Using Information Theory to Improve the Robustness of Trust Systems

Dongxia Wang, Tim Muller, Athirai A. Irissappane, Jie Zhang, Yang Liu
School of Computer Engineering, Nanyang Technological University, Singapore
{wang0915,athirai001}@e.ntu.edu.sg, {tmuller,zhangj,yangliu}@ntu.edu.sg

ABSTRACT

Unfair rating attacks to trust systems can affect the accuracy of trust evaluation when trust ratings (recommendations) about trustee agents are sought by truster agents from others (advisor agents). A robust trust system should remain accurate, even under the worst-case attacks which yield the least useful recommendations. In this work, we base on information theory to quantify the utility of recommendations. We analyse models where the advisors have the worst-case behaviour. With these models, we formally prove that if the fraction of dishonest advisors exceeds a certain threshold, recommendations become completely useless (in the worst case). Our evaluations on several popular trust models show that they cannot provide accurate trust evaluation under the worst-case as well as many other types of unfair rating attacks. Our way of explicitly modelling dishonest advisors induces a method of computing trust accurately, which can serve to improve the robustness of the trust models.

Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Intelligent agents

Keywords

Trust System; Unfair Rating; Robustness; Information Leakage; Worst-Case Attack

1. INTRODUCTION

In a trust system, a truster agent evaluates the trustworthiness of a trustee agent with which it interacts, based on its direct experiences and the recommendations about the trustee provided by other trusters (called advisors). However, some dishonest advisors may launch attacks by providing misleading recommendations, also known as unfair ratings [9, 20, 6, 2]. Thus, the accuracy of trust evaluation depends on the robustness of the trust system, that is, whether it can function properly under all situations – particularly in the worst case where dishonest advisors launch the worst-case unfair rating attacks – and be capable of handling unfair rating attacks in a satisfactory manner [10, 13].

In order to properly value an advisor's recommendations, we need to establish the honesty (or trustworthiness) of the advisor. This is what the existing trust models typically achieve [20, 21, 19]. However, we argue that knowing only the trustworthiness of advisors is not sufficient. For a complete picture, we also need to understand how the advisors behave when they are dishonest. Therefore, in this paper, we focus on analysing the behaviors of dishonest advisors (i.e., attacks), especially the worst-case attack, aiming to improve the robustness of the existing trust models.

More specifically, for the truster who wants to learn from recommendations, the worst-case unfair rating attack is to deliberately produce recommendations from which the truster learns the least[1]. In this work, we use information theory to quantify how much information the truster can learn as the *information leakage* of recommendations. The worst-case unfair rating attack is then identified as minimising the information leakage of the recommendations. Based on this thinking, we find out the strategies of dishonest advisors that lead to the worst case.

Some notable theoretical contributions are: 1) we prove that in the worst case, even if the fraction of dishonest advisors is larger than the commonly asserted threshold 0.5, the truster can still obtain information from recommendations; 2) we prove that, even in the case where the truster obtains zero-information, dishonest advisors may still sometimes report the truth; 3) we also prove that, for dishonest advisors, to minimise the information leakage of their true observations and that of the trustworthiness (or integrity) of trustees, they need to perform different attacking strategies.

Based on the theoretical analysis and the explicit modelling of the worst-case attacking strategies of the dishonest advisors, we propose an induced trust computation (ITC) method, which can ensure the accuracy of trust evaluation under the worst case. The experimental results demonstrate that under the worst case, ITC predicts either the integrity of trustees or the true observations of dishonest advisors with much higher accuracy compared to three representative trust models: TRAVOS [20], BLADE [17] and MET [6]. To defend unfair rating attacks, always assuming the worst case is a safe but may not always be the most accurate choice. Hence, we also investigate and compare the performance of ITC with TRAVOS, BLADE and MET under various unfair rating attacks. The experimental results show that although our ITC method assumes the worst-case attack in computation, it still presents the higher accuracy in trust evaluation

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May, 4–8, 2015, Istanbul, Turkey.*

[1] Recommendations opposite to the truth, for example, may carry useful information. BLADE is a trust model capable of learning from such opposite recommendations [17].

than TRAVOS, BLADE and MET when dealing with those types of unfair rating attacks that are not the worst case. All these results confirm that our method can effectively improve the robustness of the trust models.

2. RELATED WORK

The unfair rating problem has been recognized as an important and challenging problem in trust systems [1, 9]. To deal with this problem, many approaches attempt to accurately model the trustworthiness of advisors in giving ratings. We introduce two representative examples. Based on the beta probability density function, TRAVOS [20] estimates the trustworthiness of an advisor by examining the reliability of the previous recommendations provided by this advisor. BLADE [17] builds a Bayesian network model to learn the advisors' evaluation function which provides the probability of ratings given the trustee's features. However, these trust models have the common assumption that dishonest advisors only adopt some simple strategies.

Only recently, the robustness issue has drawn the attention in the trust community [7, 5, 13], demanding that a trust system should be robust to all potential attacks[2]. There are trust models trying to address some specific unfair rating attacks. For example, Feng et al. [3] study three attacks, namely RepBad, RepSelf and RepTrap, and propose defenses against them. Jiang et al. [6] propose a trust model based on evolutionary computation (called MET) to effectively cope with four typical attacks and their combinations. Liu et al. [11] propose a fuzzy logic based trust model to effectively resist the attacks that exist in a cyber competition where human participants compete to break down a trust system. However, it is still difficult to say that these trust models will be robust to all possible unfair rating attacks.

Instead of looking at some specific attacks, we consider all possible attacks and identify the worst case, because a trust system is robust to unfair rating attacks if it functions well under the worst-case attack. By explicitly modelling and analyzing the worst-case unfair rating attacks, trusters can then derive accurate trust evaluation.

3. PRELIMINARIES

Our approach is mostly supported by concepts and theorems in information theory, as presented below.

Definition 1. (Shannon entropy [12]) The Shannon entropy of a discrete random variable X is given:

$$H(X) = \mathbf{E}(I(X)) = -\sum\nolimits_{x_i \in X} P(x_i) \cdot \log(P(x_i))$$

The Shannon entropy gets maximum when all possible outcomes are equiprobable. Further, it can be generalised to differential entropy for continuous random variables Y as:

$$h(Y) = \mathbf{E}(I(Y)) = -\int_Y p(y) \cdot \log(p(y))\ \mathrm{d}y$$

The Shannon entropy measures the expected amount of information carried in a random variable, which is decided by the uncertainty of the random variable. The base of the logarithm is set as 2, without loss of generality. Since $x\log(x)$ is a common term, we introduce the shortcut $\mathbf{f}(x) = x\log(x)$. For practical reasons, we let $0\log(0) = 0$.

Definition 2. (Conditional entropy [12]) The conditional entropy of discrete random variables X under Y is given as:

$$H(X|Y) = -\sum\nolimits_{y_j \in Y} P(y_j) \cdot \sum\nolimits_{x_i \in X} \mathbf{f}(P(x_i|y_j))$$

It can be generalised to continuous X and Y as:

$$H(X|Y) = -\int_Y p(y) \cdot \int_X \mathbf{f}(p(x|y))\ \mathrm{d}x\ \mathrm{d}y$$

The conditional entropy measures the expected amount of information in one random variable when another random variable is known. $H(X|Y) = H(X)$ iff X and Y are independent. For brevity, we leave out the cases where only one of X and Y is continuous. Note that $0 \leq H(X|Y) \leq H(X)$.

Definition 3. (Cross entropy [16]) The cross entropy for two distributions P and Q is given as:

$$H(P,Q) = E_P[-log(Q)] = H(P) + D_{KL}(P||Q)$$

The cross entropy measures the distance between the probability distribution the data actually follows and the distribution that is assumed. $D_{KL}(P||Q)$ named Kullback - Leibler divergence is a non-symmetric measure of the difference between distributions P and Q [18]. When $P = Q$, $H(P,Q) = H(P)$, $D_{KL}(P||Q) = 0$, which are their minimal.

Definition 4. (Information leakage) The information leakage of X under Y is given as: $H(X) - H(X|Y)$.

Information leakage is the gain of information about one random variable by learning another random variable. This definition is the same with mutual information [15]. Information leakage is zero, iff the two variables are independent.

Proposition 1. For any random variables X, Y: $H(X) - H(X|Y) = 0$ iff $P(X) = P(X|Y)$.

Theorem 1. (Jensen's inequality) For a convex function f:

$$f(\frac{\sum_i a_i \cdot x_i}{\sum_i a_i}) \leq \frac{\sum_i a_i f(x_i)}{\sum_i a_i}$$

Equality holds iff $x_1 = x_2 = \ldots = x_n$ or f is linear. Two instances of convex functions are $\mathbf{f}(x)$ and $-\log(x)$.

4. THE WORST CASE: MINIMISING INFORMATION LEAKAGE

A trust system is robust if it can function properly under all situations [7, 13]. For unfair rating attacks, we argue that a trust system is robust if it can function well under the worst case. To make a trust system robust to unfair rating attacks, we thus need to find the worst-case attacks first.

A truster aims to learn (or obtain information) from recommendations, based on which it constructs trust opinions about trustees. Note that this does not simply mean the truster would believe the recommendations. The truster can calibrate the interpretation of recommendations based on the trustworthiness or the strategies of advisors, to construct accurate trust opinions. For example, BLADE proposes to re-interpret ratings based on the evaluation functions used by advisors [17]. Therefore, whenever there is information in recommendations, there can be the way for a truster to make use of it. The worst case then is: *there is little information in recommendations, or misbehaving advisors try to minimise that information.*

Below, we quantify the worst case and analyse what kinds of recommending strategies constitute that, based on the recommendation model introduced as follows.

[2] Besides unfair rating attacks, other typical attacks are playbooks, reputation lag attack, etc. [10, 7].

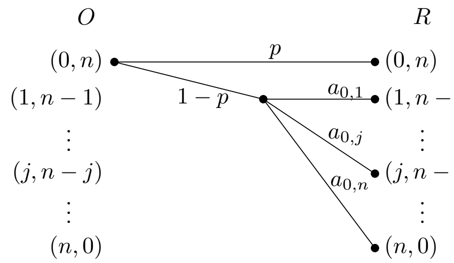


Figure 1: The naive recommendation model

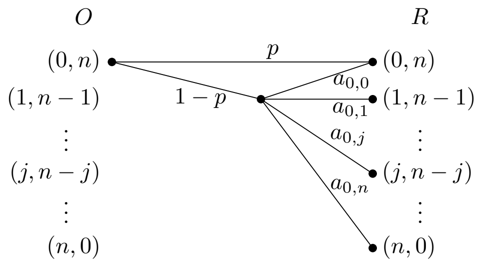


Figure 2: The extended recommendation model

4.1 Recommendation Model

There are trusters, trustees and advisors in a trust system, of which an e-marketplace is a popular example. It includes three kinds of agents: buyers, sellers and advisors. Buyers and advisors may have interactions with sellers, where the seller may deliver goods in a satisfactory manner (success) with probability T, or fail to do so (failure) with probability $1-T$, where T is called the trustworthiness (or integrity) of the seller, and is unknown to the other agents. Below, we analyse the worst-case recommendations provided by advisors to a single buyer regarding a single seller.

Figure 1 presents the set-up. We first consider the set-up with a single advisor. The advisor recommends its interaction history with the seller to the buyer. We assume that for the buyer, the number of interactions between the advisor and the seller is a known quantity, $n \in \mathbb{N}$; the only thing unknown is what fractions are successes and failures. The random variables O and R represent the true and the claimed interaction history of the advisor about the seller, respectively. We assume that before getting R, O and T have the highest uncertainty to the buyer, thus they are uniformly distributed based on the maximum entropy principle.

The advisor may not always report the truth to the buyer. We set a parameter p to describe the probability that the advisor is honest. Honesty can refer to "free of deceit" as well as "truthful". We in this paper interpret it as the former. Hence, dishonesty means that the advisor strategically gives recommendations, and we will use these two words alternatively. Correspondingly, $1-p$ represents the probability that the advisor is dishonest/strategical. Given an observation $O = (i, n-i)$, with i as the number of successful interactions, the probability that the advisor reports $R = (j, n-j)$ is $a_{i,j}$. For example, $a_{0,1}$ represents the probability that the advisor reports $R = (1, n-1)$ when $O = (0, n)$ is observed. As $R = (j, n-j), (j = 0, 1, \ldots, n, j \neq i)$ constitutes all possible recommendations when the advisor is dishonest, we have $\sum_{j \neq i} a_{i,j} = 1$. Matrix $a_{i,j}$ decides the recommending strategy of an advisor. For simplicity, below we use $O = i$, $R = j$ to represent $O = (i, n-i)$ and $R = (j, n-j)$ respectively.

The set-up with a single advisor can be generalised to multiple advisors. To find the worst-case strategy, we need to maintain the consistency and uniformity of their honesty and strategies. Thus we assign the same $n, p, a_{i,j}$ to them. Here, p ($1-p$) can also be approximately treated as the rate of honest (strategical) advisors Also, $a_{i,j}$ can be treated as the rate of advisors reporting $R = j$ when $O = i$ is observed. In this way, our analysis for a single advisor is also explainable for multiple advisors.

In this paper, we consider two types of worst-case unfair rating attacks performed by advisors: misbehaving advisors aiming at minimising (hiding) the information of their true observations, and misbehaving advisors aiming at minimising the information of the integrity of the seller.

4.2 Advisors Hiding their True Observations

This kind of advisors aims to hide their observations from the buyer. The rationale is that, by hiding the true observations, advisors make it difficult for the buyer to construct an accurate trust opinion about the seller. In the worst case, the advisor can completely hide the true observations, i.e., the recommendation is independent of the true observations.

Theorem 2. In the naive recommendation model shown in Figure 1, the recommendation R is independent of the true observation O iff $p = \frac{1}{n+1}$ and $a_{i,j} = \frac{1}{n}(i \neq j)$.

Proof If recommendation R is independent of O, then $P(R{=}j|O{=}i) = P(R{=}j|O{=}i')$, for all j, i, i'.

$$P(R{=}j|O{=}i) = \begin{cases} p & \text{if } i = j \\ (1-p)a_{i,j} & \text{if } i \neq j \end{cases} \quad (1)$$

Therefore $p = (1-p)a_{i,j}$ and $a_{i,j} = \frac{p}{1-p}$, where $i \neq j$. Since $\sum_j a_{i,j} = 1$ where $i \neq j$, $n \cdot \frac{p}{1-p} = 1$ and $p = \frac{1}{n+1}$. As the result, $a_{i,j} = \frac{1}{n}$, where $i \neq j$.

On the other hand, if $p = \frac{1}{n+1}$ and $a_{i,j} = \frac{1}{n}$, $i \neq j$, then

$$P(R{=}j|O{=}i) = \frac{1}{n+1} \quad (2)$$

$$P(R{=}j) = \sum{}_i P(O = i) \cdot P(R{=}j|O{=}i) = \frac{1}{n+1} \quad (3)$$

As $P(R{=}j|O{=}i) = P(R{=}j)$ holds for any i and j, R and O are independent. □

Intuitively, we expect that the lower values of p (less honest advisors) make it easier to hide O. However, Theorem 2 implies that when $p < \frac{1}{n+1}$ the true observations cannot be perfectly hidden, whereas for $p = \frac{1}{n+1}$, it can. Therefore, we need to alter the naive model to accommodate for the case $p < \frac{1}{n+1}$. When $p < \frac{1}{n+1}$, the independence of O and R implies $\sum_{j \neq i} a_{i,j} < 1$, which is impossible in the naive model. This is caused by the fact that the advisor is forced to lie (with n fixed) if the advisor is strategical in the naive model. Therefore, we must allow strategical/dishonest advisors to report the truth with non-zero probability. In fact, it is nature that strategical advisors may sometimes tell the truth, as part of deceit. As a real-world scenario: consider a card game with only one Ace, King, Queen – the highest wins. Alice asks her (dishonest) opponent Bob about what his card is. If Bob always lies and if he states Queen, and Alice has the King, Alice would know that Bob has the Ace. Thus, as a strategical player, Bob should sometimes report the truth to deceive Alice. Hence, here we introduce an alternative option $a_{j,j}$ (e.g., $a_{0,0}$ when $j = 0$), as depicted in the extended recommendation model in Figure 2.

Theorem 3. In the extended recommendation model shown in Figure 2, the recommendation R is independent of the true observation O iff $0 \leq p \leq \frac{1}{n+1}$ and $a_{ij} = \frac{p}{1-p} + a_{jj}$.

Proof If recommendation R is independent of O, then $P(R{=}j|O{=}i) = P(R{=}j|O{=}i')$, for all j, i, i'.

$$P(R{=}j|O{=}i) = \begin{cases} p + (1-p)a_{i,j} & \text{if } i = j \\ (1-p)a_{i,j} & \text{if } i \neq j \end{cases} \tag{4}$$

Therefore $p + (1-p)a_{j,j} = (1-p)a_{i,j}$, $a_{i,j} = \frac{p}{1-p} + a_{j,j}$. Since $\sum_{j \neq i} a_{i,j} = 1 - a_{i,i}$, $\frac{np}{1-p} + \sum_{j \neq i} a_{j,j} = 1 - a_{i,i}$, we get $\sum_j a_{j,j} = \frac{1-(n+1)p}{1-p}$. Since $\sum_j a_{j,j} \geq 0$ and $0 \leq p \leq 1$, we get $0 \leq p \leq \frac{1}{n+1}$.

On the other hand, if $0 \leq p \leq \frac{1}{n+1}$ and $a_{ij} = \frac{p}{1-p} + a_{jj}$

$$P(R{=}j|O{=}i) = P(R{=}j) = p + (1-p)a_{j,j} \tag{5}$$

holds for any i, j. Hence, R and O are independent. □

When $\sum_j a_{j,j} = 0$ and $a_{i,j} = \frac{p}{1-p}$, Theorem 3 becomes Theorem 2. Note that $\sum_j a_{j,j} > 0$ is allowed when R is independent of O, which implies even when the buyer learns nothing, still some dishonest advisors may tell the truth.

Intuitively, recommendations are only useful when less than half of the advisors are dishonest. Remarkably, Theorem 3 proves otherwise. It implies that R and O cannot be independent when $p > \frac{1}{n+1}$. This means that, for $n > 1$, over half of the advisors can be dishonest (i.e., $(1-p) > \frac{1}{2}$), yet the buyer can still learn from the recommendations.

Although no strategy can achieve the independency when $p > \frac{1}{n+1}$, some strategies are still better at hiding the true observations than others. To capture this, we generalise the measure of dependency between recommendations and true observations to information leakage (Definition 4 in Section 3). The independency of R and O holds iff R leaks zero information about O. Low information leakage about O means that O is hidden well. Below, we aim to find the strategy that minimises the information leakage for $p > \frac{1}{n+1}$. As $H(O)$ is unchangeable to the buyer, to minimise information leakage, it suffices to minimise $-H(O|R)$.

Definition 5. (Level strategy) is the strategy where: for all $0 \leq j \leq n$, $a_{j,j} = 0$, and for all $0 \leq i \neq j \leq n$, $a_{i,j} = \frac{1}{n}$.

Theorem 4. The level strategy minimises information leakage of O given R for $p \geq \frac{1}{n+1}$.

Proof Given $h_j = p + (1-p)\sum_i a_{i,j}$, $0 \leqslant i, j \leqslant n$,

$$\begin{aligned} -H(O|R) &= \sum\nolimits_j P(R{=}y_j) \sum\nolimits_i P(O{=}x_i|R{=}y_j) \log P(O{=}x_i|R{=}y_j) \\ &=^1 \frac{1}{n+1} \sum\nolimits_j \Big(\sum\nolimits_{i \neq j} (1-p) \cdot a_{i,j} \log(\frac{(1-p) \cdot a_{i,j}}{h_j}) \\ &\quad + (p + (1-p)a_{j,j}) \log(\frac{p + (1-p)a_{j,j}}{h_j}) \Big) \qquad (6) \\ &\geq^2 \frac{n}{n+1} \sum\nolimits_i \frac{(1-p)(1-a_{i,i})}{n} \log(\frac{(1-p)(1-a_{i,i})}{n}) \\ &\quad + (p + \frac{\sum_j (1-p) \cdot a_{j,j}}{n+1}) \cdot \log(p + \frac{\sum_j (1-p) \cdot a_{j,j}}{n+1}) \\ &\geq^3 p \cdot \log(p) + (1-p) \cdot \log(\frac{1-p}{n}) \end{aligned}$$

Inequality 2 is derived based on the Jensen's inequality (Theorem 1 in Section 3). Inequality 3 is derived based on the property that $xlog(x)$ is superlinear and $p \geq \frac{1}{n+1}$.

Finally, note that applying the strategy from Definition 5 to term 1 yields term 3. Thus, term 3 represents the information leakage under the level strategy. Since term 3 is the minimum, the level strategy minimises information leakage. For $p = \frac{1}{n+1}$, the level strategy leads to zero information leakage, as we proved in Theorem 2. □

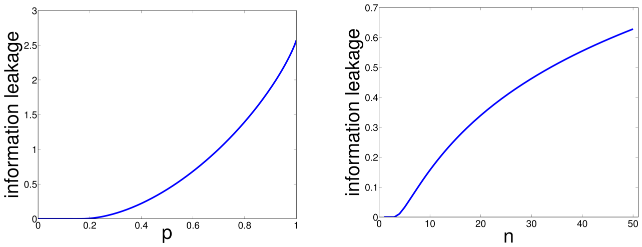


Figure 3: The minimal information leakage of O varies with p and n

In summary, we have found the strategies that minimise the information leakage about O for all $p \in (0, 1)$. Specifically, for $p < \frac{1}{n+1}$, the strategy requires a fraction of dishonest advisors to report the truth. For $p \geq \frac{1}{n+1}$, the strategy requires dishonest advisors to uniformly choose a lie. Further, zero information leakage (independency) is only achieved when $p \leq \frac{1}{n+1}$. The buyer can still get some information for $p > \frac{1}{n+1}$.

To illustrate our results, we plot the information leakage of O in the worst case, as a variable of p (with $n = 5$) and n (with $p = 0.25$), in Figure 3. From the figure, we learn that when $p \leq \frac{1}{n+1}$ or $n \leq \frac{1}{p} - 1$, the information leakage is zero. And when the difference between p and $\frac{1}{n+1}$ increases, the information leakage increases. This will further be demonstrated in the experiments in Section 5.

4.3 Advisors Hiding the Integrity of the Seller

This kind of advisors aims to hide the integrity, T, of the seller from the buyer. The rationale is that the buyer's trust opinion is about the integrity of the seller. Therefore, to make it difficult for the buyer to construct an accurate trust opinion about the seller, the advisor aims to hide information about the integrity of the seller.

Intuitively, hiding the true observations may seem equivalent to hiding the integrity of the seller. As we will prove in Theorem 6, they are not the same. However, they do coincide whenever they can avoid information leakage.

Theorem 5. There is zero information leakage of T, iff there is zero information leakage of O.

Proof From Proposition 1 in Section 3, zero information leakage of T (O) given R is equivalent to T (O) being independent of R. If O is independent of R, we have

$$\begin{aligned} P(T{=}t|R{=}j) &=^1 \sum\nolimits_i P(T{=}t|R{=}j, O{=}i) \cdot P(O{=}i|R{=}j) \\ &=^2 \sum\nolimits_i P(T = t|O = i) \cdot P(O = i|R = j) \\ &=^3 \sum\nolimits_i P(T = t|O = i) \cdot P(O = i) \qquad (7) \\ &=^4 P(T = t) \end{aligned}$$

which holds for any t, j, implying that T is independent of R. Term 2 follows because T and R are conditionally independent given O.

On the other hand, if T is independent of R, we have

$$\begin{aligned} P(O{=}i|R{=}j) &=^1 \sum\nolimits_i P(O{=}i|T{=}t, R{=}j) \cdot P(T{=}t|R{=}j) \\ &=^2 \sum\nolimits_i P(O = i|T = t) \cdot P(T = t|R = j) \\ &=^3 \sum\nolimits_i P(O = i|T = t) \cdot P(T = t) \qquad (8) \\ &=^4 P(O = i) \end{aligned}$$

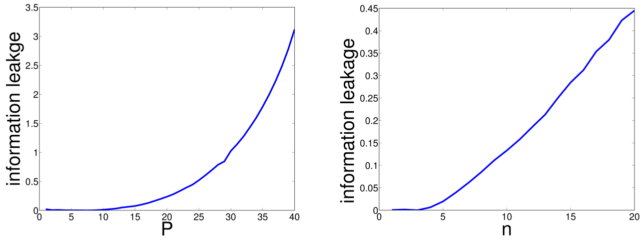


Figure 4: The minimal information leakage of T varies with p and n

which holds for any i, j, implying that O is independent of R. Term 2 follows because O and R are conditionally independent given T. Thus we prove Theorem 5. □

Note that since zero information leakage of O requires $p \leq \frac{1}{n+1}$, zero information leakage of T also requires $p \leq \frac{1}{n+1}$.

Theorem 6. The level strategy does not minimise information leakage of T, for all n, p that satisfy $p > \frac{1}{n+1}$.

Proof It suffices to provide a counterexample. For $n = 2$, $p = \frac{2}{3}$, using the level strategy, we obtain $-H(T|R) = 0.2192$. When we set

$$a = \begin{pmatrix} 0 & 0.2938 & 0.7063 \\ 0.4922 & 0.0156 & 0.4922 \\ 0.7063 & 0.2938 & 0 \end{pmatrix},$$

$-H(T|R) = 0.1934$. Since $0.1934 < 0.2192$, the level strategy does not minimise information leakage of T. □

Below, we aim to find the lying strategy that minimises information leakage of T given R when $p > \frac{1}{n+1}$. As $H(T)$ is unchangeable, it suffices to minimise $-H(T|R)$.

$$\begin{aligned} -H(T|R) &= -\sum\nolimits_j P(R=j) H(T|R=j) \\ &= \sum\nolimits_j P(R=j) \int_0^1 f_T(t|R=j) \cdot \log f_T(t|R=j)\ \mathrm{d}t, \end{aligned}$$

where $P(R = j)$ as before, and

$$\begin{aligned} f_T(t|R=j) &= \sum\nolimits_i f_T(t|O{=}i, R{=}j) \cdot P(O{=}i|R{=}j) \qquad (9) \\ &= \sum\nolimits_i f_\beta(t; i{+}1, n{-}i{+}1) \cdot \begin{cases} \frac{p+(1-p)a_{j,j}}{h_j} & \text{if } i = j \\ \frac{(1-p)a_{i,j}}{h_j} & \text{if } i \neq j \end{cases} \end{aligned}$$

Note that $P(O = i|R = j)$ is the posteriori probability about O known R, which can be computed from $P(R = j|O = i)$ based on Bayes' theorem. And $P(R = j|O = i)$ is decided by the recommending strategy.

For our analysis, we use a local search heuristic to find good strategies for the advisors. Our heuristic is initialised with the level strategy. We iterate over all $a_{i,j}$, where, for each $a_{i,j}$, we increase $a_{i,j}$ with a fixed value (at the expense of the other $a_{i,j'}$) until $-H(T|R)$ stops decreasing. We perform the iteration multiple times, with decreasing step sizes. In the limit, the heuristic is a gradient search.

To illustrate the analysis above for T, we plot the information leakage of T in the worst case, as a variable of p (with $n = 5$) and n (with $p = 0.25$), in Figure 4. From the figure, we learn that when $p \leq \frac{1}{n+1}$ or $n \leq \frac{1}{p} - 1$, the information leakage is zero. And when the difference between p and $\frac{1}{n+1}$ increases, the information leakage increases. This will also be further demonstrated in the experiments in Section 5.

4.4 Induced Trust Computation (ITC)

Given the worst-case strategies of the advisors, the buyer can construct accurate trust opinions. A trust opinion is a distribution $f_T(t|\phi)$, where ϕ consists of the knowledge of the buyer (direct experiences and recommendations) [8].

In [14], the authors prove the following theorem under the assumptions that if recommendations and observations are conditionally independent given the strategies of the sellers and advisors, and that their strategies are independent:

Theorem 7. For any collection of recommendations and direct observations φ and ψ, $f_T(t|\varphi, \psi) \propto f_T(t|\varphi) \cdot f_T(t|\psi)$.

With Theorem 7, the knowledge of the buyer can be broken down into cases for which we have explicit computations. The case where the knowledge of the buyer is direct experience, has already been solved [8]. If the knowledge of the buyer is a single recommendation, then $\phi = R$ and the trust opinion is $f_T(t|R)$. In the worst-case attack, $f_T(t|R)$ can be computed known the strategy (matrix $a_{i,j}$) of the advisors based on Equation (9).

Note that the accuracy of computing $f_T(t|S)$ is influenced by the accuracy of p. As a description of the trustworthiness of an advisor, p is usually estimated by the trust models (as done by TRAVOS [20] and many other classic models [21, 19]). *In this work, we are not trying to build a new robust trust model. We are solving a sub-problem of defending the worst-case unfair rating attack to make a trust model more robust.* Hence, we simply assume that p is already accurately estimated by the trust models. In fact, we also demonstrate through experimentation in Section 5 that even when p is not entirely accurately estimated by the trust models (e.g., TRAVOS and MET), their robustness can still be improved by our ITC method.

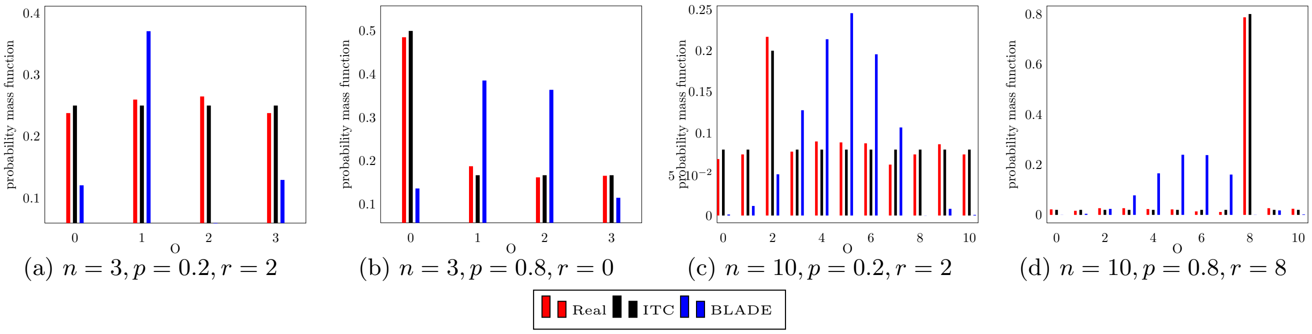
In this way, by being aware of the worst-case strategies in advance, the buyer gains the initiative to derive accurate trust opinions under the worst case.

5. ROBUSTNESS ANALYSIS

As surveyed in Section 2, TRAVOS [20], BLADE [17] and MET [6] are three state-of-the-art trust models to address the unfair rating problem, where TRAVOS and BLADE assume some simple attacking strategies for advisors but MET tries to cope with some typical attacks and their combinations. In this section, we evaluate the robustness of these trust models, and more importantly to demonstrate that our induced trust computation (ITC) method can further improve the robustness of these trust models.

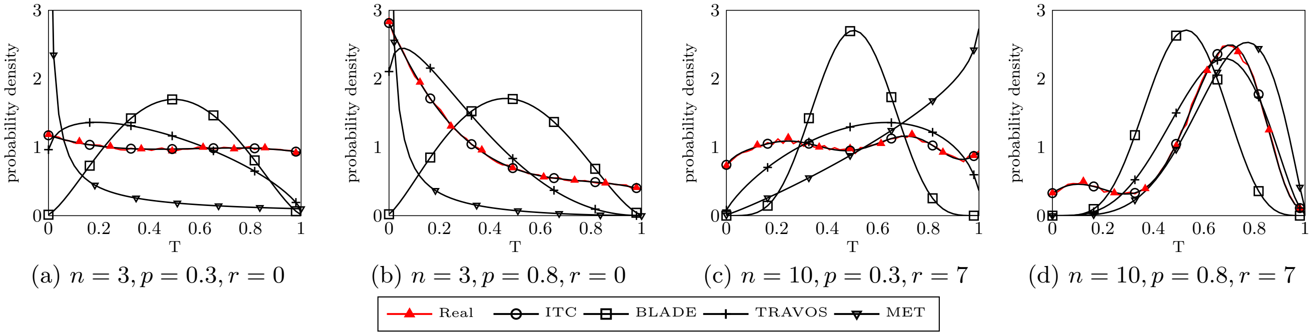
More specifically, we conduct a set of experiments based on simulations[3]. In the first experiment, we compare the trust opinions about sellers that the trust models and ITC construct, under two types of the worst-case attacks: advisors hiding true observations (O) and hiding seller integrity (T). Because ITC always assumes the worst case, to have more fair comparison, in the second experiment, we compare the accuracy of trust opinions given by ITC and the three models, under other random attacking strategies which are not the worst case. Modelling the honesty of advisors accurately is not the focus of this work, hence in the first two experiments, we simply assume that the honesty of advisors is accurately estimated by all trust models, which appears

[3] We did not use existing testbeds such as the ART testbed [4] because they are often only used to study the quality of expectations about trust evaluation.



(a) $n = 3, p = 0.2, r = 2$ (b) $n = 3, p = 0.8, r = 0$ (c) $n = 10, p = 0.2, r = 2$ (d) $n = 10, p = 0.8, r = 8$

Figure 5: Comparing predictions on distributions of O [best viewed in color]



(a) $n = 3, p = 0.3, r = 0$ (b) $n = 3, p = 0.8, r = 0$ (c) $n = 10, p = 0.3, r = 7$ (d) $n = 10, p = 0.8, r = 7$

Figure 6: Comparing predictions on distributions of T

as a same parameter (p) to all models. In the third experiment, we further study whether our ITC method can improve the robustness of the trust models given whatever advisor honesty output by these trust models.

All of the simulations above rely on the true behaviour of the seller. To address this, we run the Monte Carlo simulation. In each run, $t \in [0, 1]$ is uniformly randomly chosen as a sample of T for the seller. Then, n Bernoulli samples are drawn with the probability t, which provides us an o as the true value of O. Based on o and an advisor's strategy a, a recommendation r is generated as the true value of R (recommendations of the advisor). The trust models are provided with recommendation r, which is used to construct the trust opinion about the seller.

5.1 Under the Worst Case

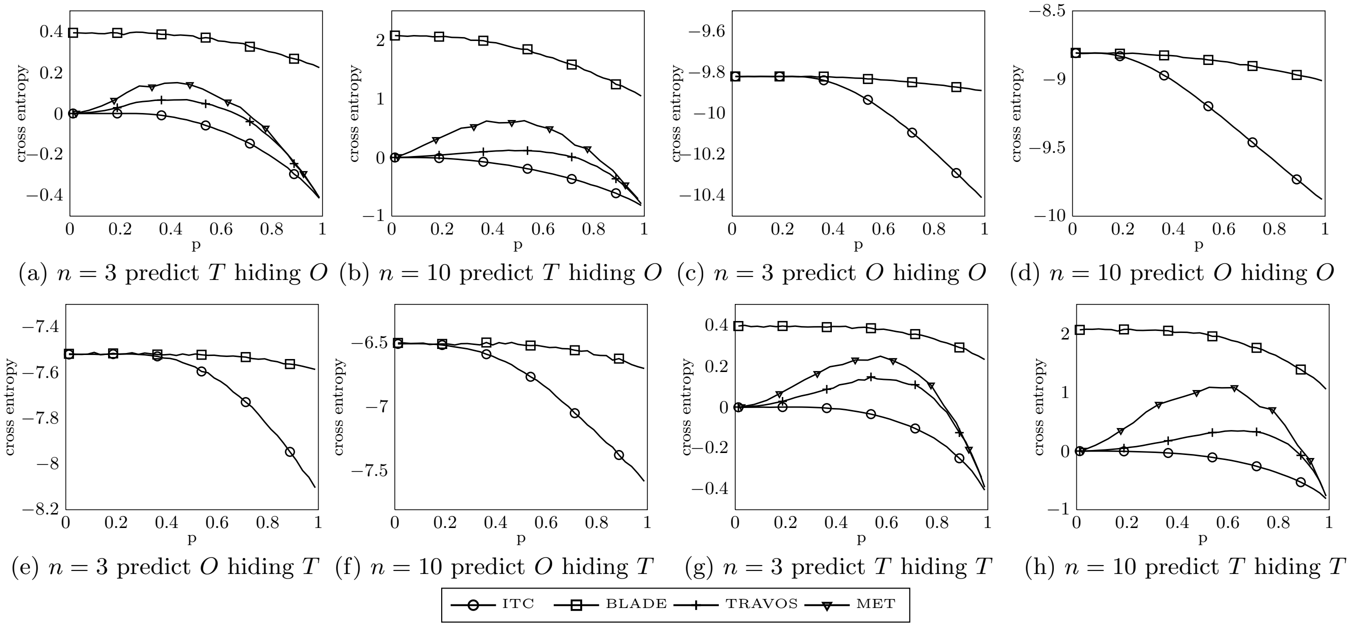
In the first experiment, we compare the predictions on T and O against the truth, under the worst-case strategies of hiding T and O respectively. The values for parameters n, p, r^* are manually chosen as the number of transactions, probability of advisor honesty, and recommendation. We then run the simulation, but reject the sample of T (and the corresponding sample of O) if the resulting $R{\neq}r^*$. In this way, we get the true probability distributions of T and O: $P(T|R{=}r^*)$ and $P(O|R{=}r^*)$, which are used compare with that predicted by TRAVOS, BLADE, MET and ITC.

For comparison about O, we select four groups of values for n, p, r^*: $(3, 0.8, 0)$, $(3, 0.2, 2)$, $(10, 0.8, 8)$, $(10, 0.2, 2)$. Figure 5 presents the results. TRAVOS and MET are not considered here, as they do not generate the prediction of O. The predictions of ITC have much smaller difference with the real distributions compared with BLADE. Larger p leads to more converged predictions. Comparing Figures 5(a) and 5(c), although $p{=}0.2, s{=}2$ are the same, prediction of ITC given $n{=}10$ is converged on $O{=}2$ while that given $n{=}3$ is uniformly distributed. According to the theoretical proof in the former section, when $n{=}3$, $p < \frac{1}{n+1} = 0.25$, there is no information leakage of O under the worst-case attack. Hence, ITC predicts maximum uncertainty of O.

For comparison about R, we select four groups of values for parameters n, p, r^*: $(3, 0.8, 0)$, $(3, 0.3, 0)$, $(10, 0.8, 7)$, $(10, 0.3, 7)$. Figure 6 presents the results. The probability distributions of T predicted by ITC are much closer to the real distributions than that of TRAVOS, BLADE and MET. For ITC, TRAVOS and MET, the shapes of predicted distributions are mainly decided by p and r, while BLADE is largely influenced by n instead. Comparing Figures 6(a,c) with 6(b,d), larger p leads predictions of ITC and TRAVOS to be more converged and aligned with the recommendations, because the buyer tends to believe the advisor more.

Figures 5 and 6 are restricted to a fixed r, n, p and a. To make more meaningful comparisons, we use cross entropy (Definition 3 in Section 3) to measure the quality of a prediction so that we can compare a multitude of outcomes simultaneously. In a good prediction, cross entropy is low. We generate a true integrity of a seller t, a true observation o and a recommendation r in each run, and r is used as input for the models to yield a trust opinion about the seller. To generate the graphs, we let $n = 3$ and $n = 10$, and let $0 < p < 1$ be the x-axis. We study four scenarios: predicting O (T) under the worst-case strategies of hiding O (T), and predicting T (O) under the worst-case strategy of hiding O (T). Because TRAVOS and MET do not output predictions of O, they do not appear in Figure 7(c-f). Figure 7 provides the following information.

First, when $p \leq \frac{1}{n+1}$, all the ITC graph segments are flat, meaning that uniform distribution is predicted. This corroborates our proofs: when $p \leq \frac{1}{n+1}$, there is no information leakage about T (O) given R in the worst case, thus $H(T|R)$ (or $H(O|R)$) reaches the maximum, which implies uniform distribution. Note that for continuous distributions, the uniform distribution has entropy zero, explaining why ITC has



(a) $n = 3$ predict T hiding O (b) $n = 10$ predict T hiding O (c) $n = 3$ predict O hiding O (d) $n = 10$ predict O hiding O

(e) $n = 3$ predict O hiding T (f) $n = 10$ predict O hiding T (g) $n = 3$ predict T hiding T (h) $n = 10$ predict T hiding T

Figure 7: Comparing accuracy of predicting T (or O) under the worst-case of hiding T (or O)

cross entropy of 0 for small p, in Figures 7(a, b, g, h). In Figures 7(c-f), the uniform distribution is over discrete variables, meaning that the entropy depends on n, which explains the difference in cross entropy for value p near 0.

Second, when $p > \frac{1}{n+1}$, ITC shows lower cross entropy than BLADE, TRAVOS and MET, for equal p and n. Moreover, we can identify the trends that BLADE and ITC have decreasing cross entropy over p (and n), whereas for TRAVOS and MET, the cross entropy increases before it decreases, over p. The reason that ITC is decreasing, is simply because $H(T|R)$ (and $H(O|R)$) are decreasing over p. Recall (Definition 3) that cross entropy is the entropy of the truth plus KL-divergence, and that ITC has KL-divergence of 0, because it computes correct $H(T|R)$ (and $H(O|R)$) by knowing p and the worst-case strategies of advisors. TRAVOS and MET first increase because they over-predict – causing to assign unreasonably low probability to unlikely events (as shown in Figure 6). As p tends to 1, their over-predictions start to match the true distribution. BLADE suffers the same problem of over-predicting. However, its over-predicting is not linked with p. Therefore, we observe a decreasing cross-entropy, as reality tends towards more polarised outcomes. Note that using the same real p value, the accuracy of TRAVOS is higher than MET, indicating that the method of aggregating recommendations in TRAVOS is better than that of MET under the worst-case attack. In fact, MET adopts a simple weighted average method to aggregate advisors' recommendations.

Third, when p is close to 1, the curves of TRAVOS, MET and ITC with the same n get to converge at a same point. With p being close to 1, nearly all of advisors report the truth. The predictions of TRAVOS and MET thus get closer to the truth, which is the prediction of ITC.

From the analysis above, it is obvious that our prediction of T (O) is much more accurate than TRAVOS, BLADE and MET under the worst case. Using our ITC method could improve the robustness of these trust models.

5.2 Under other Attacks

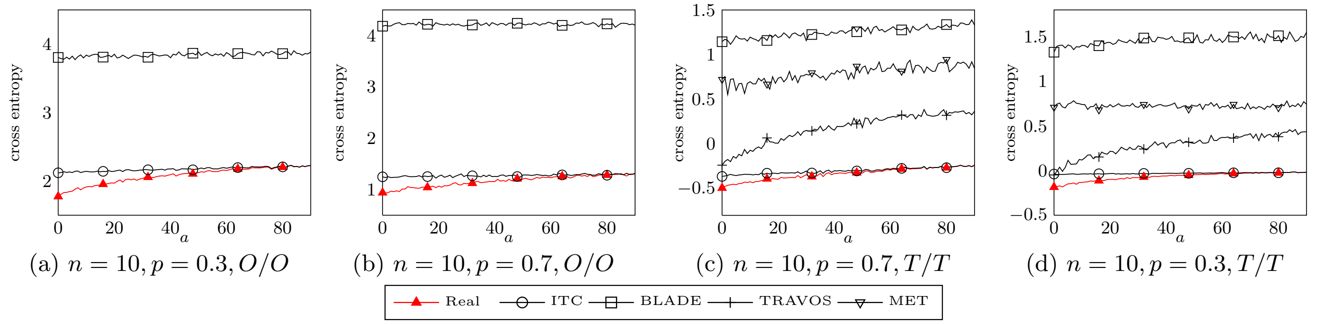
The real strategies of advisors cannot be known. To always assume the worst case is a safe choice, but may not be the most accurate choice. Hence, we investigate the performance of ITC, which assumes the worst-case strategies, under other types of attacks. Recall that a strategy of an advisor is represented as a matrix $a_{i,j}$ where $0 \leq i, j \leq n$ (see Section 4.1). We randomly generate ninety such strategies. Then, these strategies are combined with the worst-case strategy by assigning the worst case a weight varying from 0 to 1. In so doing, the strength of the resulting strategies approximately increases. We then compare the cross entropy regarding the predictions of T (O) given by ITC, TRAVOS, BLADE and MET, under all of these strategies. Figure 8 presents the result.

For the truth (the red line), the cross entropy is equal to the entropy of true distribution of T (O) given R because KL-divergence is 0. As the recommending strategy tends to be worse, the entropy of T (O) given R increases towards the maximum, which is exactly the worst case. In Figure 8, ITC has much smaller cross entropy with the truth, compared to the three models, indicating that ITC predicts much closer to the truth. And as the generated attacking strategy gets closer to the worst case, ITC predicts more and more accurately. Notice that there is little variance in the cross entropy of BLADE and MET as the attacking strategies change, implying that their performance does not change much for all those strategies. On the other hand, the cross entropy of TRAVOS increases as the attacking strategy gets closer to the worst case, showing that the performance of TRAVOS gets worse as the attacks become stronger.

From this experiment, even always assuming the worst case, our ITC method can still improve the robustness of the trust models against various other types of attacks.

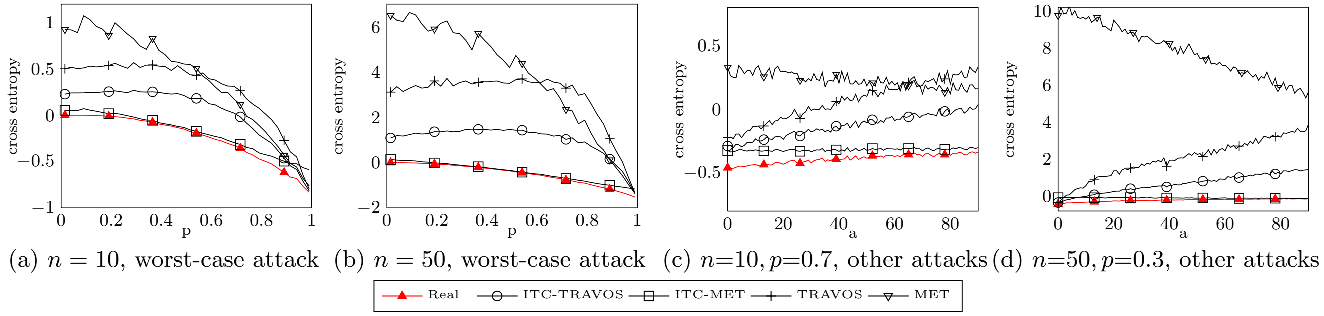
5.3 Inaccurate Estimation of Advisor Honesty

The above experiments are conducted by assuming the accurate estimation of advisor honesty (i.e., true p). In this experiment, we investigate how ITC performs when p is predicted by other trust models, which may not be completely accurate. BLADE does not estimate p, so we only compare the accuracy of ITC (ITC-TRAVOS and ITC-MET)



(a) $n = 10, p = 0.3, O/O$ (b) $n = 10, p = 0.7, O/O$ (c) $n = 10, p = 0.7, T/T$ (d) $n = 10, p = 0.3, T/T$

Figure 8: Under various other types of attacks. O/O: predict O by hiding O; T/T: predict T by hiding T



(a) $n = 10$, worst-case attack (b) $n = 50$, worst-case attack (c) n=10, p=0.7, other attacks (d) n=50, p=0.3, other attacks

Figure 9: Using p estimated by other trust models under the worst-case and other types of attacks

with TRAVOS and MET, based on their predicted p respectively. We consider the prediction of seller integrity T under two scenarios: 1) the worst-case strategy of hiding O, with the real p value varying from 0 to 1 (Figure 9(a-b)); 2) other types of attacks with $p = 0.7$ and $p = 0.3$ (Figure 9(c-d)).

Based on p predicted by the corresponding trust models, ITC still has much higher accuracy indicated by the lower cross entropy of ITC-TRAVOS and ITC-MET as shown in the figure, confirming that ITC can effectively improve the robustness of TRAVOS and MET even when the estimation of p may not be entirely accurate, and when the advisor attacking strategies may not be the worst case.

Similar as the results in Figure 7, larger p leads to more accurate prediction because the advisor is more trustworthy. In addition, when the estimation of p is more accurate, the prediction of seller integrity T should also be more accurate. With this, compare ITC-TRAVOS and ITC-MET. ITC performs better when using p output by MET than when using p from TRAVOS, indicating that MET predicts the honesty of advisors more accurately than TRAVOS. This is also supported by the results in [6]. However, with the p value from MET, ITC cannot accurately predict the truth even under the worst case (see Figure 9(a-b)), indicating that advisor honesty estimated by MET is not completely accurate.

On the other hand, TRAVOS performs better than MET when $p < 0.6$ in Figure 9(a-b,d). Also, recall the results in Figure 7 where given the same true p, the predictions of TRAVOS are more accurate than MET. These results indicate that TRAVOS has a nice method for aggregating recommendations from the advisors. However, when $p > 0.6$, MET outperforms TRAVOS, indicating that when the advisors are more trustworthy, the effect of that method becomes less important. This can also be observed from Figure 9(c) that when $p = 0.7$ and under the worst-case attack (attack #90), MET provides more accurate prediction than TRAVOS. In fact, for other types of attacks that are close to the worst case, MET also outperforms TRAVOS.

6. CONCLUSION AND FUTURE WORK

In this work, we used information theory to measure how helpful recommendations are to trusters that receive them. A fraction of advisors giving recommendations is dishonest: attackers. We identified and analysed which attacking strategies reduce the overall helpfulness of recommendations. Our techniques and results can increase the robustness of existing trust models against unfair rating attacks.

We introduced two information theoretic measures for the quality of a recommendation, concerning how much a recommendation by an advisor reveals about the true observations of that advisor and about the true integrity of the trustee, respectively. We find that the two measures coincide iff recommendations reveal nothing; that the recommendations cannot always reveal nothing, even with more attackers than honest advisors; and that it may be rational for an attacker to report the truth, to obscure the truth.

We derived how to compute trust opinions, assuming the worst-case attacking strategies. The results of our experiments show that our method's predictions are more accurate than TRAVOS, BLADE and MET, meaning our method is more robust, and more importantly that our method complements the trust models in improving their robustness.

We have not yet considered collusion between agents. It is a difficult problem, for which our approach of defining the worst case may help. We will investigate this potential.

Acknowledgement

This research is supported (in part) by the National Research Foundation, Prime Minister's Office, Singapore under its National Cybersecurity R & D Program (Award No. NRF2014NCR-NCR001-30) and administered by the National Cybersecurity R & D Directorate. This research is also partially supported by "Formal Verification on Cloud" project under Grant No: M4081155.020, and the ASTAR / I2R - SERC, Public Sector Research Funding (PSF) Singapore (M4070212.020) awarded to Dr. Jie Zhang.

7. REFERENCES

[1] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proceedings of the Second ACM Conference on Electronic Commerce (EC)*, 2000.

[2] H. Fang, J. Zhang, and N. Thalmann. Subjectivity grouping: Learning from users' rating behavior. In *Proceedings of the 13th International Autonomous Agents and Multi Agent Systems (AAMAS)*, 2014.

[3] Q. Feng, Y. L. Sun, L. Liu, Y. Yang, and Y. Dai. Voting systems with trust mechanisms in cyberspace: Vulnerabilities and defenses. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(12):1766–1780, 2010.

[4] K. K. Fullam, T. B. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. S. Barber, J. S. Rosenschein, L. Vercouter, and M. Voss. A specification of the agent reputation and trust (art) testbed: experimentation and competition for trust in agent societies. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 512–518, 2005.

[5] K. Hoffman, D. Zage, and C. Nita-Rotaru. A survey of attack and defense techniques for reputation systems. *ACM Computing Surveys (CSUR)*, 42(1):1, 2009.

[6] S. Jiang, J. Zhang, and Y.-S. Ong. An evolutionary model for constructing robust trust networks. In *Proceedings of the 12th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2013.

[7] A. Jøsang. Robustness of trust and reputation systems: Does it matter? In *Proceedings of the 6th IFIP International Conference on Trust Management (IFIPTM)*, 2012.

[8] A. Josang and R. Ismail. The beta reputation system. In *Proceedings of the 15th bled electronic commerce conference*, pages 41–55, 2002.

[9] A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[10] R. Kerr and R. Cohen. Smart cheaters do prosper: defeating trust and reputation systems. In *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 993–1000, 2009.

[11] S. Liu, H. Yu, C. Miao, and A. C. Kot. A fuzzy logic based reputation model against unfair ratings. In *Proceedings of the 12th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2013.

[12] R. J. McEliece. *Theory of Information and Coding.* Cambridge University Press New York, NY, USA, 2nd edition, 2001.

[13] T. Muller, Y. Liu, S. Mauw, and J. Zhang. On robustness of trust systems. In *Proceedings of the 8th International Conference on Trust management (IFIPTM)*, 2014.

[14] T. Muller and P. Schweitzer. On beta models with trust chains. In *Proceedings of the 7th International Conference on Trust management (IFIPTM)*, 2013.

[15] A. Papoulis and S. U. Pillai. *Probability, random variables, and stochastic processes.* Tata McGraw-Hill Education, 2002.

[16] K. Plunkett and J. L. Elman. *Exercises in rethinking innateness: A handbook for connectionist simulations.* MIT Press, 1997.

[17] K. Regan, P. Poupart, and R. Cohen. Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, 2006.

[18] K. Solomon. Letter to the editor: The kullbackÍcleibler distance. *The American Statistician*, 41(4):338–341, 1987.

[19] W. T. L. Teacy, M. Luck, A. Rogers, and N. R. Jennings. An efficient and versatile approach to trust and reputation using hierarchical bayesian modelling. *Artificial Intelligence*, 193:149–185, 2012.

[20] W. T. L. Teacy, J. Patel, N. R. Jennings, and M. Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.

[21] J. Zhang and R. Cohen. Evaluating the trustworthiness of advice about seller agents in e-marketplaces: A personalized approach. *Electronic Commerce Research and Applications*, 7(3):330–340, 2008.