo, 2010), a_j peut être un agent honnête diffamé si $d_i(a_j) = 0$ ou a_j peut être un agent malveillant faisant de l'auto-promotion si $d_i(a_j) = 1$. 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_i à prendre la « bonne » décision, c'est-à-dire faire confiance s'il y a eu de l'auto-promotion ou ne pas faire confiance s'il y a eu de la diffamation.

En cas d'auto-promotion (ligne 2), a_i sélectionne le sous-ensemble de témoins qui ont confiance dans a_j (ligne 3). Pour chaque couple de témoins a_k et a_l, a_i génère deux agents Sybil (ligne 4 et 5) qui vont interroger les potentiels agents malveillants. Ces agents Sybil demandent à a_k s'il a confiance dans a_l (ligne 6) et inversement (ligne 7). Si a_k (respectivement a_l) est honnête, il répondra honnêtement ; si a_k est malveillant, il sera incité à répondre selon la définition 5.

En cas de diffamation (ligne 10), a_i sélectionne le sous-ensemble de témoins qui n'ont pas confiance dans a_j (ligne 11). Pour chaque témoin a_k, a_i génère deux agents Sybil (lignes 12 et 13). Un des agents Sybil demande à a_k s'il a confiance dans le deuxième agent Sybil (ligne 14). Si a_k est honnête, il répondra ne pas connaître l'agent

Sybil représenté par une valeur de confiance par défaut ; si a_k est malveillant, il sera incité à répondre selon la définition 7.

Une fois que a_i a obtenu l'ensemble des valeurs de suspicion pour les témoins (lignes 8 et 15), il les utilise pour réviser $f_{\mathcal{G}}(a_i, a_j)$ (ligne 18).

3.4. Réviser la valeur de réputation

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

Algorithme 2 Révision

1: $\mathcal{G}' \leftarrow \mathcal{G}$
2: **pour tout** $a_k \in W'$ **faire**
3: **si** $d_i(a_j) = 1$ **alors**
4: $M_i \leftarrow \max_{a_l \in W'} M_i^{\uparrow}(a_k, a_l)$
5: **sinon**
6: $M_i \leftarrow M_i^{\downarrow}(a_k)$
7: **fin si**
8: $V' \leftarrow V' \setminus \{a_k\}$ avec une probabilité M_i
9: **fin pour**
10: **retourne** $f_{\mathcal{G}'}(a_i, a_j)$

Le mécanisme procède comme suit : l'agent a_i retire du graphe d'interaction chaque témoin a_k avec une probabilité égale à la plus haute valeur de suspicion qui lui a été attribuée. 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

Comme ce protocole est une couche implantée au-dessus d'un système de réputation existant, il accroît nécessairement 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 9. — *Le coût de communication du protocole donné par l'algorithme 1 est* $\mathcal{O}(4.|W-1|^2)$.

PREUVE 10. — Chaque fois que la réputation d'un agent a_j est calculée par un agent a_i, chaque paire de témoins distincts dans W est interrogée, ce qui génère quatre messages (deux questions puis deux réponses) en cas d'auto-promotion et deux messages

(une question puis une réponse) en cas de diffamation. Ainsi trivialement, dans le pire des cas le protocole ajoute $4.|W-1|^2$ nouveaux messages. ■

La nouvelle réputation $f_{\mathcal{G}'}(a_i, a_j)$ calculée par ce protocole n'identifie pas formellement les agents malveillants mais atténue l'influence des témoignages des agents présentant les signes d'une manipulation. 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 tente une manipulation lorsqu'un agent tiers l'interroge. 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, le tableau 1 représente, du point de vue de l'agent malveillant, le dilemme comme un jeu à somme nulle sous forme stratégique.

DÉFINITION 11. — *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.*

Tableau 1. Jeu sous forme stratégique

Agent	**Manipuler**	**Ne pas manipuler**
Honnête	$(1-\delta)g$	0
Sybil	$-\delta p$	δg

Informellement, le dilemme est le suivant. D'un côté, si l'agent malveillant manipule un agent honnête, il manipule effectivement cet agent et est donc récompensé pour atteindre son objectif. Cette récompense correspond au gain à manipuler le système. De même, si l'agent malveillant ne manipule pas un agent Sybil alors il manipule en fait l'agent honnête qui a généré la Sybil et obtient donc la même récompense car il est parvenu à manipuler le système. D'un autre côté, si l'agent malveillant ne manipule pas un agent honnête, son gain est nul car il ne pourra pas manipuler le système. Dans le dernier cas, si l'agent malveillant manipule 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.

Notons que le protocole donné par l'algorithme 1 utilise des agents Sybil différents pour chaque témoin à contre-interroger. Ainsi, chaque jeu est indépendant des autres et chaque agent malveillant – même au sein d'un coalition – doit décider pour son jeu seul. Un agent malveillant rationnel va donc chercher à partir de cette matrice une stratégie (manipuler ou ne pas manipuler) qui maximisera son gain. Or, cette matrice correspond à un 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 (Dang, 2009). 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 M l'action de manipuler et $\bar{M}$ celle de ne pas manipuler. 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 $\pi_a = \langle \sigma(M) = (1-m), \sigma(\bar{M}) = m \rangle$ 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 12. — *L'*équilibre de Nash en stratégie mixte *du jeu caractérisé par le tableau 1 est :*

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

PREUVE 13. — L'utilité espérée de π_a en fonction de m et δ est :

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

Comme l'agent malveillant désire maximiser $u(\pi_a)$, les racines des dérivées partielles de $u_{\pi_a}(m,\delta)$ 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 $u(\pi_a)$, les racines des dérivées partielles de $-u_{\pi_a}(m,\delta)$ 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, il découle de la proposition 12 que, 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

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 ne pas manipuler un agent honnête revient à abandonner la manipulation et manipuler un agent Sybil revient à être sanctionné. En conséquence, une manipulation réussie est définie comme suit.

DÉFINITION 14. — *Une manipulation est* réussie *si et seulement si l'agent malveillant manipule l'agent honnête (stratégie M) puis manipule l'agent Sybil qui a été généré (stratégie M).*

Tableau 2. Occurrences des stratégies jointes

Agent	**Manipuler**	**Ne pas manipuler**
Honnête	$m(1-m)$	m^2
Sybil	$(1-m)^2$	$m(1-m)$

Si 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 alors le tableau 2 donne les probabilités d'occurrence de ces stratégies jointes pour la stratégie mixte déterminée précédemment (et $m = 1 - \delta$).

PROPOSITION 15. — *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 12, 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 peut conduire les agents malveillants à être identifiés comme suspects s'ils suivent une stratégie pure qui consiste à toujours manipuler. De même si les agents malveillants jouent une stratégie pure qui consiste à ne jamais manipuler alors ils ne posent aucun problème au système de réputation. Ceci place alors 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 des systèmes FlowTrust (Cheng, Friedman, 2005) et BetaReputation (Josang, Ismail, 2002). Nous avons ensuite 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

Dans ces expérimentations, 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 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 $10\,000$ 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 problème 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 manipuler (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 $10\,000$ simulations.

Nous avons choisi de nous comparer à deux systèmes : FlowTrust présenté dans l'exemple 2 car il est robuste aux manipulations lorsque les agents ont une vision globale du système (bien que peu informatif) et BetaReputation (Josang, Ismail, 2002) car il s'agit d'un système de référence dans la littérature. En effet, BetaReputation est un système asymétrique personnalisé qui utilise un fonction de densité beta pour estimer la probabilité qu'un agent exhibe un bon comportement dans le futur.

De plus, pour nous placer dans le cadre le moins favorable, nous considérons que les agents malveillants n'ont pas de pénalité de collusion ($p = 0$). Sur chacune des figures présentant les résultats, les courbes pleines représentent les manipulations réussies sans notre protocole, les courbes en tirets (gras) représentent les manipulations réussies sous stratégie pure tandis que les courbes en pointillets illustrent la stratégie mixte.

5.2. Du nombre d'agents malveillants

Afin d'étudier l'influence du nombre d'agents malveillants, nous avons fait varier la proportion de ces derniers entre 10 et 50 % du système tandis qu'un agent honnête

évalue 5 agents sélectionnés aléatoirement. Les résultats pour les auto-promotions sont donnés en figure 1 et en figure 2 pour les diffamations.

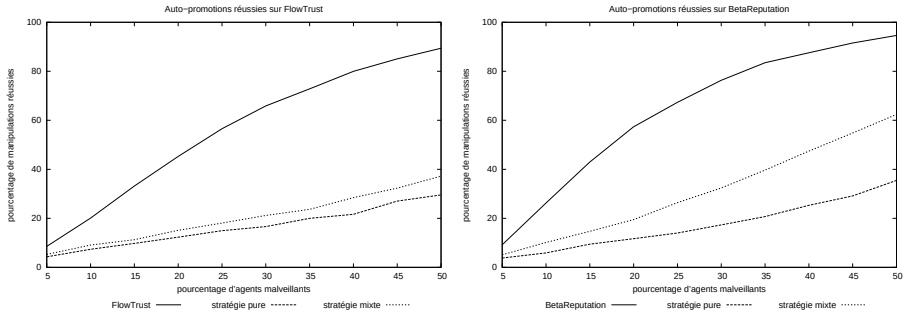

Figure 1. Auto-promotions réussies selon le nombre d'agents malveillants

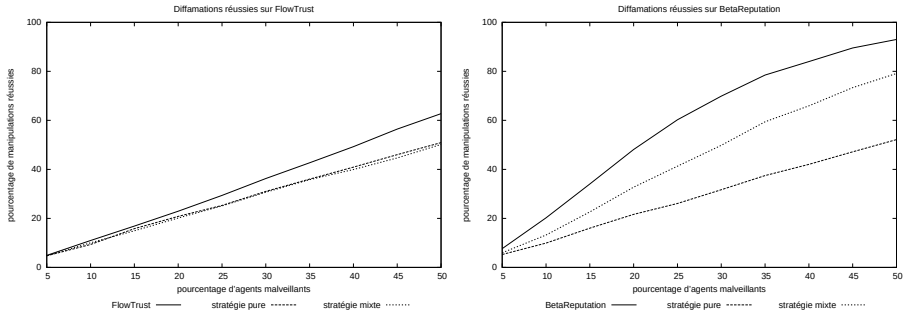

Figure 2. Diffamations réussies selon le nombre d'agents malveillants

Sans surprise, dans tous les cas, le nombre de manipulations réussies augmente lorsque le nombre d'agents malveillants augmente. Dans le cas des auto-promotions, la stratégie mixte réduit en moyenne de 55 % le nombre de manipulations réussies sur les deux systèmes tandis que la stratégie pure les réduit de 62 % sur FlowTrust et 72 % sur BetaReputation. Dans le cas des diffamations, FlowTrust qui est déjà robuste ne voit les manipulations réduites par notre protocole que de 18 % en stratégie pure et 13 % en stratégie mixte. Notre protocole est bien plus efficace sur BetaReputation où les diffamations sont réduites de 50 % en stratégie pure et 25 % en stratégie mixte. En conséquence, notre protocole est efficace bien que cette efficacité dépend en partie de la robustesse initiale du système de réputation. Toutefois, nos résultats théoriques sont confirmés : les agents malveillants doivent jouer une stratégie mixte pour maximiser le nombre de manipulations réussies. Dans tous les cas, 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 faire le choix de réduire la force de sa manipulation afin d'éviter d'être suspecté par les agents honnêtes. Pour cela, la coalition malveillante va présenter des valeurs de confiance plus faible en cas d'auto-promotion ou plus forte en cas de diffamation afin de réduire la valeur de 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 $[0{,}5, 1]$ pour les auto-promotions et l'intervalle $[0{,}05, 0{,}5]$ pour les diffamations. Les résultats pour les auto-promotions sont donnés en figure 3 et en figure 4 pour les diffamations.

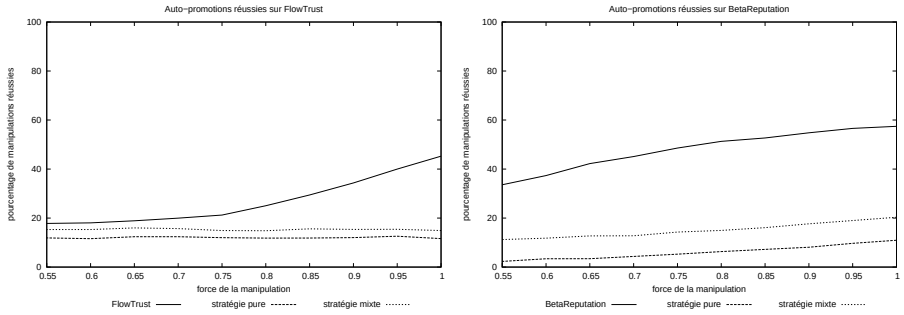

Figure 3. Auto-promotions réussies selon la force de la manipulation

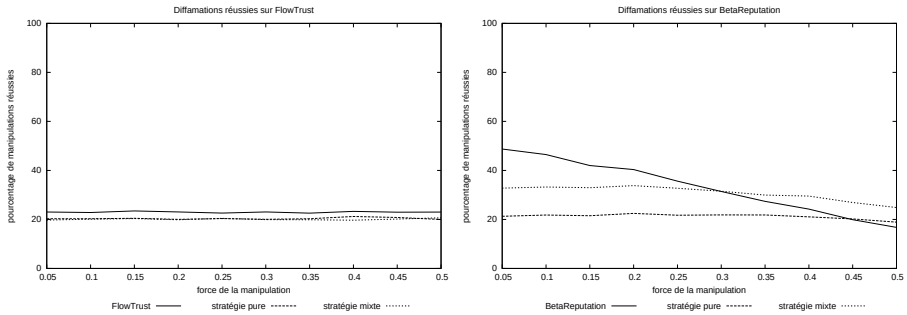

Figure 4. Diffamation réussies selon la force de la manipulation

Sans surprise encore, sur les deux systèmes, le nombre d'auto-promotions réussies sans notre protocole augmente au fur et à mesure que les agents malveillants augmente la force de leur auto-promotion. Nous pouvons remarquer que notre protocole est très efficace sur FlowTrust car, quelle que soit la force des auto-promotions, le nombre de manipulations réussies reste autour de 15 % pour la stratégie mixte et 5 % pour la stratégie pure. Ces valeurs restent en moyenne les mêmes sur BetaReputation.

Notons que, dans les cas de diffamation sur FlowTrust, notre approche n'est que peu efficace en raison de la robustesse initiale du système mais ne dégrade en aucun cas les performances. Enfin, nous pouvons remarquer que notre approche produit 20 % de manipulations supplémentaires sur BetaReputation dans le cas où la force de la diffamation est faible (de $0{,}3$ à $0{,}5$) : un certain nombre de faux positifs retire des agents honnêtes et permet aux agents malveillants d'obtenir les meilleures valeurs de réputation.

Dans tous les cas, et de manière surprenante, la performance absolue de notre protocole reste stable quelle que soit la valeur exprimée par les agents malveillants et quel que soit le système considéré. Ainsi, notre protocole peut non seulement prévenir les manipulations mais il ne semble pas pouvoir être manipulé en retour. Cependant, comme l'algorithme 2 retire les suspects stochastiquement, notre approche peut dégrader les performances de certains systèmes de réputation si les agents malveillants adoptent un comportement proche des agents honnêtes.

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 faisons 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 pour les auto-promotions sont donnés en figure 5 et en figure 6 pour les diffamations.

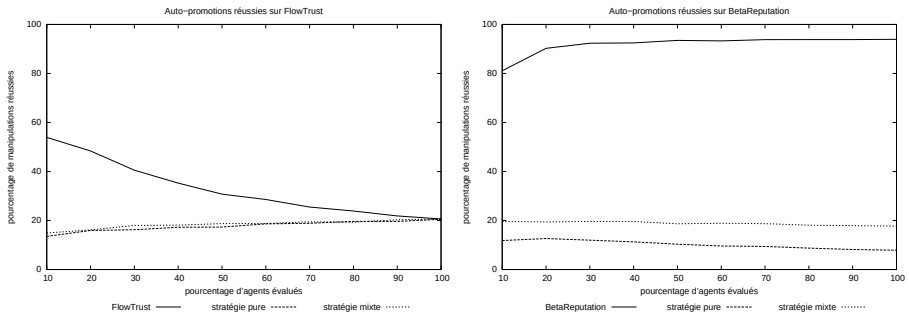

Figure 5. Auto-promotions réussies selon le nombre d'agents évalués

De manière surprenante encore, nous pouvons remarquer que la performance absolue de notre protocole reste stable quel que soit le nombre d'agents évalués et quel que soit le système considéré. Sur FlowTrust, en raison de la robustesse du système, le nombre d'auto-promotions réussies sans notre protocle décroît au fur et à mesure que le nombre d'agents évalués augmente. Le nombre de diffamations reste stable quant à lui. Dans ces deux cas, nous pouvons aussi remarquer que la stratégie mixte est

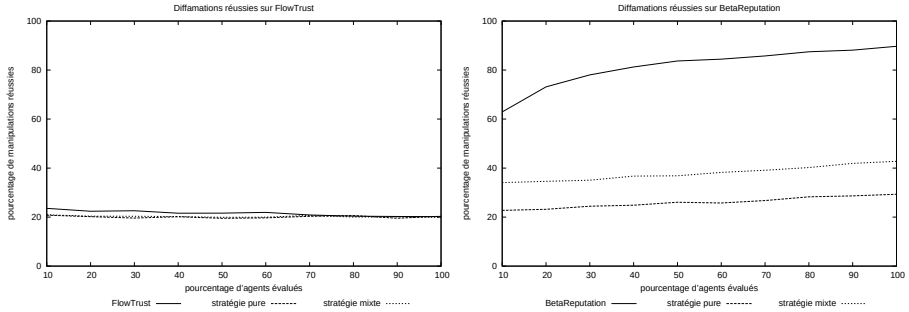

Figure 6. Diffamations réussies selon le nombre d'agents évalués

aussi efficace que la stratégie pure : les agents malveillants ne peuvent pas augmenter le nombre de manipulations réussies en jouant sur leur stratégie. De plus, les performances de notre protocole convergent pour atteindre celles du système de réputation originel. Toutefois, rappelons encore que FlowTrust est robuste aux manipulations lorsqu'il est centralisé. Or, dans un système réel, il n'est pas toujours possible de faire cette hypothèse. Si nous considérons BetaReputation, nous obtenons les mêmes résultats de stabilité mais en permettant une réduction moyenne des auto-promotions de 88 % et 78 % pour la stratégie pure et mixte respectivement, et une réduction de 67 % et 52 % des diffamations. Aussi, notre protocole reste aussi très efficace dans ce cas.

5.5. Limites de l'approche

Bien que nos expérimentations suggèrent que notre protocole est efficace pour limiter les manipulations, il peut toutefois se heurter à la vulnérabilité intrinsèque du système de réputation sur lequel il est implanté. Par exemple, le système EigenTrust est un système de réputation asymétrique global fondé sur le même principe que le PageRank : soit un graphe pondéré représentant les liens de confiance entre agents, la réputation d'un agent est la probabilité qu'une marche aléatoire passe par le nœud de cet agent (Kamvar *et al.*, 2003). Sur EigenTrust, l'auto-promotion et la diffamation sont équivalentes car chaque agent distribue une même quantité de confiance sur un sous-ensemble d'agents du système. Ainsi, privilégier certains pour les promouvoir revient à priver d'autres agents de confiance et donc à les diffamer. Il suffit alors qu'un agent au sein d'une coalition réussisse une manipulation pour qu'il puisse automatiquement obtenir une haute valeur de réputation pour l'ensemble de sa coalition, valeur d'autant plus haute que la taille de la coalition est grande (Cheng, Friedman, 2006). Or comme notre approche consiste à retirer stochastiquement des agents suspects, il suffit qu'un agent malveillant ne soit pas détecté pour que notre protocole soit sans effet. La figure 7 met en lumière ces constats.

Dans cette expérience, nous avons fait varier la proportion d'agents malveillants entre 10 et 50 % tandis qu'un agent honnête évalue 5 agents du système. Trivialement, plus il y a d'agents malveillants, plus le nombre de manipulations réussies augmente.

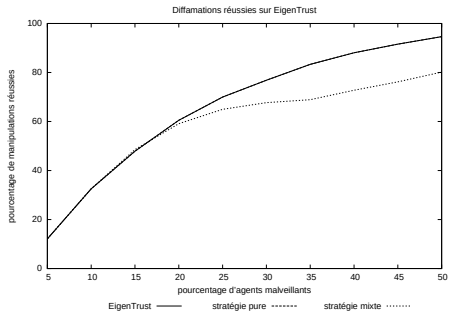

Figure 7. Manipulations réussies sur EigenTrust

Nous pouvons remarquer que la stratégie pure produit autant de manipulations que le protocole original. Notre approche est donc sans effet. De manière surprenante, la stratégie mixte n'est plus la stratégie optimale pour les agents malveillants qui voient leurs manipulations réduites de 8 % environ. Cela est du au fait que lorsqu'un agent joue la stratégie qui consiste à ne pas manipuler, il réduit la taille de la coalition et par conséquent la valeur maximale de réputation que les agents malveillants peuvent obtenir. Ainsi, certaines manipulations qui auraient pu réussir avec une stratégie pure sont inefficaces avec une stratégie mixte.

Remarquons toutefois que, d'une part, notre approche ne dégrade pas les performances du système de réputation et que, d'autre part, EgeinTrust est un système global (où la réputation d'un agent ne dépend pas de celui qui l'évalue) qui n'est pas capturé par le modèle de (Cheng, Friedman, 2005) sur lequel nous nous sommes fondés.

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 manipulation.

Dans ce contexte, nous proposons un protocole mettant en œuvre un dilemme pour détecter et prévenir les auto-promotions et les diffamations dans un système de réputation personnalisé. 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 soit à révéler des valeurs de confiance caractéristiques d'une manipulation, utilisées pour détecter des collusions, soit à ne pas manipuler afin de dissimuler leur présence. Notre analyse théorique du protocole montre que les agents malveillants, s'ils sont rationnels, doivent jouer une stratégie mixte et, en conséquence, nécessairement 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 confiance entre agents honnêtes sont correlées et qu'il existe des *clusters* d'agents pouvant augmenter le nombre de faux positifs retournés. Dans un deuxième temps, il nous semble important d'étendre notre protocole aux systèmes de réputation globaux et en particulier à EigenTrust sur lequel les manipulations se comportent différemment de notre caractérisation.

Répondre à cette question nécessite de se 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 donc 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 (Yu *et al.*, 2010) afin d'améliorer la performance du protocole. Une autre voie consiste à considérer 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.

Bibliographie

Abdul-Rahman A., Hailes S. (1997). Using recommendations for managing trust in distributed systems. In *3rd IEEE Malaysia International Conference on Communication.*

Altman A., Tennenholtz M. (2005). Ranking systems: the PageRank axioms. In *6th ACM Conference on Electronic Commerce*, p. 1-8.

Altman A., Tennenholtz M. (2010). An axiomatic approach to personalized ranking systems. *Journal of the Association for Computing Machinery*, vol. 57, n° 4, p. 1-26.

Bachrach Y., Elkind E. (2008). Divide and conquer: false-name manipulations in weighted voting games. In *7th International Conference on Autonomous Agents and Multiagent System*, p. 975-982.

Barberà S. (2010). *Strategy-proof social choice*. Rapport technique. Universitat Autònoma de Barcelona.

Bonnet G. (2012). A protocol based on a game-theoretic dilemma to prevent malicious coalitions in reputation systems. In *28th European Conference on Artificial Intelligence*, p. 187-192.

Bonnet G. (2013). Un protocole fondé sur un dilemme pour se prémunir des collusions dans les systèmes de réputation. In *21es Journées Francophones sur les Systèmes Multi-Agents*, p. 9-18.

Borisov N. (2006). Computational puzzles as Sybil defenses. In *6th International Conference on Peer-to-Peer Computing*, p. 171-176.

Brandt F., Conitzer V., Endriss U. (2012). Computational social choice. In G. Weiss (Ed.), *Multiagent systems: A modern approach to distributed artificial intelligence*, p. 213-283. MIT Press.

Castro M., Drusche P., Ganesh A., Rowstron A., Wallach D.-S. (2002). Secure routing for structured peer-to-peer overlay networks. In *5th Symposium on Operating Systems Design and Implementation*, p. 299-314.

Cheng A., Friedman E. (2005). Sybilproof reputation mechanisms. In *3rd ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems*, p. 128-132.

Cheng A., Friedman E. (2006). Manipulability of PageRank under Sybil strategies. In *1st Workshop of Networked Systems*.

Cholez T., Chrisment I., Festor O. (2010). Efficient DHT attack mitigation through peers' ID distribution. In *24th International Workshop on Hot Topics in Peer-to-Peer Systems*, p. 1-8.

Conitzer V., Immorlica N., Letchford J., Munagala K., Wagman L. (2010). False-name-proofness in social networks. In *6th International Conference on Internet and Network Economics*, p. 1-17.

Conitzer V., Yokoo M. (2010). Using mechanism design to prevent false-name manipulations. *Artificial Intelligence Magazine*, vol. 31, n° 4, p. 65-77.

Dang T. (2009). *Gaming or guessing: mixing and best-responding in matching pennies*. Rapport technique. University of Arizona.

Dini F., Spagnolo G. (2009). Buying reputation on eBay: do recent changes help? *International Journal of Electronic Business*, vol. 7, n° 6, p. 581-598.

Douceur J.-R. (2002). The Sybil attack. In *1st International Workshop on Peer-to-Peer Systems*, p. 251-260.

Faliszweki P., Hemaspaandra E., Hemaspaandra L. (s. d.). Using complexity to protect elections. *Communications of the ACM*, vol. 53, n° 11.

Hoffman K., Zage D., Nita-Rotaru C. (2009). A survey of attack and defense techniques for reputation systems. *ACM Computing Survey*, vol. 42, n° 1, p. 1-31.

Josang A., Ismail R. (2002). The Beta reputation system. In *15th Bled Electronic Commerce Conference*.

Kamvar S.-D., Schlosser M.-T., Garcia-Molina H. (2003). The EigenTrust algorithm for reputation management in P2P networks. In *12th International Conference on World Wide Web*, p. 640-651.

Levine B.-N., Shields C., Margolin N.-B. (2006). *A survey of solutions to the Sybil attack*. Rapport technique. University of Massachussets Amherst.

Liao X., Hao D., Sakurai K. (2011). Classification on attacks in wireless ad hoc networks: A game theoretic view. In *7th International Conference on Networked Computing and Advanced Information Management*, p. 144-149.

Margolin N.-B., Levine B.-N. (2007). Informant: detecting Sybils using incentives. In *11th International Conference on Financial Cryptography*, p. 192-207.

Pal A.-K., Nath D., Chakreborty S. (2010). A discriminatory rewarding mechanism for Sybil detection with applications to Tor. *World Academy of Science, Engineering and Technology*, vol. 39, p. 29-36.

Sheldon D. (2010). *Manipulation of PageRank and collective hidden Markov models*. Thèse de doctorat non publiée, Cornell University.

Sirivianos M., Park J.-H., Cheng R., Yang X. (2001). *Free-riding in BitTorrent networks with the large view exploit*. Rapport technique. California Irvine.

Vallée T., Bonnet G., Zanuttini B., Bourdon F. (2014). A study of Sybil manipulations in hedonic games. In *13th International Conference on Autonomous Agents and Multiagent Systems*, p. 21-28.

Viswanath B., Post A., Gummadi K.-P., Mislove A. (2010). An analysis of social network-based Sybil defenses. In *16th SIGCOMM Conference*, p. 363-374.

Von Ahn L., Blum M., Hopper N., Langford J. (2003). CAPTCHA: Using hard AI problems for security. In *22nd International Conference on the Theory and Applications of Cryptographic Techniques*, p. 294-311.

Walsh T. (2011). Is computational complexity a barrier to manipulation? *Annals of Mathematics and Artificial Intelligence*, vol. 62, p. 7-26.

Yokoo M., Sakurai Y., Matsubara S. (2004). The effect of false-name bids in combinatorial auctions: new fraud in Internet auctions. *Game and Economic Behavior*, vol. 46, p. 174-188.

Yu H., Gibbons P.-B., Kaminsky M., Feng X. (2010). SybilLimit: a near-optimal social network defense against Sybil attacks. *IEEE/ACM Transactions on Networking*, vol. 18, n° 3, p. 885-898.

Zacharia G., Maes P. (2000). Trust management through reputation mechanisms. *Applied Artificial Intelligence*, vol. 14, p. 881-907.