On the Impact of Witness-Based Collusion in Agent Societies

Amirali Salehi-Abari and Tony White

School of Computer Science, Carleton University,
1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6, Canada
{asabari,arpwhite}@scs.carleton.ca

Abstract. In ways analogous to humans, autonomous agents require trust and reputation concepts in order to identify communities of agents with which to interact reliably. This paper defines a class of attacks called witness-based collusion attacks designed to exploit trust and reputation models. Empirical results demonstrate that unidimensional trust models are vulnerable to witness-based collusion attacks in ways independent multidimensional trust models are not. This paper analyzes the impact of the proportion of witness-based colluding agents on the society. Furthermore, it demonstrates that here is a need for witness interaction trust to detect colluding agents in addition to the need for direct interaction trust to detect malicious agents. By proposing a set of policies, the paper demonstrates how learning agents can decrease the level of encounter risk in a witness-based collusive society.

1 Introduction

By analogy with human societies in which trust is one of the most crucial concepts driving decision making and relationships, *trust* is indispensable when considering interactions among individuals in artificial societies such as are found in e-commerce marketplaces. According to Jarvenpaa et al. [5], trust is an essential aspect of any relationship in which the trustor does not have direct control over the actions of a trustee, the decision is important, and the environment is uncertain.

We use the experience that we gain in interacting with others to judge how they will perform in similar situations. However, when we need to assess our trust in someone of whom we have no direct personal experience, we often ask others regarding their personal experience with this individual. This collective opinion of others regarding the specific individual is known as an individual's reputation.

As reputation and trust have recently received considerable attention in many diverse domains several definitions exist. Mui et al. define trust as "a subjective expectation an agent has about another's future behavior based on the history of their encounters" [8]. While trust definitions focus more on the history of agents' encounters, reputation is based on the aggregated information from other individuals. For instance, Sabater and Sierra [10] declared that "reputation is the opinion or view of someone about something".

J.-J. Yang et al. (Eds.): PRIMA 2009, LNAI 5925, pp. 80–96, 2009.

Sabater and Sierra [11] categorized computational trust and reputation models based on various intrinsic features. From their perspective, a trust and reputation model can be cognitive or game-theoretical in terms of its conceptual model. Trust and reputation models might use different sources of information such as direct experiences, witness information, sociological information and prejudice. Direct experience and witness information are pertinent to this paper. Direct experiences deal with agent-to-agent interactions while witness information is information that comes from members of the community about others. Trust and reputation models vary in terms of individual behavior assumptions; in some models, cheating behaviors and malicious individuals are not considered at all whereas in others possible cheating behaviors are taken into account. There are many computational models of trust, a review of which can be found in [11].

Regret [10] is a decentralized trust and reputation system oriented to e-commerce environments. The system takes into account three different sources of information: direct experiences, information from third party agents and social structures. Yu and Singh [16] developed an approach for social reputation management, in which they represented an agent's ratings regarding another agent as a scalar and combined them with testimonies using combination schemes similar to certainty factors. Huynh et al. [4] introduced a trust and reputation model called FIRE that incorporates interaction trust, role-based trust, witness reputation, and certified reputation to provide a trust metric.

Most recently, researchers have identified the existence of cheaters (exploitation) in artificial societies employing trust and reputation models [6,12], and the existence of inaccurate witnesses [15,2], and [17]. Kerr and Cohen [6] examined the security of several e-commerce marketplaces employing a trust and reputation system. To this end, they proposed several attacks and examined their effects on each marketplace. Unfortunately, Kerr and Cohen assume that buyers are honest in the witness information provided to one another and consequently do not consider collusion attacks. Salehi-Abari and White [12] introduced and formally modeled the con-man attack and demonstrated the vulnerability of several trust models against this attack. This work also did not consider any collusion attacks.

There are few trust models which consider the existence of an adversary in providing witness information and present solutions for dealing with inaccurate reputation. TRAVOS [14] models an agent's trust in an interaction partner. Trust is calculated using probability theory that takes account of past interactions and reputation information gathered from third parties while coping with inaccurate reputations. Yu and Singh [17] is similar to TRAVOS, in that it rates opinion source accuracy based on a subset of observations of trustee behavior.

To our knowledge, there is no formal model of witness-based collusion and analysis of the level of encounter risk for trust-aware agents in witness-based collusive societies. In a witness-based collusion attack, an unreliable witness provider in spite of being cooperative in its direct interactions provides high ratings for other malicious agents (other members of the colluding group), thus resulting in motivating the victim agent to interact with them. This lack of

study on witness-based collusion attacks motivates the work reported in this paper. This paper expands on a preliminary workshop paper [13].

Our contributions include the introduction of witness-based collusion attacks; a formal agent-based model of this attack class; an analysis of the impact of this attack class on agent societies, and on the level of encounter risk for trust-aware individuals; and a proposal for strategies of trust-aware agents to deal with this attack class.

The remainder of this paper proceeds as follows. Before describing the Witness-based Collusion Attack in Section 3, we discuss the environment model of agents in Section 2. We describe the agent model in Section 4, and experiments in Section 5. Finally, conclusions and future work are explained in Section 6.

2 Environment Model

The majority of open distributed computer systems can be modeled as multi-agent systems (MAS) in which each agent acts autonomously to achieve its objectives. Autonomy is represented here by the evaluation of policies that cause changes in agent trust and reputation models and subsequent changes in societal structure. Our model incorporates heterogeneous agents interacting in a game theoretic manner. The model is described in the following 3 subsections.

2.1 Interactions

An agent interacts with a subset of all agents that are the neighbors of the given agent. Two agents are *neighbors* if both accept each other as a neighbor and interact with one another continuously. An agent maintains the *neighborhood* set which is dynamic, changing when an agent is determined to be untrustworthy or new agent interactions are required. Agents have bounded sociability as determined by the maximal cardinality of the neighborhood set. Agents can have two types of interactions with their neighbors: *Direct Interaction* and *Witness Interaction*.

Direct Interaction. Direct interaction is the most frequently used source of information for trust and reputation models [11,9]. Different fields have their own interpretation of direct interaction. For example, in e-commerce, direct interaction might be considered to be buying or selling a product, whereas in file sharing systems direct interaction is file exchange.

Witness Interaction. An agent can ask for an assessment of the trustworthiness of a specific agent from its neighbors and then the neighbors send their ratings of that agent to the asking agent. We call this asking for an opinion and receiving a rating, a Witness Interaction.

2.2 Games: IPD and GPD

Direct and witness interactions are modeled using two extensions of the Prisoner's Dilemma. The Prisoner's Dilemma is a non-zero-sum, non-cooperative, and simultaneous game in which two players may each "cooperate" with or "defect" from

the other player. In the Iterated Prisoner's Dilemma (IPD) [1], the game is played repeatedly. As a result, players have the opportunity to "punish" each other for previous uncooperative play. The IPD is closely related to the evolution of trust because if both players trust each other they can both cooperate and avoid mutual defection. We have modeled the direct interaction using IPD.

Witness Interaction is modeled by the Generalized Prisoner's Dilemma (GPD). GPD is a two-person game which specifies the general forms for an asymmetric payoff matrix that preserves the social dilemma [3]. GPD is compatible with client/server structure where one player is the client and the other one is the server in each game. The decision of the server alone determines the ultimate outcome of the interaction.

2.3 Cooperation and Defection

We define two kinds of **Cooperation** and **Defection** in our environment: (1) Cooperation/Defection in Direct Interaction (CDI/DDI) and (2) Cooperation/Defection in Witness Interaction (CWI/DWI).

CDI/DDI have different interpretations depending on the context. For example, in e-commerce, defection in an interaction can be interpreted as the agent not satisfying the terms of a contract, selling poor quality goods, delivering late, or failing to pay the requested amount of money to a seller [9]. CWI means that the witness agent provides a reliable rating for the asking agent regarding the queried agent. In contrast, DWI means that the witness agent provides an unreliable rating for the asker agent regarding the queried agent.

3 Witness-Based Collusion Attack

Collusion can be defined as a collaborative activity that gives to members of a colluding group benefits they would not be able to gain as individuals. Collusion attacks occur when one or more agents conspire together to take advantage of breaches in trust models to defraud one or more agents. It can be the case that agents in the colluding group adopt a sacrificial stance in collusion attacks in order to maximize the utility of the colluding group.

Collusion attacks often work based on the basic idea that one or more agents show themselves as trustworthy agents in one type of interaction (usually direct interaction). Afterward, they will be untrustworthy in other type of interaction (e.g., witness interaction) by providing false information in favor of other members of the colluding group. This false information usually encourages a victim to interact with members of the colluding group. The members of the colluding group will cheat the victim, if victim interacts with them.

As depicted in Figure 1, we formally define three roles in the Witness-based Collusion Attack: *victim agent*, *enticer agent*, and *malicious agent*. Enticer agents and malicious agents form the colluding group to exploit victim agents. The enticer agents show themselves trustworthy in direct interactions to victim agents and consequently they become trustworthy neighbors of victim agents. Afterward, when victim agents are looking for ratings (reputation) of malicious agent

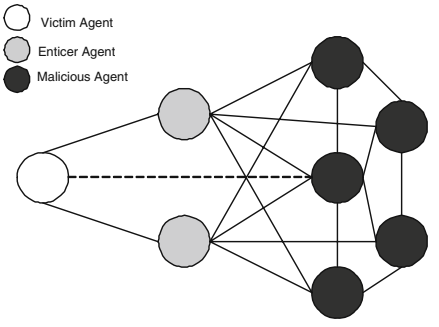


Fig. 1. Witness-based Collusion Attack

by asking from their trustworthy neighbors, the enticer agents provide high ratings for malicious agents (other members of the colluding group) in order to encourage the victim agents to interact with them, and consequently the victim agents will be exploited by them. The dashed line in Figure 1 shows the start of interaction of a victim agent with a malicious agent as a result of high ratings provided by enticer agents.

It can be observed that when the victim agent bases its assessment of witness information on the cooperations (trustworthiness) in direct interactions, this attack will be successful. In particular, the success of this attack is the result of the inappropriate assumption that whoever is cooperative (trustworthy) in direct interactions will be cooperative (trustworthy) in providing witness information regarding other agents.

It is the hypothesis of this paper that the Witness-based Collusion Attack can be prevented if the asker agent utilizes an independent multi-dimensional trust model. In this sense, the asker agent will assess the witness providers based on their cooperations in witness interactions.

4 Agent Model

This section presents two types of trust variables and a type of reputation variable that assist agents in determining with whom they should interact. Furthermore, three policy types will be presented: direct interaction policy, witness interaction policy, and connection policy which assist agents in deciding how and when they should interact with another agent.

4.1 Trust Variables

Based on the aforementioned cooperation/defection explained in section 2.3, two modeled dimensions of trust are proposed. The motivation for having two trust variables is that we believe trustworthiness has different independent dimensions. For instance, an agent who is trustworthy in a direct interaction is not necessarily trustworthy in a witness interaction.

Each trust variable is defined by $T_{i,j}(t)$ indicating the trust rating assigned by agent i to agent j after t interactions between agent i and agent j, while $T_{i,j}(t) \in [-1, +1]$ and $T_{i,j}(0) = 0$. One agent in the view of the other agent can have one of the following levels of trustworthiness: *Trustworthy*, *Not Yet Known*, or *Untrustworthy*. Following Marsh [7], we define an upper and a lower threshold for each agent to model different levels of trustworthiness. The agent i has its own upper threshold $-1 \leq \omega_i \leq 1$ and lower threshold $-1 \leq \Omega_i \leq 1$. Agent j is *Trustworthy* from the viewpoint of agent i after t times of interactions if and only if $T_{i,j}(t) \geq \omega_i$. Agent i sees agent j as an *Untrustworthy* agent if $T_{i,j}(t) \leq \Omega_i$ and if $\Omega_i < T_{i,j}(t) < \omega_i$ then the agent j is in the state *Not Yet Known*.

Direct Interaction Trust (DIT). Direct Interaction Trust (DIT) is the result of CDI/DDI. Each agent maintains $DIT_{i,j}(t)$ variables for the agents with which they have had direct interactions. We used the following trust updating scheme motivated by that described in [16]:

if $DIT_{i,j}(t) > 0$ **and** CDI **then**

$DIT_{i,j}(t+1) = DIT_{i,j}(t) + \alpha_D(i)(1 - DIT_{i,j}(t))$

if $DIT_{i,j}(t) < 0$ **and** CDI **then**

$DIT_{i,j}(t+1) = (DIT_{i,j}(t) + \alpha_D(i))/(1 - min(|DIT_{i,j}(t))|, |\alpha_D(i)|)$

if $DIT_{i,j}(t) > 0$ **and** DDI **then**

$DIT_{i,j}(t+1) = (DIT_{i,j}(t) + \beta_D(i))/(1 - min(|DIT_{i,j}(t))|, |\beta_D(i)|)$

if $DIT_{i,j}(t) < 0$ **and** DDI **then**

$DIT_{i,j}(t+1) = DIT_{i,j}(t) + \beta_D(i)(1 + DIT_{i,j}(t))$

Where $\alpha_D(i) > 0$ and $\beta_D(i) < 0$ are positive evidence and negative evidence weighting coefficients respectively for updating of the direct interaction trust variable of agent i. The value of $DIT_{i,j}(t)$, ω_i^{DIT} and Ω_i^{DIT} determine that the agent j is either *trustworthy*, *Not Yet Known* or *Untrustworthy* in terms of direct interaction from the perspective of agent i.

Witness Interaction Trust (WIT). Witness Interaction Trust (WIT) is the result of the cooperation/defection that the neighbors of an agent have with the agent regarding witness interaction (CWI/DWI). Agent i maintains a $WIT_{i,j}(t)$ variable for the agent j from whom it has received witness information. The updating scheme of $WIT_{i,j}(t)$ is similar to the one presented for $DIT_{i,j}(t)$ but CDI and DDI should be replaced by CWI and DWI respectively and $\alpha_D(i) > 0$ and $\beta_D(i) < 0$ is replaced with $\alpha_W(i) > 0$ and $\beta_W(i) < 0$ respectively. Where $\alpha_W(i) > 0$ and $\beta_W(i) < 0$ are positive evidence and negative evidence weighting coefficients respectively for updating of the witness interaction trust variable of agent i. The value of $WIT_{i,j}(t)$, ω_i^{WIT} and Ω_i^{WIT} determine that the agent j is either *Trustworthy*, *Not Yet Known* or *Untrustworthy* in terms of witness interaction.

4.2 Witness-Based Reputation (WR)

As agents need to predict the trustworthiness of those agents with whom they have never interacted, we use witness-based reputation (WR) for predicting trustworthiness of these agents. This reputation is calculated based on the witness information received from the neighbors.

Witness-based reputation for a specific agent is calculated based on the ratings of other agents. The asking agent stores the ratings of other agents in an *Opinion* variable. $Opinion(j,k)$ shows the rating issued by agent j regarding agent k. WR of agent k from the perspective of agent i after reception of t opinions (ratings) is denoted by $WR_{i,k}(t)$ and can be calculated by either Equation 1 or Equation 2:

$$WR_{i,k}(t) = \frac{\sum_{j \in OpinionSenders} (\phi(DIT_{i,j}) \times Opinion(j,k))}{\sum_{j \in OpinionSenders} \phi(DIT_{i,j})} \tag{1}$$

$$WR_{i,k}(t) = \frac{\sum_{j \in OpinionSenders} (\phi(WIT_{i,j}) \times Opinion(j,k))}{\sum_{j \in OpinionSenders} \phi(WIT_{i,j})} \tag{2}$$

In both formulae, the *OpinionSenders* variable includes indices of the neighbors of agent i who sent their ratings about agent k and $WIT_{i,j}$ is the current value of WIT variable of agent j from the perspective of agent i. Note that, $\phi(r)$ is a converter function that is calculated by Equation 3.

$$\phi(r) = \begin{cases} 0 & -1 \leq r < \Omega \\ \frac{r-\Omega}{\omega-\Omega} & \Omega \leq r \leq \omega \\ 1 & \omega \leq 1 \end{cases} \tag{3}$$

The value of $WR_{i,k}(t)$, ω_i^{WR} and Ω_i^{WR} determine that the agent k is either *Trustworthy*, *Not Yet Known* or *Untrustworthy* in terms of witness-based reputation from the perspective of agent i.

4.3 Agent Policy Types

The perceptions introduced above allow agents to determine the trustworthiness of other agents. Policies make use of agent perceptions, trust and reputation models in order to decide upon the set of agents with which they will interact and in what ways they will interact. Policies may cause the agent interaction neighborhood to change, for example. Several policy classes have been defined for the research reported here; they are explained in the following subsections.

Direct Interaction Policy (DIP). This type of policy assists an agent in making decisions regarding its direct interactions.

Witness Interaction Policy (WIP). This type of policy exists to aid an agent in making three categories of decisions related to its witness interactions. First, agents should decide how to provide the witness information for another agent on receiving a witness request. Should they manipulate the real information and forward false witness information to the requester (an example of defection) or should they tell the truth? The second decision made by the Witness Interaction Policy is related to when and from whom the agent should ask for witness information. Should the agents ask for the witness information when it has a connection request from an unknown party? Should the agents ask for witness information from a subset or all of its neighbors? The third decision is on how agents should aggregate the received ratings. For example, should the agent calculate the simple average of ratings or a weighted average of ratings?

We defined three sub witness interaction policies: Answering Policy (AP), Querying Policy (QP), and Information-Gathering policy (IGP). Answering Policy intends to cover the the first category of decisions mentioned above while Querying Policy and Information-Gathering policy apply to the second and third categories respectively.

Connection Policy (CP). This policy type assists an agent in making decisions regarding whether it should make a request for connection to other agents and whether the agents should accept/reject a request for a connection.

4.4 Experimentally Evaluated Policies

This section described policies that were evaluated experimentally.

Direct Interaction Policies. Three kinds of DIPs used in our experiments are: Always Cooperate (AC), Always-Defect (AD), and Trust-based Tit-For-Tat (TTFT). Agents using the AC policy for their direct interactions will cooperate with their neighbors in direct interactions regardless of the action of their neighbor. In contrast, agents using the AD policy will defect in all neighbor interactions. Agents employing TTFT will start with cooperation and then imitate the neighbors' last move as long as the neighbors are neither trustworthy nor untrustworthy. If a neighbor is known as untrustworthy, the agent will defect and immediately disconnect from that neighbor. If a neighbor is known as trustworthy, the agent will cooperate with it.

Connection Policy. Three kinds of connection polices used in our experiments are: Conservative (C), Naive (N), and Greedy (G). There is an internal property for each of these policies called Socializing Tendency (ST) which affects decisions regarding making connection requests and the acceptance of a connection request. Both Naive and Greedy policies use Algorithm 1 with different values for ST.

According to Algorithm 1, any connection request from another agent will be accepted regardless of ST value but the agent will attempt to connect to unknown agents if its number of neighbors is less than ST.

Algorithm 1. Greedy and Naive Policies

```
1: {CRQ is a queue containing the connection requests}
2: if CRQ is not empty then
3:    j = dequeue(CRQ)
4:    connectTo(j)
5: end if
6: if size(neighborhood) < ST then
7:    j = get unvisited agent from list of all known agents
8:    if ∃j ≠ null then
9:       requestConnectionTo(j)
10:   end if
11: end if
```

Algorithm 2. Conservative Connection Policy

```
1: if CAQ is not empty then
2:    j = dequeue(CAQ)
3:    connectTo(j)
4: end if
5: if size(neighborhood) < ST then
6:    if SIQ is not empty then
7:       j = dequeue(SIQ)
8:    else
9:       if CRQ is not empty then
10:         j = dequeue(CRQ)
11:      else
12:         j = get unvisited agent from list of all known agents
13:      end if
14:   end if
15:   if ∃j ≠ null then
16:      connectTo(j)
17:   end if
18: end if
19: if CRQ is not empty then
20:   j = dequeue(CRQ)
21:   enque(SIQ,j)
22: end if
```

Using the Conservative policy presented in Algorithm 2, the agents connect to confirmed agents regardless of the number of their neighbors. CAQ contains the list of agent IDs confirmed; this confirmation of an agent might be accomplished by a witness interaction policy. If the number of neighbors is less than ST, the agent connects to the agents requested for connections or to an unvisited agent. Finally, if there are any agent IDs in CRQ (a queue of connection requests), the first agent ID will be inserted in SIQ (a list of agents whose reputations should be investigated).

We set the value of ST 5, 25, and 100 for Conservative, Naive, and Greedy connection policies respectively.

Algorithm 3. Answering Policy

```
1: if receiving a witness request about j from k then
2:    opinion = ∗
3:    send opinion to k
4:    if |opinion − DIT_{i,j}(t)| < DT then
5:       Send CWI to k after T_W time steps
6:    else
7:       Send DWI to k after T_W time steps
8:    end if
9: end if
```

Witness Interaction Policy. Three kinds of answering policies are modeled: Honest (Ho), Liar (Li), and Misleader (Mi). All these sub-policies use the pseudo-code presented in Algorithm 3 while differentiating in the assignment of opinion variable (refer to * in Algorithm 3). The asterisk should be replaced by $DIT_{i,j}(t)$, "$-1 * DIT_{i,j}(t)$", or 1 for Honest, Liar, or Misleader policy respectively. An agent employing the Liar policy gives manipulated ratings to other agents by giving high ratings for untrustworthy agents and low ratings for trustworthy ones. The Misleader policy ranks all other agents as trustworthy but the Honest policy always tells the truth to everyone. CWI/DWI will be sent based on whether the forwarding opinion agrees with the internal trust value of an agent or not. If the difference between them is less than the Discrimination Threshold (DT), an agent will send CWI otherwise DWI is sent. We can therefore say that: Liar always defects, Honest always cooperates, and Misleader sometimes defects (by rating high untrustworthy agents) and sometimes cooperates (by rating low trustworthy agents) in providing the witness information. In the experiments reported here DT is set to 0.25.

Using the Querying Policy presented in Algorithm 4, the agents ask for witness information from their neighbors regarding agents which are in the SIQ queue. SIQ contains a list of agents whose reputations should be investigated. After asking for witness information regarding a specific agent, the ID of that agent is inserted in the WIFQ queue. WIFQ contains the list of agents waiting for the

Algorithm 4. Querying Policy

```
1: if SIQ is not empty then
2:    k = dequeue(SIQ)
3:    for all j ∈ Neighborhood do
4:       Ask for witness information about k from j
5:    end for
6:    enque(WFIQ, k)
7: end if
8: for all k ∈ WFIQ do
9:    if WR_{i,k} > ω_i^{WR} then
10:      enque(CAQ, k)
11:      remove(WFIQ, k)
12:   else
13:      if WR_{i,k} < Ω_i^{WR} then
14:         remove(WFIQ, k)
15:      else
16:         if ShouldBeReInvestigated(k) then
17:            for all j ∈ Neighborhood do
18:               Ask for witness information about k from j
19:            end for
20:         end if
21:      end if
22:   end if
23: end for
```

result of an investigation. If an agent in the WIFQ is known as trustworthy in the context of WR, then the ID of that agent will be added to CAQ which contains the list of confirmed agents. The agents known as untrustworthy in terms of WR will be removed from WIFQ. If an agent in WIFQ is known neither as trustworthy nor as untrustworthy and the primitive $ShouldBeReInvestigated(k)$ returns a *true* value, then again the agent will request witness information from their neighbors regarding the given agent (the agent k). This primitive can be easily implemented relying on whether agent k remains in WIFQ for more than a specific amount of time.

We have specified two information-gathering policies: DIT-based Weighted (DTW), and WIT-based Weighted (WTW). Both use Algorithm 5 while differentiating in the calculation of $WR_{i,k}$ (refer to * in Algorithm 5). DTW calculate it by using the formula presented in the Equation 1, Whereas WTW use the formula presented in the Equation 2.

Algorithm 5. Information-Gathering Policy

1: {Suppose that agent i is executing this code}
2: **if** receiving opinion about k from j **then**
3: Calculate $WR_{i,k}(t)$ based on *
4: **end if**

5 Experiments

We have empirically analyzed our agent types at both microscopic and macroscopic levels. On the macro level, we studied how society structure changes over the course of many interactions. On the micro level, the utility of agents and the number of dropped connections are examined. $\overline{U_{AT}(i)}$, the average of utilities for agents with the type of AT at time step i, is calculated by:

$$\overline{U_{AT}(i)} = \frac{\sum_{a \in AT} U_{Avg}(a, i)}{N_{AT}} \tag{4}$$

where $U_{Avg}(a, i)$ is the average of utility of agent a over its interactions at time step i and N_{AT} is the total number of agents in the society whose type is AT. The utility of each interaction is calculated as follows: If agent i defects and agent j cooperates, agent i gets the Temptation to Defect payoff of 5 points while agent j receives the Sucker's payoff of 0. If both cooperate each gets the Reward for Mutual Cooperation payoff of 3 points, while if both defect each gets the Punishment for Mutual Defection payoff of 1 point.

$\overline{D_{AT}(i)}$, the average of dropped connections for agents with the type of AT at time step i, is calculated by:

$$\overline{D_{AT}(i)} = \frac{\sum_{a \in AT} D_{total}(a, i)}{N_{AT}} \tag{5}$$

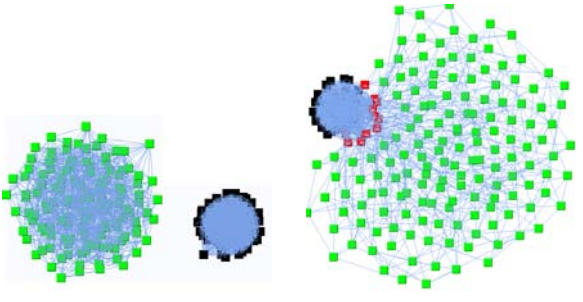
where $D_{total}(a, i)$ is the total number of connections broken for agent a from the start time to time step i and N_{AT} is total number of agents of society whose type are AT.

We have modeled the witness-based collusion attacks by using Enticer and Malicious agents as explained in Section 3. In addition to these two types of agents, the agent society includes Trust-Aware agents which are equipped with perception variables (trust and reputation variables) to assess trustworthiness of others and with policies to properly interact with others. We have defined two classes of Trust-Aware agents for our experiments: Trust-Aware (TA_w) and Trust-Aware$^+$ (TA_w^+) where TA_w uses a unidimensional trust model as opposed to TA_w^+ which uses a multi-dimensional trust model. Table 1 presents all agent types used for all experiments.

Experiment 1. We have run two simulations of 200 agents for this experiment. In the first simulation, which models a non-collusive society, TA_w agents comprise 75% of the population and the rest are Malicious agents (for our convenience, we refer to this simulation as Sim1). The second simulation represents the witness-based collusive society in which Enticer and Malicious agents comprise 5% and 20% of the populations respectively and the rest are TA_w agents (for our convenience, we refer to this simulation as Sim2). The objective of this experiment is to understand the effect of a witness-based collusion attack on the structure of agent society and on the level of encounter risk. Encounter risk is defined to be linearly related to the average number of dropped connections.

Table 1. Agent Types and Specifications

Name	Enticer	Malicious	TA_w	TA_w^+
Trust	-	-	DIT	DIT&WIT
DIP	AC	AD	TTFT	TTFT
CP	N	G	C	C
AP	Mi	Li	Ho	Ho
QP	-	-	QP	QP
IGP	-	-	DTW	WTW



(a) Non-collusive Society (b) Collusive Society

Fig. 2. The Final Society Structure in Exp. 1

The structures of the agent society after 400 time steps for Sim1 and Sim2 are presented in Figure 2 where TA_w agents and Malicious agents are in green (light gray in white-black print) and in black respectively. Red squares with white "-" represent Enticer agents. In non-collusive societies as shown in Figure 2a, we have two isolated groups of TA_w and Malicious agents. In the witness-based collusive society (see Figure 2b), we could not achieve separation of Malicious and TA_w agents seen in Sim1. Since TA_w agents perceived Enticer agents as trustworthy agents in direct interaction so they maintain their connections with Enticer agents. As illustrated in Figure 2b, TA_w agents are connected indirectly to Malicious agents by means of Enticer agents while acting a buffer between the Malicious agents and TA_w agents.

Figure 3 illustrates $\overline{D}$ of TA_w over the course of two simulations of Sim1 and Sim2. TA_w agents in Sim1 (non-collusive society) have considerably fewer dropped connections when compared to the TA_w agents in Sim2 (witness-based collusive society). In this sense, TA_w agents expose themselves to higher level of risk of being exploited by malicious agents in Sim2 as a result of ongoing witness-based attacks, when compared to Sim1. This high level of risk is due to the fact that each TA_w agent is surrounded by Enticer agents, resulting in receiving more manipulated opinions about other malicious agents while the senders of all opinions are trustworthy in terms of direct interactions.

Experiment 2. We have run two simulations of 200 agents for this experiment, in each of which Enticer, Malicious agents are 5% and 20% of the populations respectively. The remainder of the population (75%) is either TA_w or TA_w^+. Both TA_w and TA_w^+ benefit from using the Conservative connection policy and witness interaction policies for inquiring about the trustworthiness of the connection requester from neighbors. Note that, these two types employ various witness information-gathering policies and different trust models. TA_w utilizes a uni-dimensional trust model (i.e., DIT), whereas TA_w^+ utilizes a multi-dimensional trust model (i.e., DIT and WIT).

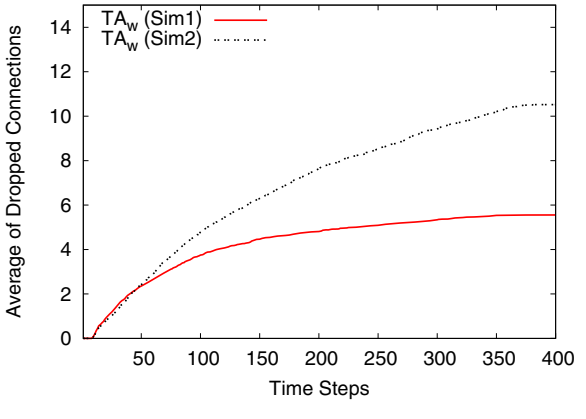


Fig. 3. $\overline{D}$ of agent types over simulation

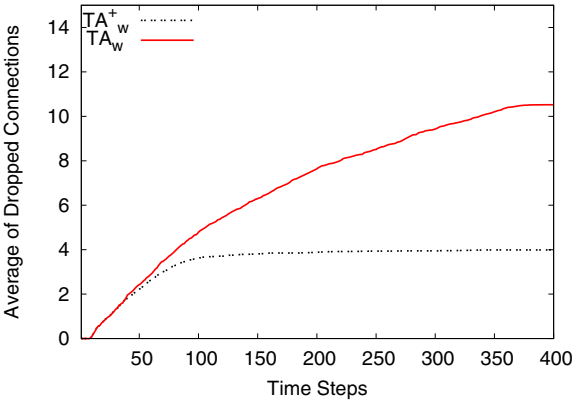


Fig. 4. $\overline{D}$ of agent types over the simulation

This experiment is intended to demonstrate the benefit of using multi-dimensional trust where there are witness-based collusion attacks. More precisely, the intention behind this experiment is to show that TA_w^+ agents by using multi-dimensional trust and appropriate witness interaction policies (e.g., WTW) can decrease the impact of Enticer and Malicious agents (colluding groups) on aggregating the reputation ratings. As a result, the TA_w^+ agents can decide more reliably regarding the trustworthiness of other agents and expose themselves to a lower level of risk.

As shown in Figure 4, TA_w^+ agents have considerably fewer dropped connections when compared to TA_w. Policies used by this agent type result in successful acceptance/rejection of connection requests. In this sense, TA_w^+ agents expose themselves to smaller numbers of untrustworthy agents and consequently lower the level of risk of being exploited by these agents.

Figure 5 illustrates $\overline{U}$ for TA_w and TA_w^+ types over the course of the simulations. $\overline{U}$ reaches the value of 3 faster for TA_w^+ than TA_w and will not fall below it later. This is evidence of the learning capability of TA_w^+ agents especially by using WIT for aggregating opinions in witness interaction policies. Each TA_w^+ agent, by updating WIT, will learn which of its neighbors are trustworthy in terms of witness information and then weight their opinions based on their WIT which is completely independent of DIT. As a result, false opinions of neighbors cannot mislead them several times whereas TA_w agents can be deceived several times by false opinions from the same neighbors (Enticer agents) because of the lack of this trust dimension.

Experiment 3. This experiment intends to show the effect of population proportion of Enticer agents on the efficacy of Witness-based collusion attacks and on the robustness of TA_w and TA_w^+. We have run 2 sets of 4 simulations. Each set consists of 200 agents with different proportions of Enticer and Malicious agents while keeping the proportion of either TA_w or TA_w^+ agents unchanged as shown in Table 2. The simulations are run for 400 time steps.

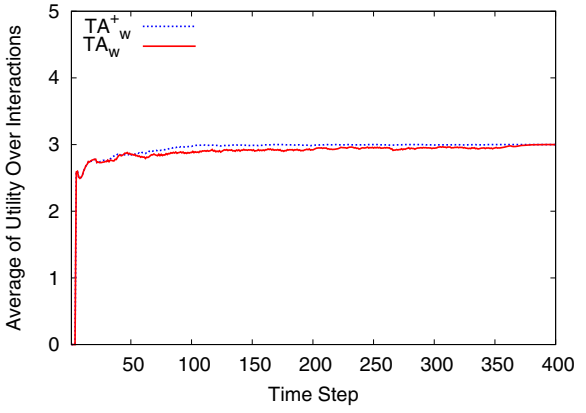


Fig. 5. $\overline{U}$ of agent types over the simulation

Table 2. Population Distributions of Experiment 3

Agent Type	Population			
	Pop1	Pop2	Pop3	Pop4
Malicious	20%	15%	10%	5%
Naive	5%	10%	15%	20%
TA_w or TA_w^+	75%	75%	75%	75%

Figure 6 presents $\overline{D}$ of each agent type at time step 400 for each of the runs. By increasing the proportion of Enticer agents (i.e., decreasing the proportion of Malicious agents), the $\overline{D}$ of TA_w and TA_w^+ will be decreased. Moreover, it can be observed that in all runs the number of dropped connection for TA_w is greater when compared with the number of dropped connections of TA_w^+. This is evidence of the fact that TA_w^+ has better robustness against this attack.

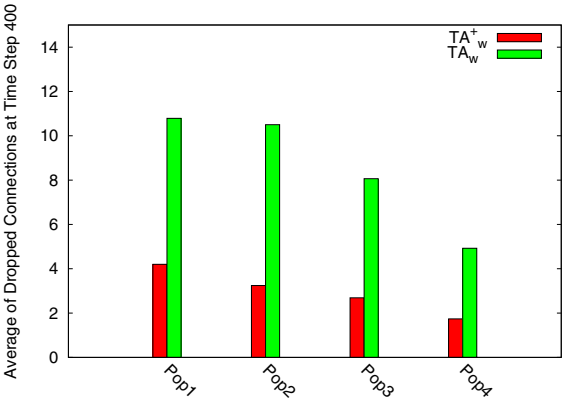


Fig. 6. $\overline{D}$ for all runs

6 Conclusion

Witness-based collusion attacks degrade the value of DIT in trust-aware agent societies. In these attacks, agents which are trustworthy in their direct interactions, collude with malicious agents by providing a good rating for them. We experimentally show how a unidimensional trust model is vulnerable against witness-based collusion attacks. This vulnerability results in TA_w agents, which use a unidimensional trust model to weight the ratings, exposing themselves to a higher level of encounter risk. Furthermore, TA_w^+ agents, by using WIT, weight the rating of Enticer agents and decrease the impact of them in their final assessment. This results in exposing themselves to a lower level of risk in their interactions. We empirically demonstrate that the efficacy of TA_w^+ over TA_w is better for different population proportions of Enticer and Malicious agents. Finally, we conclude multi-dimensionality is a crucial factor for resistance against witness-based collusion attacks.

Collusion attacks are an emerging area of research in trust and reputations systems. Future work will uncover new classes of such attacks and ways in which they can be defeated.

References

1. Axelrod, R.: The Evolution of Cooperation. Basic Books, New York (1984)
2. Dellarocas, C.: Mechanisms for coping with unfair ratings and discriminatory behavior in online reputation reporting systems. In: ICIS, pp. 520–525 (2000)
3. Feldman, M., Lai, K., Stoica, I., Chuang, J.: Robust incentive techniques for peer-to-peer networks. In: EC 2004, pp. 102–111. ACM, New York (2004)
4. Huynh, T.D., Jennings, N.R., Shadbolt, N.R.: An integrated trust and reputation model for open multi-agent systems. Autonomous Agents and Multi-Agent Systems 13(2), 119–154 (2006)
5. Jarvenpaa, S.L., Tractinsky, N., Vitale, M.: Consumer trust in an internet store. Inf. Technol. and Management 1(1-2), 45–71 (2000)
6. Reid, K., Robin, C.: Smart cheaters do prosper: Defeating trust and reputation systems. In: AAMAS 2009, Budapest, Hunagry, ACM, New York (2009)
7. Marsh, S.: Formalising trust as a computational concept (1994)
8. Mui, L., Mohtashemi, M., Halberstadt, A.: A computational model of trust and reputation for e-businesses. In: HICSS 2002, Washington, DC, USA, p. 188. IEEE Computer Society Press, Los Alamitos (2002)
9. Ramchurn, S.D., Huynh, D., Jennings, N.R.: Trust in multi-agent systems. Knowl. Eng. Rev. 19(1), 1–25 (2004)
10. Sabater, J., Sierra, C.: Regret: A reputation model for gregarious societies. In: Fourth Workshop on Deception Fraud and Trust in Agent Societies, pp. 61–70 (2001)
11. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artif. Intell. Rev. 24(1), 33–60 (2005)
12. Salehi-Abari, A., White, T.: Towards con-resistant trust models for distributed agent systems. In: IJCAI 2009: Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence, Pasadena, CA, USA, pp. 272–277 (2009)

13. Amirali, S.-A., Tony, W.: Witness-based collusion and trust-aware societies. In: SPOSN 2009: the Workshop on Security and Privacy in Online Social Networking (2009)
14. Luke Teacy, W.T., Patel, J., Jennings, N.R., Luck, M.: Coping with inaccurate reputation sources: experimental analysis of a probabilistic trust model. In: AAMAS 2005, pp. 997–1004. ACM Press, New York (2005)
15. Whitby, A., Jsang, A., Indulska, J.: Filtering out unfair ratings in bayesian reputation systems. In: Proceedings of 7th International Workshop on Trust in Agent Societies (2004)
16. Yu, B., Singh, M.P.: A social mechanism of reputation management in electronic communities. In: Klusch, M., Kerschberg, L. (eds.) CIA 2000. LNCS (LNAI), vol. 1860, pp. 154–165. Springer, Heidelberg (2000)
17. Yu, B., Singh, M.P.: Detecting deception in reputation management. In: AAMAS 2003, pp. 73–80. ACM, New York (2003)