Trust Models and Con-man Agents: From Mathematical to Empirical Analysis

Amirali Salehi-Abari and Tony White
School of Computer Science
Carleton University, Canada
{asabari, arpwhite}@scs.carleton.ca

Abstract

Recent work has demonstrated that several trust and reputation models can be exploited by malicious agents with cyclical behaviour. In each cycle, the malicious agent with cyclical behaviour first regains a high trust value after a number of cooperations and then abuses its gained trust by engaging in a bad transaction. Using a game theoretic formulation, Salehi-Abari and White have proposed the AER model that is resistant to exploitation by cyclical behaviour. Their simulation results imply that FIRE, Regret, and a model due to Yu and Singh, can always be exploited with an appropriate value for the period of cyclical behaviour. Furthermore, their results demonstrate that this is not so for the proposed adaptive scheme. This paper provides a mathematical analysis of the properties of five trust models when faced with cyclical behaviour of malicious agents. Three main results are proven. First, malicious agents can always select a cycle period that allows them to exploit the four models of FIRE, Regret, Probabilistic models, and Yu and Singh indefinitely. Second, malicious agents cannot select a single, finite cycle period that allows them to exploit the AER model forever. Finally, the number of cooperations required to achieve a given trust value increases monotonically with each cycle. In addition to the mathematical analysis, this paper empirically shows how malicious agents can use the theorems proven in this paper to mount efficient attacks on trust models.

1. Introduction

Recently, researchers have identified the existence of cheaters (exploitation) in artificial societies employing trust and reputation models (Kerr and Cohen 2009; Salehi-Abari and White 2009b; 2009a). (Kerr and Cohen 2009) 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. (Salehi-Abari and White 2009b) introduced and formally modeled the con-man attack. In the con-man attack, the con-man has cyclical behaviour such that in each cycle he first regains a high trust value with a number of cooperations and then misuses the trust gained by engaging in a bad transaction. (Salehi-Abari and White 2009b) empirically demonstrated the vulnerability of several trust models (i.e., the model due to Yu and Singh, Regret, and FIRE) against this attack. Moreover, their proposed adaptive updating scheme, called AER, prevented such exploitation as supported by empirical simulation evidence. Furthermore, AER does not rely on reputation obtained in terms of gain-in-exchange as other systems do. However, Salehi-Abari and White could not demonstrate that their results are valid for all parameter settings. More specifically, it was not clear whether the success or failure of the con-man attack against the examined trust models is the result of specific parameter settings or the design of those models.

This paper is motivated by the need to develop trust and reputation schemes that have provable properties. While simulation can often provide insights into average case trust and reputation model performance, mathematical analysis based upon known or potential attacks are important to increase confidence in the true utility of such models. To this end, this paper provides mathematical analysis of the con-man attack against several prominent trust models.

There are two types of contribution in this paper. To begin, we define what is meant by an attack on a trust and reputation model and what it is meant for such models to be vulnerable to an attack or exhibit exploitation resistance to the attack. Our principal contributions are analytical and consist of 5 results. First, we prove that the Yu and Singh model and FIRE can be exploited indefinitely if malicious agents are aware of the model's parameter settings. Second, Regret and probabilistic trust models can be exploited indefinitely by malicious agents mounting a con-man attack even when malicious agents are not aware of the model's parameters. Third, malicious agents can not indefinitely exploit AER. Fourth, the number of cooperations required to achieve a given trust value increases monotonically without any upper bound in AER, while this is not true for the other models. Fifth, as forgiveness is a frequently noted aspect of trust and reputation theory (Sabater and Sierra 2001; Axelrod 1984), it is proven that the AER scheme is forgiving but that forgiveness is slower when several defections have happened. In addition, this paper empirically shows how malicious agents can use the theorems provided in this paper to mount efficient attacks on trust models.

The remainder of this paper proceeds as follows. Section 2 provides background material and briefly describes five trust and reputation models whose properties in the face of the con-man attack are analyzed in this paper. Section 3

introduces definitions for vulnerability and exploitation resistance and provides a formal model of the con-man attack. We describe our hypotheses and conjectures in Section 4. Section 5 presents lemmas and theorems. Section 6 presents simulation results and discussion. Finally, concluding remarks are explained in Section 7.

2. Background and Terminology

2.1 Direct Interaction Components

Direct interaction is the most popular source of information for trust and reputation models (Ramchurn, Huynh, and Jennings 2004). Trust and reputation models usually have a direct interaction trust variable that indicates the level of an agent's trustworthiness. We discuss the direct interaction trust components of Yu and Singh's model, Regret, FIRE, and probabilistic trust models in the following subsections.

Yu and Singh Yu and Singh's (Yu and Singh 2000) 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, with $T_{i,j}(t) \in [-1, +1]$ and $T_{i,j}(0) = 0$.

An agent will update this variable based on the perception of cooperation/defection. Cooperation by the other agents generates positive evidence of α, with $1 > \alpha > 0$ and defection generates negative evidence of β, with $-1 < \beta < 0$.
If $T_{i,j}(t) > 0$ **and** $Cooperation$ **then**
$T_{i,j}(t+1) := T_{i,j}(t) + \alpha(1 - T_{i,j}(t))$
If $T_{i,j}(t) < 0$ **and** $Cooperation$ **then**
$T_{i,j}(t+1) := (T_{i,j}(t) + \alpha)/(1 - min(|T_{i,j}(t))|, |\alpha|)$
If $T_{i,j}(t) > 0$ **and** $Defection$ **then**
$T_{i,j}(t+1) := (T_{i,j}(t) + \beta)/(1 - min(|T_{i,j}(t))|, |\beta|)$
If $T_{i,j}(t) < 0$ **and** $Defection$ **then**
$T_{i,j}(t+1) := T_{i,j}(t) + \beta(1 + T_{i,j}(t))$

Regret Regret defines an impression as the subjective evaluation made by an agent on a certain aspect of an outcome and bases its trust model upon it. The variable $r_{i,j}(t)$, with $r_{i,j}(t) \in [-1, 1]$, is the rating associated with the impression of agent i about agent j as a consequence of specific outcome at time t. $R_{i,j}$ is the set of all $r_{i,j}(t)$ for all possible t. A subjective reputation at time t from agent i's point of view regarding agent j is noted as $T_{i,j}(t)$[1]. To calculate $T_{i,j}(t)$, Regret uses a weighted mean of the impressions' rating factors, giving more importance to recent impressions. The formula to calculate $T_{i,j}(t)$ is:

$$T_{i,j}(t) = \sum_{w_k \in R_{i,j}} \rho(t, t_k).w_k \quad (1)$$

where t_k is the time that w_k is recorded, t is the current time, $\rho(t, t_k) = \frac{f(t_k,t)}{\sum_{r_l \in W_{i,j}} f(t_l,t)}$, and $f(t_k, t) = \frac{t_k}{t}$ which is called the rating recency function.

FIRE FIRE (Huynh, Jennings, and Shadbolt 2006) utilizes the direct trust component of Regret but does not use its rating recency function. FIRE introduced a rating recency function based on the time difference between current time and the rating time. The parameter λ is introduced into the rating recency function to scale time values. FIRE's rating recency function is:

$$f(t_k, t) = e^{-\frac{t-t_k}{\lambda}} \quad (2)$$

Probabilistic Trust Models Considerable progress has recently been made in the development of probabilistic trust models, the Beta Reputation System (BRS) and TRAVOS being two examples (Josang and Ismail 2002; Teacy et al. 2005). Probabilistic trust models are built based on observations of past interactions between agents mapping observations to cooperations and defections.

In probabilistic trust models, the probability that agent j satisfies its obligations for agent i is expressed by $B_{i,j}$. The trust value of agent i for agent j at time t, denoted by $T_{i,j}(t)$, is the expected value of $B_{i,j}$ given the set of outcomes $O_{i,j}(t)$ at time t.

$$T_{i,j}(t) = E[B_{i,j}|O_{i,j}(t)] \quad (3)$$

As the standard equation for the expected value of a beta distribution is $E[B|\alpha, \beta] = \frac{\alpha}{\alpha+\beta}$, the trust value $T_{i,j}(t)$ after t interactions is:

$$T_{i,j}(t) = E[B_{i,j}|\alpha, \beta] = \frac{\alpha}{\alpha + \beta} \quad (4)$$

where $\alpha = n_c(t) + 1$ and $\beta = n_d(t) + 1$. $n_c(t)$ and $n_d(t)$ denote the number of cooperations (successful interactions) and the number of defections (unsuccessful interactions)[2].

AER AER extended the direct trust of (Yu and Singh 2000) by introducing the following update schema for a positive evidence weighting coefficient of $\alpha > 0$ and a negative evidence weighting coefficient $\beta < 0$ when the agent perceives defection:

$$\begin{aligned} \alpha(i) &= \alpha(i-1) \times (1 - |\beta(i-1)|) \\ \beta(i) &= \beta(i-1) - \gamma_d \times (1 + \beta(i-1)) \end{aligned}$$

Where γ_d is the discounting factor and is in the range of $[0, 1]$. Note that, $\alpha(i)$ and $\beta(i)$ will be updated when the i^{th} defection occurs.

3. Definitions

In this paper we consider an agent's trust and reputation model, *M*, to be characterized by two attributes, S and P; S is the trust and reputation strategy being employed and P is the set of parameter values that are used to operate it. This paper deals with the concept of vulnerability. We define more precisely *vulnerability* and the levels of vulnerability of trust models against an attack as follows:

Definition Attack. An attack, *A*, is a sequence of cooperations and defections used by a malicious agent, *ma*, to achieve or maintain a trustworthy status as maintained by an agent, *ta*, with which it is interacting.

Definition Vulnerability. A trust model, *M*, is vulnerable to an attack, *A*, if a malicious agent, *ma*, adopting some strategy and with full or partial knowledge of an agent, *ta*, and its associated trust model, *M*, can be trustworthy as determined by *ta*.

[1] For the purpose of simplification, we have changed the original notations from (Sabater and Sierra 2001).

[2] It is worth mentioning that the trust value in probabilistic models is in the range of [0, 1] as opposed to Yu and Singh, Regret, and FIRE models in which trust is in the range of [-1, 1].

We define the following levels of vulnerability in this paper:

Definition Low-level. A trust model, *M*, is vulnerable to an attack, *A*, with low-level risk, if it is vulnerable only for some specific model parameter settings and *ma* needs to be aware of the parameter values used by *ta* to mount a successful attack.

Definition Medium-level. A trust model, *M*, is vulnerable to an attack, *A*, with medium-level risk, if it is vulnerable for any parameter settings and *ma* needs to be aware of the value of parameters used by *ta* to successfully mount an attack.

Definition High-level. A trust model, *M*, is vulnerable to an attack, *A*, with high-level risk, if *ma* is able to successfully mount an attack under any conditions even when *ma* is not aware of the values of parameters.

Finally, we say that a trust and reputation model, *M*, exhibits exploitation resistance to an attack, *A*, if it is not vulnerable to that attack. We also refer to a trust and reputation model, *M*, as being exploitation resistant when faced with an attack, *A*.

3.1 Con-man Attack and Terminology

In the con-man attack introduced in (Salehi-Abari and White 2009b), a con-man is modeled by the parameter θ. The con-man will defect after cooperating θ times. After each defection, the con-man will again cooperate θ times possibly repeating this interaction pattern several times.

In this paper, there is a slight modification in the con-man interaction pattern when compared to (Salehi-Abari and White 2009b). Here, the con-man has a higher level of intelligence such that it will defect in an interaction with the victim agent whenever its trust value is equal to or greater than a threshold, denoted by T_c. In other words, the con-man will cooperate until its trust value reaches T_c. We formally model this interaction pattern with Equation 5:

$$L = \left\{ (C^{\theta_i} D)^{+} | i = 0...n, \theta_i \in \mathbb{N} \right\} \tag{5}$$

Where C and D represent cooperation and defection respectively. The main difference in this interaction pattern when compared with that presented in (Salehi-Abari and White 2009b) is that θ_i is subject to change for each cycle of cooperation and defection instead of being a constant. The value of θ_i is determined by the number of the cooperations which the con-man needs to increase its trust value above T_c.

For the purpose of simplification in the proofs which follow, we rewrite the interaction pattern in such a way that the i^{th} cycle of interactions starts with a defection and followed by θ_i cooperations. In this sense, the first cooperations of con-man, θ_0, which results in an increment of trust from T_0 to T_c is not modeled. In other words, we consider the con-man has already built up its trust to T_c from T_0 by θ_0 cooperations. The modified interaction pattern of the con-man is presented in Equation 6.

$$L = \left\{ (DC^{\theta_i})^{+} | i = 1...n, \theta_i \in \mathbb{N} \right\} \tag{6}$$

More precisely, we herein highlight the terminology that is used in this paper and is illustrated in Figure 1. The variable θ_i is the number of cooperations that the con-man will have in the i^{th} cycle of defection-cooperations. The i^{th} cycle includes the i^{th} defection and θ_i cooperations. The trust value at the end of the i^{th} cycle is T_c or greater; i.e., T_c defines the criterion for ending the i^{th} cycle. The trust value of the con-man before the i^{th} defection is denoted by $T_b(i)$. $T_d(i)$ denotes the trust value of the con-man after the i^{th} defection.

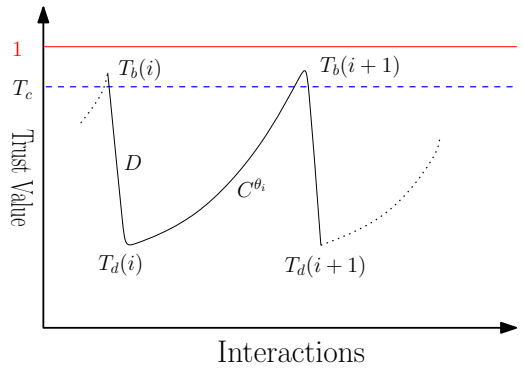

Figure 1: Trust value changes in the i^{th} cycle of defection-cooperations

4. Hypothesis and Conjectures

This paper intends to prove, for the con-man attack, that:

- Yu and Singh model is vulnerable with medium-level risk.
- Regret is vulnerable with high-level risk.
- FIRE is vulnerable with medium-level risk.
- Probabilistic models are vulnerable with high-level risk.
- AER is not vulnerable to the con-man attack.

5. Lemmas and Theorems

We here provide principal lemmas and theorems proved using mathematical analysis. The proofs from (Salehi-Abari and White 2010) are omitted owing to space limitations.

5.1 Yu and Singh

Our proof strategy is to find θ_i for any parameter setting in any defection-cooperations cycle and show that $\theta_{i+1} \leq \theta_i$. It is not straightforward to calculate θ_i for the Yu and Singh model since it includes several recurrent formulae. The proof is broken down into a series of cases that mirror the distinct forms of the formulae. Lemmas, 1, 2 and 3 present results for the cases. Theorem 1 provides a closed form solution for the number of cooperations that are required in a con-man attack to reach a given level of trust that is simply a function of α and β. This results in Theorem 2 that proves the middle-level risk vulnerability of this trust model.

Lemma 1 *Given that* $T_c > 0$ *and* $T_c \geq |\beta|$, *then* $\theta_i = \frac{\ln(1+\beta)}{\ln(1-\alpha)}$.

Lemma 2 *Given that* $0 < T_c < |\beta|$ *and* $-\alpha < T_d(i)$, *then* $\theta_i = \frac{\ln(1+\beta)}{\ln(1-\alpha)}$.

Lemma 3 *Given that* $0 < T_c < |\beta|$ *and* $-\alpha \geq T_d(i)$, *then* $\theta_i = \frac{\ln(1+\beta)}{\ln(1-\alpha)}$.

Theorem 1 *If $1 > T_c > 0$, θ_i will be calculated by $\frac{\ln(\beta+1)}{\ln(1-\alpha)}$ for Yu and Singh's model.*

Theorem 2 *Yu and Singh model is vulnerable to the con-man attack with a medium-level risk for any α and β.*

5.2 Regret

Our proof strategy is to find θ_i for any parameter setting in any defection-cooperations cycle and show that $\theta_{i+1} \leq \theta_i$. Throughout our analysis and proofs, cooperation and defection are mapped to 1 and -1 respectively for the Regret model and the appropriate value is used as an input of the trust model.

Lemma 4 *Given t_i is the number of interactions at the beginning of the i^{th} cycle, θ_i for the Regret model will be calculated by $\theta_i = \frac{-3-2t_i+\sqrt{(2t_i+\frac{T_c-5}{T_c-1})^2-\frac{8(T_c+1)}{(T_c-1)^2}}}{2}$*

Theorem 3 *Let $1 > T_c > 0$, $\theta_i < \frac{T_c+1}{1-T_c}$ for any $i \in \mathbb{N}$ in the Regret model.*

Theorem 4 *The Regret model is vulnerable to the con-man attack with high-level risk.*

5.3 FIRE

Our proof strategy is to find θ_i for any parameter setting in any defection-cooperations cycle and show that $\theta_i \leq \theta_c$, where θ_c is a constant. Throughout our analysis and proofs, the cooperation and defection is mapped to 1 and -1 respectively for the FIRE model and the value is used as an input of the trust model.

Lemma 5 *Let $T(t)$ and T_{ta} be the starting trust value and the target trust value as a result of θ cooperation respectively, where t is the number of interactions used for calculation of $T(t)$. θ is calculated by $\theta = \lambda \ln\left(1 - \frac{1-e^{\frac{t}{\lambda}}}{T_{ta}-1} \times (T(t)-1)\right) - t$.*

Lemma 6 *Let t_i be the number of interactions at the beginning of the i^{th} cycle, the trust value after the i^{th} defection, $T_d(i)$ will be calculated by $T_d(i) = \frac{1-e^{\frac{t_i}{\lambda}}}{1-e^{\frac{t_i+1}{\lambda}}} \times (T_b(i)+1) - 1$ where $T_b(i)$ is the trust value before the i^{th} defection.*

Theorem 5 *Given t_i be the number of interactions at the beginning of the i^{th} cycle, θ_i for the FIRE model will be calculated by:*

$$\theta_i = \lambda \ln\left(\frac{T_c + 1 - 2e^{\frac{1}{\lambda}}}{T_c - 1}\right) - 1 \qquad (7)$$

Theorem 6 *The FIRE model is vulnerable to the con-man attack with medium-level risk for any value of λ.*

5.4 Probabilistic Trust models

Our proof strategy is to show that the value of θ is a simple function of the trust threshold, T_c, thereby implying that the trust model is vulnerable to a con-man attack.

Lemma 7 *Let $T(t)$ and T_{ta} be the starting trust value and the target trust value as a result of θ cooperations respectively, where t is the number of interactions used for calculation of $T(t)$. θ is calculated by $\theta = \frac{(t+2)(T_{ta}-T(t))}{1-T_{ta}}$.*

Theorem 7 *Given t_i as the number of interactions at the beginning of the i^{th} cycle, θ_i for a probabilistic trust model will be calculated by $\theta_i = \frac{T_c}{1-T_c}$*

Theorem 8 *Probabilistic models are vulnerable to the con-man attack with high-level risk.*

5.5 AER

We will prove here that the AER update scheme for α and β is con-resistant in such a way that the con-man requires more cooperations in each cycle of defection-cooperations when compared to the previous cycle in order to reach to T_c. In other words, $\theta_i < \theta_{i+1}$.

Lemma 8 *$\beta(i)$ is ever-decreasing (i.e., $\beta(i) > \beta(i+1)$).*

Lemma 9 *$\alpha(i)$ is ever-decreasing (i.e., $\alpha(i) > \alpha(i+1)$).*

Theorem 9 *Let $T_c > 0$, θ_i will be calculated by $\frac{\ln(\beta(i)+1)}{\ln(1-\alpha(i))}$ for AER.*

Theorem 10 *AER will not let the con-man regain a high trust value easily with the same or smaller number of cooperations (i.e., $\theta_i < \theta_{i+1}$). AER is exploitation resistant to the con-man attack.*

Corollary 1 *AER is forgiving in any cycle of defection-cooperations but is more strict after each defection.*

6. Simulation Experiments

We here demonstrate how a con-man agent efficiently mounts an attack using the theorems proven in previous sections using simulation. All simulations were run with one trust-aware agent (TAA) which utilizes a specific computational trust model and a con-man agent (CA). The interaction of agents with each other can be either cooperation or defection. The interaction strategy of TAAs is tit-for-tat which starts by cooperation and then imitates the opponent's last move. The interaction strategy of CAs follows the formal language presented in Section 3.1 which is solely dependent on the parameter θ_i. CAs calculate the optimized θ_i using the theorems presented in Section 5.

6.1 Yu and Singh

We here demonstrate how a CA can efficiently mount an attack when θ_c is calculated by using Theorem 1. We assume that the con-man knows the values of α and β for calculating θ_c. The values of (α,β) for the 4 experiments were set to $(0.1, -0.2)$, $(0.075, -0.25)$, $(0.05, -0.3)$, or $(0.025, -0.35)$. The CA has set $T_c = 0.8$ for itself.

Figure 2 shows the variation of the trust value of the TAA. As shown, the con-man agent by using Theorem 1 could successfully calculate θ_c and consequently regain lost trust with the same number of cooperations in each cycle. The θ_c values for different settings of α and β are shown in Table 1. As θ_c should be an integer, we calculate the ceiling of θ_c, denoted by $\lceil \theta_c \rceil$, to use in our simulations. The side effect of this ceiling appears in Figure 2 for $\alpha = 0.1$ and $\beta = -0.2$ settings when $T_b(i)$ for high number of cycles reaches 1 as opposed to $T_c = 0.8$ because of rounding up of 2.1179 to 3. It is interesting to note that when the magnitude of β is

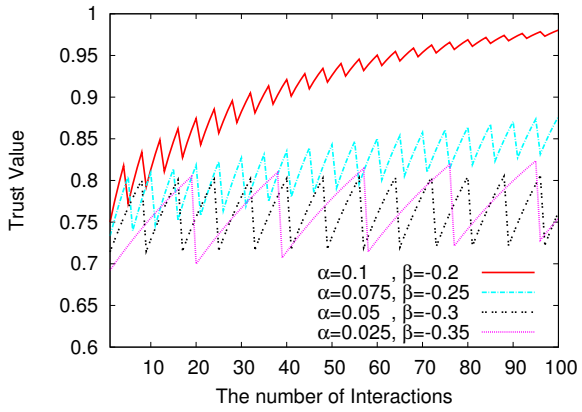

Figure 2: Exploitation of Yu & Singh model by a con-man.

much larger than that of α (e.g., $\alpha = 0.025$ and $\beta = -0.35$), which leads to a small improvement for a cooperation and a big drop for a defection, the con-man needs to choose a higher θ_c (e.g., 18) value.

	$\alpha = 0.1$ $\beta = -0.2$	$\alpha = 0.075$ $\beta = -0.25$	$\alpha = 0.05$ $\beta = -0.3$	$\alpha = 0.025$ $\beta = -0.35$
θ_c	2.1179	3.6901	6.9536	17.0150
$\lceil\theta_c\rceil$	3	4	7	18

Table 1: The Values of θ_c for various α and β settings.

6.2 Regret

We repeated the previous experiment where the trust-aware agent employs the Regret model and the con-man agent uses Theorem 3. Figure 3 demonstrates how a CA can efficiently mount an attack when θ_c is calculated by using Theorem 3. We ran 4 simulations in each of which the con-man agent has set T_c to 0.7, 0.8, 0.9, or 0.95 respectively for itself. As the consequence of this T_c setting, the following θ_c is calculated by the con-man agent: 5.67, 9, 19, 39. It is worth noting that the con-man agent does not require knowledge of the model's parameter settings to mount a successful attack and calculate an efficient θ_c. This is because θ_c is only dependent on T_c which is the parameter set by the con-man itself.

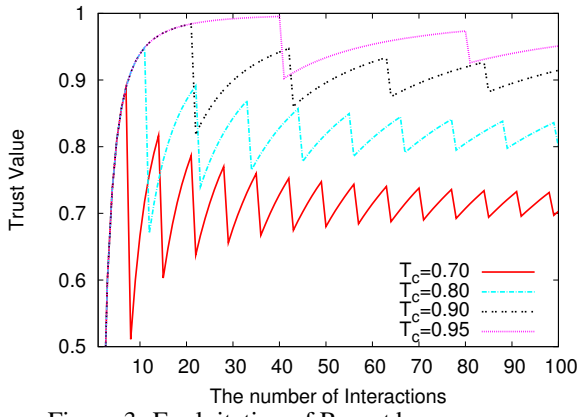

Figure 3: Exploitation of Regret by a con-man.

6.3 FIRE

The previous experiments were repeated with the trust-aware agent employing FIRE and the con-man agent using Theorem 5. The con-man is assumed to know the value of λ to calculate θ_c. Figure 4 depicts the variation of the trust value of TAA over the of the first set of simulations where we set λ to 4, 8, 16, or 32 while $T_c = 0.8$. The con-man could successfully calculate the following θ_c for the examined λs and T_c: 4.38, 5.77, 6.96, 7.82. Although, FIRE is more sensitive to defection for lower values of λ (e.g., the sharp drop of trust value after each defection for $\lambda = 4$), θ_c is lower for lower values of λ. In addition, It is interesting to note that θ_c changes linearly in spite of an exponential increase of λ. We set T_c to 0.7, 0.8, 0.9, or 0.95 while $\lambda = 8$.

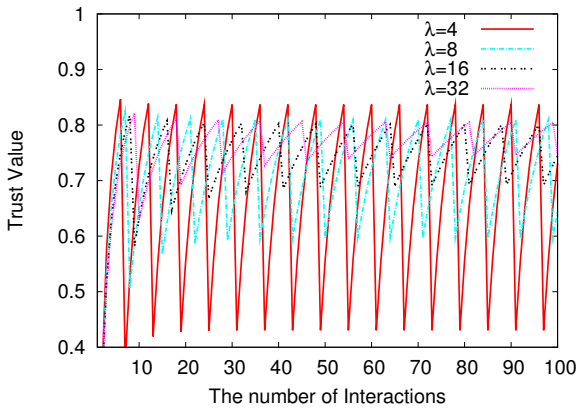

Figure 4: Exploitation of FIRE by a CA for various λ.

For these settings, the following values of θ_c are calculated by the con-man: 4.08, 5.77, 9.39, 13.76.

6.4 Probabilistic Trust Models

We ran 4 simulations where the trust-aware agent employs the probabilistic trust model. For each simulation, the con-man agent has set T_c to 0.7, 0.8, 0.9, or 0.95 respectively for itself. These T_c settings yielded the following θ_c which are calculated by the con-man agent using Theorem 7: 2.33, 4, 9, 19. Figure 5 demonstrates how the con-man agent can efficiently mount an attack by using Theorem 7 to calculate θ_c. As with Regret, the con-man agent does not need to know any trust model's parameter settings to mount a successful attack by calculating an efficient θ_c. This is because θ_c is only dependent on T_c which is the parameter set by the con-man itself.

6.5 AER

We ran 4 simulations with the same settings of previous experiments with the difference that the trust-aware agent uses AER and the con-man agent uses Theorem 9 to calculate θ_i for each cycle of defection-cooperations. For each simulation, we set different values of (α_0,β_0) for TAA as follows $(0.1, -0.2)$, $(0.075, -0.25)$, $(0.05, -0.3)$, and $(0.025, -0.35)$. For all simulations, $\gamma_d = 0.1$ and the CA has set $T_c = 0.8$ for itself. As the con-man needs the values (α_i,β_i) in each cycle of cooperation to calculate θ_i, we assume that the con-man is aware of these values.

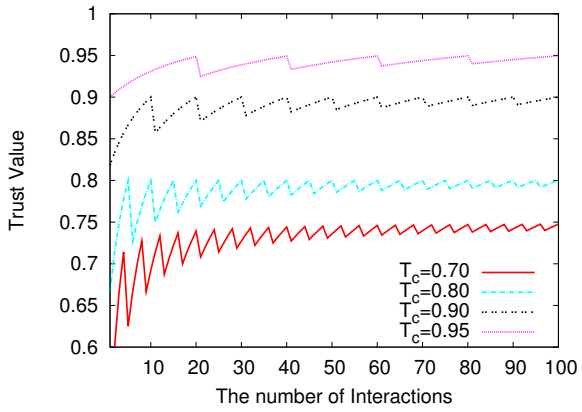

Figure 5: Exploitation of probabilistic trust models by a CA.

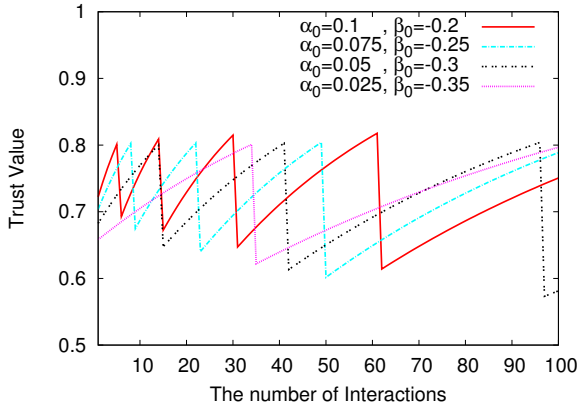

Figure 6: The con-man vs. AER

Figure 6 shows the trust value variation of the TAA for various α_0 and β_0 settings. Note that the con-man can be forgiven after each defection but with the larger number of cooperations and a change in its pattern of interactions (i.e., θ_i should be increased in each cycle of defection-cooperations). The θ_i values for different settings of α_0 and β_0 in each cycle of defection-cooperations are shown in Figure 7. Note that θ_i is increasing exponentially after each defection regardless of α_0 and β_0.

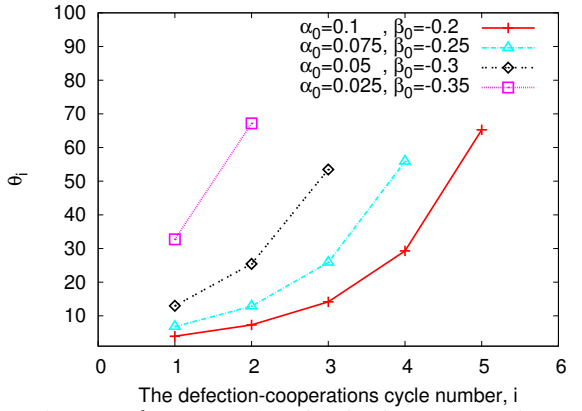

Figure 7: θ_i over cycles of defection-cooperations.

7. Conclusions and Future work

This paper is motivated by the dire need to develop trust and reputation schemes that have provable properties for artificial societies, especially e-commerce. This paper has proven that simple malicious agents with cyclical behaviour can exploit Yu and Singh's trust model, Regret, FIRE, and probabilistic trust models regardless of the model's parameters. However, AER has been shown to be exploitation resistant. Furthermore, malicious agents with cyclical behaviour will have to increase the number of cooperations in each and every cycle with AER in order to achieve a specific trust value. It is proven that AER is forgiving but that the rate of forgiveness slows with every defection. This paper has also empirically demonstrated how con-man agents can mount efficient attacks using theorems presented in this paper. Future work will design adaptive schemes similar to AER for Regret, FIRE, and probabilistic trust models and their exploitation resistance to the con-man attack proven.

References

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Huynh, T. D.; Jennings, N. R.; and Shadbolt, N. R. 2006. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 13(2):119–154.

Josang, A., and Ismail, R. 2002. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*.

Kerr, R., and Cohen, R. 2009. Smart cheaters do prosper: Defeating trust and reputation systems. In *AAMAS '09*.

Ramchurn, S. D.; Huynh, D.; and Jennings, N. R. 2004. Trust in multi-agent systems. *Knowl. Eng. Rev.* 19(1):1–25.

Sabater, J., and Sierra, C. 2001. Regret: A reputation model for gregarious societies. In *Fourth Workshop on Deception Fraud and Trust in Agent Societies*, 61–70.

Salehi-Abari, A., and White, T. 2009a. On the impact of witness-based collusion in agent societies. In *PRIMA '09: Proceedings of the 12th International Conference on Principles of Practice in Multi-Agent Systems*, 80–96. Berlin, Heidelberg: Springer-Verlag.

Salehi-Abari, A., and White, T. 2009b. Towards con-resistant trust models for distributed agent systems. In *IJCAI '09: Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence*, 272–277.

Salehi-Abari, A., and White, T. 2010. A mathematical analysis of computational trust models with the introduction of con-man agents. Technical Report, TR-10-01, School of Computer Science, Carleton University.

Teacy, W. T. L.; Patel, J.; Jennings, N. R.; and Luck, M. 2005. Coping with inaccurate reputation sources: experimental analysis of a probabilistic trust model. In *AAMAS '05*, 997–1004. New York, NY, USA: ACM.

Yu, B., and Singh, M. P. 2000. A social mechanism of reputation management in electronic communities. In *CIA '00*, 154–165. London, UK: Springer-Verlag.