Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Is It Harmful When Advisors Only Pretend to Be Honest?

Dongxia Wang and **Tim Muller** and **Jie Zhang** and **Yang liu**
School of Computer Engineering, Nanyang Technological University, Singapore
wang0915@e.ntu.edu.sg, {tmuller,zhangj,yangliu}@ntu.edu.sg

Abstract

In trust systems, unfair rating attacks – where advisors provide ratings dishonestly – influence the accuracy of trust evaluation. A secure trust system should function properly under all possible unfair rating attacks; including dynamic attacks. In the literature, camouflage attacks are the most studied dynamic attacks. But an open question is whether more harmful dynamic attacks exist. We propose random processes to model and measure dynamic attacks. The harm of an attack is influenced by a user's ability to learn from the past. We consider three types of users: blind users, aware users, and general users. We found for all the three types, camouflage attacks are far from the most harmful. We identified the most harmful attacks, under which we found the ratings may still be useful to users.

Introduction

Trust systems help users select trustworthy partners (*targets*), by evaluating trustworthiness based on direct experiences and ratings from other users (*advisors*). For example, trust systems are applied in e-marketplaces to select honest sellers (Regan, Poupart, and Cohen 2006; Xiong and Liu 2004), and in secure routing for wireless sensor networks to help select reliable nodes (Shaikh et al. 2009; Bao et al. 2012). Various attacks have been found in (popular) trust systems, threatening their security (Kerr and Cohen 2009). A well-known type of attacks are unfair rating attacks, where malicious advisors (*attackers*) provide dishonest ratings (Şensoy et al. 2009). Unfair rating attacks influence the accuracy of trust evaluation, and accordingly, decision making of users (Jøsang 2010).

There are various formulations of the unfair rating attacks in the literature (Sun et al. 2006; Vogiatzis, MacGillivray, and Chli 2010; Hoffman, Zage, and Nita-Rotaru 2009). Most commonly unfair rating attacks are defined statically; meaning attackers' strategies are independent of history. For instance, in the ballot-stuffing attacks, attackers are assumed to always report positively (Yu and Singh 2003; Yu et al. 2012) regarding specific targets, regardless of when they are asked for the ratings.

A secure trust system must also function properly under dynamic unfair rating attacks – where attackers' strategies are influenced by the rating history. The commonly studied form of dynamic attacks are camouflage attacks – where attackers pretend to be honest by rating strategically (Jiang, Zhang, and Ong 2013; Kamvar, Schlosser, and Garcia-Molina 2003). To only be able to defend against camouflage attacks cannot guarantee the security of a trust system, however. In reality, it is hard to predict what dynamic attacks would happen to a trust system. Hence, it is important to know whether there are more harmful dynamic attacks compared to camouflage attacks.

In this paper, we focus on modelling dynamic attacks and measuring their harm. We apply the information-theoretic measurement of harm (*information leakage*)which we introduced in (Wang et al. 2015a). The harm of an attack depends on how much information a user can gain about the real observations of the attacker, which influences the difficulty to construct an accurate trust opinion. Such a user-based measurement is independent of specific trust systems. As we will discuss later, only static attacks are considered in (Wang et al. 2015a), which has fundamentally different assumptions.

For dynamic unfair rating attacks, we introduce a stochastic process-based model. The hindsight of a user about past ratings influences the information it can gain from a dynamic attack. Hence, we distinguish three cases: (1) the user cannot determine whether the attacker reported the truth (*blind users*), (2) the user can accurately determine whether the attacker reported the truth (*aware users*), and (3) the user can determine whether the attacker reported the truth with limited accuracy (*general users*). For each type of users, we derive the harm of dynamic unfair rating attacks, and relate it to the harm of merely pretending to be honest.

We found that initially purely pretending to be honest is not the most harmful against any types of users. For blind users who have no hindsight, camouflage attacks are no more harmful than static attacks. Even for aware users, who can perceive honest behaviour, it is not the worst to be cheated by disguised honesty. We found that, for all users, to cause the maximum harm, attackers must be honest with smaller probability than to lie, even initially. Furthermore, we found that a non-blind user can always gain some information, provided the percentage of attackers is below an exponentially large threshold. The most harmful dynamic attacks have not been introduced in the literature.

Background

Unfair rating attacks are commonly studied in trust systems (Jøsang, Ismail, and Boyd 2007). Three types of unfair rating attacks are typically considered in the literature: ballot-stuffing – for some targets, attackers always provide positive ratings (e.g., a number of buyers in eBay are bribed to rate an unreliable seller highly), bad-mouthing – for some targets, attackers always provide negative ratings – and lying – for some targets, attackers always report the opposite of their true observations (Wang et al. 2015b). These attacks are all static, with attackers' behaviour depends on the specific targets and is independent of the rating time.

We refer to unfair rating attacks where time influences attackers' behaviour as dynamic. Dynamic attackers' behaviour is closely related to their rating history. Camouflage attacks, where malicious advisors camouflage themselves as honest, are popularly studied dynamic attacks[1]. E.g. in e-commerce, some raters first provide reliable suggestions regarding arbitrary sellers, to gain the trust of a buyer, then they cheat the buyer by unfairly recommending colluding sellers. Below, we call an attack with k honest ratings followed by lies the k-camouflage attack.

Being able to defend against specific dynamic attacks can ensure the security of a trust system to some extent. In reality, it is difficult to know accurately what attacks will happen to the system. Hence, it is worth to study whether more harmful attacks exist compared with camouflage attacks.

Our methodology extends that of (Wang et al. 2015a), which is based on information theory:

Definition 1 (Shannon entropy (McEliece 2001))**.** *Let* X, Y *be discrete random variables. Let* $\mathbf{f}(z) = z\log_2(z)$.
Shannon entropy *of* X *is:* $H(X) = -\sum_{x_i \in X} \mathbf{f}(P(x_i))$.
Conditional entropy *of* X *given* Y *is:*
$$H(X|Y) = -\sum_{y_j \in Y} P(y_j) \cdot \sum_{x_i \in X} \mathbf{f}(P(x_i|y_j)).$$
Information leakage *of* X *given* Y *is:* $H(X) - H(X|Y)$.

The information leakage measures the amount of information (uncertainty) about X provided (reduced) by knowing Y, which coincides with mutual information (Cover and Thomas 1991). Considering a trust system, where an advisor provides ratings about a target to a user, a rating should serve to provide information about the advisor's observation about the target. Measuring the information leakage between the advisor's observation and its rating would enable us to quantify how much information the user can gain.

As a shortcut, for random variable X, we use x for its outcomes, and $p(x)$ to mean $p(X{=}x)$. We may write $\overline{X_i}$ to mean $X_i, \ldots, X_1$, or an empty list, when $i = 0$.

Modelling Dynamic Attacks

In a trust system, advisors provide (binary) ratings to a user. The user wants to learn the observations of advisors about a target. An honest advisor's ratings coincide with his observations. For dishonest advisors (attackers), ratings diverge from their observations. Attackers may apply various rating strategies. In dynamic unfair rating attacks, attackers' strategies have memory, hence we need states in attack modelling.

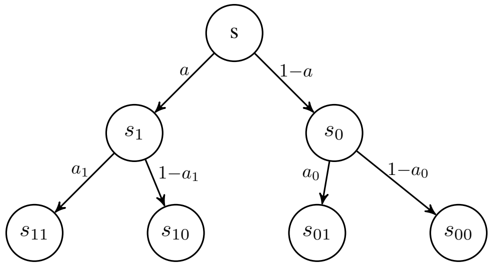

Figure 1: Attacker strategies for two ratings.

The random variables O_i and R_i represent the observation and rating of an advisor in the i^{th} iteration (or about i^{th} target). O_i only depends on the target's integrity, but R_i depends on both the honesty and strategies of the advisor. We introduce a random variable P to represent honesty of the advisor. The probability $p = p(P{=}1)$ represents the probability that the advisor is honest, and $1 - p = p(P{=}0)$ that he is dishonest. Hence, $p(R_i = O_i|P = 1) = 1$, for all i. However, $p(R_i = O_i|P = 0)$ depends on the strategies of the attacker.

Let B_i be sets of binary strings of length i, B_i^1 (B_i^0) the subset of strings ending in 1 (0), and $\hat{b}_i$ the string consisting of i 1's. We denote concatenation of b and c as bc. Finally, $b(i)$ refers to the i^{th} bit of b.

Let $\mathcal{S} = \{s_b | b \in B\}$ be a set of states. We introduce a random process $\{S_i : i \in n\}$ to model a dynamic attack of size n. An outcome s_b of random variable S_i represents the state of the attacker after the i^{th} rating, which is modelled as its behaviour history after the rating. Formally, $p(O_i = R_i|S_i = s_b) = b(i)$. The transition probability $p(S_{i+1} = s_c|S_i = s_b)$ is a_b if $c{=}b1$, $1{-}a_b$ if $c{=}b0$, and 0 otherwise. The random process $\{S_i : i \in n\}$ is a Markov chain.

As an example, the model of attacks of size 2 is shown in Figure 1. In the initial state, s, no rating occurred. In the state s_{01}, for instance, the attacker lied in the first rating and told the truth in the second rating. The probability of reaching s_{01} is $(1 - a)a_0$, which we shorthand to α_{01}. Formally, letting $\pi(x, 1) = x$ and $\pi(x, 0) = 1 - x$, we set $\alpha_b = \pi(a, b(1)) \cdot \pi(a_{b(1)}, b(2)) \cdot \ldots \cdot \pi(a_{b(1)b(2)\ldots b(i-1)}, b(i))$. Simple algebra shows that $p(s_b|P{=}0) = \alpha_b$.

As we focus on the behaviour of malicious advisors but not targets, we assume maximum uncertainty (entropy) for targets' integrity (or $\overline{O_n}$), meaning O_i are independent, and $p(o_i){=}1/2$, $H(O_i){=}1$ for all i. Under this assumption, targets' integrity is invariant and will not influence the computation of information leakage over time. Furthermore, $p(r_i){=}\sum_{o_i} p(o_i)p(r_i|o_i){=}1/2$ for $r_i{\in}\{0, 1\}$. As r_i and r_j $(i{\neq}j)$ are independent without any observations, $p(\overline{r_i}){=}1/2^i$.

Harm of dynamic attacks In (Wang et al. 2015a), we propose to measure the harm of static unfair rating attacks based on information theory. Specifically, the harm of attacks depend on the information leakage of advisors' ratings about their observations. An attack with more information leak-

[1] Our usage of the term camouflage attacks refers to advisors. We do not consider camouflaging targets (e.g. sellers), as in (Oram 2001) or "karma suicide attacks" described in (Glynos et al. 2008).

age is less harmful to a user, since it makes it easier to construct accurate trust opinions (more intuitions can be found in (Wang et al. 2015a)). We apply such a measurement of harm to dynamic attacks.

Strategies of dynamic attackers are closely related to their rating history, implying their ratings over time may be correlated, which may provide extra information to a user. Whereas in a static attack, ratings given in different iterations can be treated as independent of each other, for dynamic attacks, we must measure information leakage over the sequence of ratings, to capture this extra information. We use I_i to represent the information leakage of the rating in i^{th} iteration. The computation of I_i is influenced by ratings received in the past ($\overline{R_{i-1}}$) and the user's hindsight, as will be presented and explained in the following sections. The information leakage of an attack of size n is $\sum_{1\leqslant i\leqslant n} I_i$.

Measuring Dynamic Attacks

In this section, we study the harm of dynamic attacks based on the model, and the measurement introduced in the last section. Especially, we will figure out whether commonly studied camouflage attacks are the most harmful, and if not, what would be the most harmful attacks.

In dynamic attacks, a user's ability to judge an attackers' past behaviour influences its perception of the future behaviour. From an information theoretical perspective, the information a user can gain from future ratings is influenced by its ability to judge past ratings. Hence, the harm of an attack is closely related to the judging ability of the user.

We study three types of users in total: blind users, aware users, and general users. They differ only in abilities in judging the truthfulness of past ratings. Blind users are those who cannot distinguish truth from lies. Aware users are those who can accurately distinguish truth from lies. General users are those with limited judging abilities. The first two types are extreme cases of general users.

We study each type of users in a subsection. In each subsection, we start with dynamic attacks of size 2 as a running example. We present the information leakage for the appropriate users. Based on the definition, we then prove what would be the most harmful attacks of size 2, and some theoretical properties. Then, we generalise all those results to dynamic attacks of arbitrary size.

Attacks against Blind Users

In this section, we study cases where a user is completely unable to judge whether the attacker has told the truth in the past. This happens when the user does not or cannot verify the ratings. We call these users *blind users*. Below we study the impact of dynamic attacks on blind users.

Two iterations We start with the attacks modelled in the example from Figure 1. The information leakage of the first rating R_1 with regard to O_1 is $H(O_1){-}H(O_1|R_1)$. After receiving R_1 but before receiving R_2, the user's uncertainty about O_2 is $H(O_2|R_1)$. Upon receiving R_2, the uncertainty changes to $H(O_2|R_2,R_1)$. The information leakage of the second rating is the reduction of the user's uncertainty about O_2, i.e., $H(O_2|R_1){-}H(O_2|R_2,R_1)$. The total information leakage is simply the sum of the information leakage of the different ratings:

$$H(O_1){-}H(O_1|R_1) + H(O_2|R_1){-}H(O_2|R_1,R_2) \quad (1)$$

As the observations of the attacker only depend on the integrity of the target (which is assumed of the maximum uncertainty), O_2 is independent of R_1, and $H(O_2|R_1){=}H(O_2){=}1$. Similar independency can be proved between any O_i and $\overline{R_{i-1}}$, $(\overline{O_{i-1}},\overline{R_{i-1}})$, i.e., $H(O_i|\overline{R_{i-1}}){=}H(O_i|\overline{O_{i-1}},\overline{R_{i-1}}){=}H(O_i){=}1$.

Proposition 1. *For dynamic attacks of size* 2*, the information leakage is* 0 *iff* $p{\leqslant}\frac{1}{2}$, $a{=}\frac{1-2p}{2(1-p)}$ *and* $(1{-}2p)a_1{+}a_0{=}1{-}2p$.

Proof. Formula (1) equals 0 iff $H(O_1){=}H(O_1|R_1)$ and $H(O_2){=}H(O_1|R_1,R_2)$. These happen iff O_1 is independent of R_1, and O_2 is independent of (R_1,R_2), which means $p(o_1|r_1){=}p(o_1)$ and $p(o_2|r_1,r_2){=}p(o_2)$ must hold for all o_1,o_2,r_1,r_2. The probabilities can be rewritten via a, a_0, a_1, p. Basic algebra suffices to prove the proposition. □

Proposition 1 gives strategies that induce the most harmful dynamic attacks of size 2. For $p{=}{}^1\!/_2$, we get $a{=}a_0{=}0$. This implies when there are equal number of honest advisors and attackers, the way of absolutely hiding information is to always lie. Intuitively, this is because a randomly selected advisor in the system is equally likely to tell the truth and to lie. For $p{>}{}^1\!/_2$, where honest advisors outnumber attackers, there must be information leakage.

Formula for n iterations We generalize formula (1) to dynamic attacks of size n. Similar to the example, $\sum_i H(O_i|\overline{R_{i-1}}) = n$, which we use to simplify the definition of information leakage:

$$n - \textstyle\sum_i H(O_i|\overline{R_i}) \quad (2)$$

The conditional entropies can be represented in terms of p, n and the strategies of the attacker (i.e. the collection $a_{\ldots}$, using the shorthand α):

Theorem 1. *The information leakage of n iterations is:*

$$n + \sum_{1\leqslant i\leqslant n}\Big(\mathbf{f}((1-p)\sum_{b\in B_i^0}\alpha_b) + \mathbf{f}(p+(1-p)\sum_{b\in B_i^1}\alpha_b)\Big)$$

Proof. Unfold the conditional entropy, apply the law of total probability over the possible states, and substitute the resulting terms with α. □

Theoretical results for n iterations Blind users cannot learn an attacker's honesty from its past ratings. Hence, ratings in different iterations are independent to blind users. Information gain of a rating is not related to other ratings:

Theorem 2. *Blind users will not learn from past ratings:*

$$H(O_i){-}H(O_i|R_i) = H(O_i){-}H(O_i|\overline{R_i})$$

Proof. The main equation is equivalent to $H(O_i|R_i) = H(O_i|\overline{R_i})$. Considering $p(o_i|r_i){=}p(r_i|o_i)$, we get $p(o_i|r_i) = 2\cdot\sum_{\overline{o_{i-1}},\overline{r_{i-1}}} p(\overline{o_i},\overline{r_i})$. At the same time, $p(o_i|\overline{r_i}){=}2^i$

$\cdot \sum_{\overline{o_{i-1}}} p(\overline{o_i}, \overline{r_i})$, and $\sum_{\overline{o_{i-1}}} p(\overline{o_i}, \overline{r_i})$ remains the same for any value of $\overline{r_{i-1}}$, hence we get $p(o_i|r_i){=}p(o_i|\overline{r_i})$ hold for all $o_i, r_i, \overline{r_i}$. Furthermore, $p(r_i){=}2^{i-1}p(\overline{r_i})$, thus the equality between two conditional entropy can be derived. □

Theorem 2 implies that for a blind user, rating of a dynamic attacker can be treated as independent of the iteration in which it happens. And the total information leakage is the sum of the information leakage of individual ratings, i.e., $\sum_{0<i\leqslant n} H(O_i){-}H(O_i|R_i)$. Dynamic attacks on blind users are equivalent to repeated static attacks. The maximum harm of static attacks has already been studied in (Wang et al. 2015a). Applying the results to dynamic attacks:

Corollary 1. *Zero information leakage occurs iff* $0 \leqslant p \leqslant {}^1\!/_2$ *and* $a_b = \frac{1-2p}{2-2p}$*; whereas for* ${}^1\!/_2 \leqslant p \leqslant 1$*, information leakage is minimised when* $a_b = 0$*.*

Proof. Combining Theorem 2 with Theorems 3 and 4 in (Wang et al. 2015a) suffices to prove the corollary. □

Camouflage attacks Corollary 1 implies the most harmful attacks are not camouflage attacks, since $a_b \leq {}^1\!/_2$ – attackers lie more often than tell the truth, even initially – in fact, there is no relationship between order and probability of lying. Intuitively, accumulating trustworthiness is meaningless against blind users, that cannot distinguish truth from lies. Furthermore, Theorem 2 implies that camouflage attacks are no more harmful than the repeated strongest static attacks.

Quantitatively, the information leakage of a k-camouflage attack is $n + (n{-}k) \cdot \mathbf{f}((1-p))$. It is minimized when $k{=}0$, i.e. always lie. This means the strategy to minimise information is not to do camouflage.

What if users are not blind, but smarter? Perhaps the camouflage attacks are more harmful to smarter users, which may recognise honest behaviour, allowing attackers to accumulate trust. Next, we study users who can accurately tell whether advisors told the truth or not (i.e., *aware users*).

Attacks against Aware Users

Aware users can tell with perfect accuracy whether a previous rating was honest. For now, we ignore the issue of subjectivity. Aware users are polar opposite of blind users. In this section, we study what would be the most harmful dynamic attacks for aware users.

Two iterations Again, we start with the example in Figure 1. The information leakage of the first rating remains $H(O_1){-}H(O_1|R_1)$. Aware users know whether R_1 coincides with O_1, therefore before the second rating, the uncertainty about O_2 is $H(O_2|O_1, R_1)$. Upon receiving R_2, the uncertainty becomes $H(O_2|O_1, R_1, R_2)$. Hence, the information leakage for two ratings is:

$$H(O_1){-}H(O_1|R_1){+}H(O_2|R_1, O_1){-}H(O_2|O_1, R_1, R_2) \quad (3)$$

Proposition 2. *For dynamic attacks of size* 2*, the information leakage is* 0 *iff* $p{\leqslant}\frac{1}{4}$, $a{=}\frac{1-2p}{2(1-p)}$, $a_1{=}\frac{1-4p}{2(1-2p)}$, $a_0{=}\frac{1}{2}$.

Proof. Formula (3) equals 0 iff $H(O_1){=}H(O_1|R_1)$ and $H(O_2){=}H(O_2|O_1, R_1, R_2)$. The condition for the first equality is already proved in Proposition 1. The second equality holds iff O_2 is independent of (O_1, R_1, R_2), which means $p(o_2){=}p(o_2|o_1, r_1, r_2)$. Rewrite the equation using $a_{\dots}, p$, we can derive their values. □

Proposition 2 presents the strongest attacks that can achieve 0 information leakage for aware users. Note that $a_1{<}a{<}\frac{1}{2}$, which implies the attacker has higher chance to lie than to tell the truth in the first rating, and the chance of lying increases if the attacker just told the truth. To get more general results, we extend the attack size from 2 to n ($n{\geqslant}2$).

Formula for n iterations The information leakage for n iterations is

$$n - \sum_{1\leqslant i\leqslant n} H(O_i|\overline{R_i}, \overline{O_{i-1}}) \quad (4)$$

Based on the chain rule of conditional entropy, and the conditional independency between O_i and R_j, $(j > i)$ given $\overline{O_{i-1}}, \overline{R_i}$, we can prove the information leakage formula above is equal to

$$H(\overline{O_n}) - H(\overline{O_n}|\overline{R_n}) \quad (5)$$

As before, the conditional entropy can be represented in terms of p, n and the strategy parameters $a_{\dots}$:

Theorem 3. *The information leakage of* n *iterations is:*

$$n + \mathbf{f}(p + (1-p)\alpha_{\hat{b}_n}) + \sum_{b\in B_n, b\neq\hat{b}_n} \mathbf{f}((1-p)\alpha_b)$$

Proof. Modify the proof of Theorem 1 by adding a case distinction for the left-most branch. □

Theoretical results for n iterations The attacker can render ratings useless, only if $p \leqslant {}^1\!/_{2^n}$.

Theorem 4. *For aware users and dynamic attacks of size* n*, attacks with* 0 *information leakage occur when* $0{\leqslant}p{\leqslant}{}^1\!/_{2^n}$*,* $a_{\hat{b}_i} = \frac{1}{2} - \frac{p}{2(1-p)\prod_{j=0}^{j=i-1} a_{\hat{b}_i}}$*, and* $a_b = {}^1\!/_2$ *for* $b \neq \hat{b}_i$*.*

Proof. Considering $H(\overline{O_n}){=}n$, to minimise formula (5) we only need to maximise the subtractor. $H(\overline{O_n}|\overline{R_n}){=}\sum_{\overline{r_n}} p(\overline{r_n}) \cdot H(\overline{O_n}|\overline{r_n})$, and $p(\overline{r_n})$ constantly equal to $\frac{1}{2^n}$. Based on the Jensen's inequality (Jensen 1906), the maximum of the conditional entropy happens when $p(\overline{O_n}|\overline{r_n})$ are equal for any value of $\overline{O_n}$ and $\overline{r_n}$, and the maximum value is n. The probabilities can be rewritten using all transition probabilities, and basic algebra suffices to prove the theorem. □

When $p > {}^1\!/_{2^n}$, the strategy that causes the most harm is independent of p:

Theorem 5. *For aware users, the most harmful attack of size* n*, given* $p > {}^1\!/_{2^n}$*, occurs when* $a_{\hat{b}_i} = \frac{2^{n-i-1}-1}{2^{n-i}-1}$ *and* $a_b = {}^1\!/_2$ *for* $b{\neq}\hat{b}_i$*.*

Proof. Take the formula from Theorem 3. Since $p{>}1/2^n$, the minimum has $\alpha_{\hat{b}_n} = 0$. Using Jensen's inequality, it suffices to set the remaining α_{b_n} equal. Then, α_{b_n} must equal $\frac{1}{2^n-1}$, which happens when all a_b are set as in the theorem. □

Note that for blind users, 0 information leakage can be achieved for any $p{\leqslant}\frac{1}{2}$. But for aware users, the required value ranges of p are much smaller and get narrower as rating iterations increase ($p{\leqslant}1/2^n$ for n interations). This is in line with our intuition, that when users get smarter, it should be more difficult for attackers to hide information.

Camouflage attacks Theorem 4 presents the most harmful attacks for aware users. Note that $a_{\hat{b}_i}{\geqslant}a_{\hat{b}_j}$ for $i{<}j$, meaning the probability of continuing to tell the truth is non-increasing over time. This is also the case in camouflage attacks. However, all $a_{\hat{b}_i}, i = 0, \ldots, (n{-}1)$ are below $1/2$, meaning lying is always more probable. Hence, although camouflage attacks are not the most harmful attacks for aware users, but pretending to be honest with some decreasing probability is more harmful than a fixed probability.

Quantitatively, information leakage for k-camouflage is $n + \mathbf{f}(1-p) + \mathbf{f}(p)$, for $k \neq n$, and n, for $k = n$. Comparing to the blind users, camouflage attacks are less harmful. Moreover, provided the attacker lies at some point, it does not matter when he switches. Therefore, always lying is equally harmful as camouflage attacks.

Attacks against General Users

Blind users and aware users are two extreme examples of users. In this section, we study the impact of dynamic attacks on users in between of the extremes. We introduce random variables Q_i to represent a user's hindsight perception of attackers' honesty, with $Q_i{=}1$ ($Q_i{=}0$) denoting the attacker probably told the truth (lied) in the i^{th} rating. The accuracy of the user's hindsight depends on how much he can learn from his own interactions with the target, or from other sources in the system. Subjective ratings are an example with low accuracy, since even when the user forms an opinion different of the rating, it remains probable that the advisor was not lying.

We use $q, (0{\leqslant}q{\leqslant}1)$ to describe the accuracy of the user's hindsight. With probability q, the user's perception is correct: $p(Q_i{=}1|O_i{=}R_i) = p(Q_i{=}0|O_i{\neq}R_i) = q$, and incorrect with probability $1-q$: $p(Q_i{=}0|O_i{=}R_i) = p(Q_i{=}1|O_i{\neq}R_i) = (1-q)$. For high degree of subjectivity, we expect q to be close to $1/2$.

Two iterations Q_i expresses the new knowledge of the user about the attacker after the i^{th} rating, the amount of which is decided by q. To show how this influences the definition of information leakage, consider again the example in Figure 1. The information leakage of the first iteration remains unchanged. After the first iteration, the user knows R_1 and Q_1, and the uncertainty about O_2 is $H(O_2|R_1, Q_1)$. Upon receiving R_2, the uncertainty about O_2 changes to $H(O_2|Q_1, R_1, R_2)$. Hence, the information leakage of the second iteration is $H(O_2|R_1, Q_1){-}H(O_2|Q_1, R_1, R_2)$. The total information leakage of the attack is the sum of the first and the second iterations:

$$H(O_1){-}H(O_1|R_1){+}H(O_2|R_1, Q_1){-}H(O_2|Q_1, R_1, R_2) \quad (6)$$

Recall that O_2 only depends on the target, $H(O_2|R_1, Q_1){=}H(O_2){=}1$. We first study the maximum impact of attacks of size 2:

Proposition 3. *For dynamic attacks of size 2, the information leakage is 0 for general users iff: (1) for $q{\neq}\frac{1}{2}$, $p{\leqslant}\frac{1}{4}$, $a{=}\frac{1-2p}{2(1-p)}$, $a_1{=}\frac{1-4p}{2(1-2p)}$, $a_0{=}\frac{1}{2}$. (2) for $q{=}\frac{1}{2}$, refer to strategies in Proposition 1.*

Proof. Formula (6) is 0 iff $H(O_1){=}H(O_1|R_1)$ and $H(O_2)$ $=H(O_2|Q_1, R_1, R_2)$. We have proved in Proposition 1 that to get the first equality, $a{=}\frac{1-2p}{2(1-p)}$. The second equality holds iff O_2 is independent of (Q_1, R_1, R_2), which means $p(o_2|Q_1, r_1, r_2){=}p(o_2){=}1/2$. Rewrite the probabilities using transition probabilities, to obtain a_0, a_1. □

Note that for any $q{\neq}\frac{1}{2}$ – i.e., a user has at least some accuracy in hindsight – the strategy to completely hide information remains the same. Thus, as long as a user is not blind, a single strategy suffices to completely hide information, regardless of the user's accuracy.

Formula for n iterations Generalising the attack to size n, the information leakage of an entire attack is:

$$n - \textstyle\sum_i H(O_i|\overline{R_i}, \overline{Q_{i-1}}) \quad (7)$$

Rewrite in terms of p, q, n and $a_{\ldots}$:

Theorem 6. *Let $\beta_{b,c} = \prod_{1 \leqslant j < i} \pi(1{-}q, b(j) \oplus c(j))$. For $x \in \{0, 1\}$, let $\gamma_c^x = (1{-}p) \sum_{b \in B_i^x} \alpha_b \beta_{b,c}$. Let $\delta_{c,i} = \gamma_c^0 + \gamma_c^1 + p\beta_{\hat{b}_i,c}$. The information leakage is:*

$$n + \sum_{1 \leqslant i \leqslant n} \sum_{c \in B_i} \delta_{c,i} \cdot \left[\mathbf{f}\left(\frac{\gamma_c^1 + p\beta_{\hat{b}_i,c}}{\delta_{c,i}}\right) + \mathbf{f}\left(\frac{\gamma_c^0}{\delta_{c,i}}\right) \right]$$

Proof. In addition to the technique used in Theorem 1, apply the law of total probability over all Q_i. □

Theoretical results for n iterations The blind and aware users are special cases of the general users:

Theorem 7. *Dynamic attacks on blind users (aware users) are a special case of attacks on general users, where $q = 1/2$ ($q = 1$). Specifically:*

$$H(O_i|\overline{R_i}) = H(O_i|\overline{R_i}, \overline{Q_{i-1}}), \qquad q{=}1/2 \quad (8)$$

$$H(O_i|\overline{R_i}, \overline{O_{i-1}}) = H(O_i|\overline{R_i}, \overline{Q_{i-1}}), \qquad q{=}1 \quad (9)$$

Proof. When $q = 1/2$, $\pi(1-q, c(j) \oplus b(j))$ is a constant $\frac{1}{2}$ for all j, making $\beta_{b,c}$ is a constant, and simplifying γ to fit Theorem 1. When $q = 1$, $\beta_{b,c}$ contains a 0-factor, whenever $b \neq c$, meaning $\beta_{b,c} = 1$ iff $b = c$. This implies $\gamma_c^x = \alpha_c$, and the term $p\beta_{\hat{b}_i,c}$ equals p iff $c = \hat{b}_i$. Some formula manipulation shows it fits Theorem 3. □

Theorem 8. *For dynamic attacks of size n on general users, 0 information leakage can be achieved when for all $1 \leqslant i \leqslant n$, $c \in B_i$: $\gamma_c^1 + p\beta_{\hat{b}_i,c} = \gamma_c^0$.*

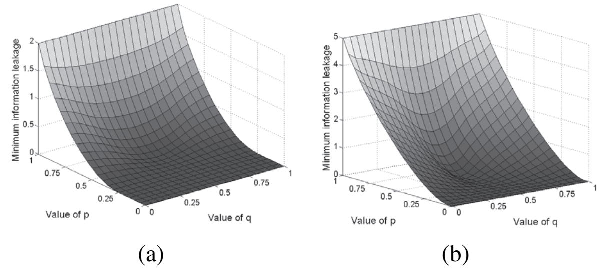

Figure 2: The information leakage in the most harmful strategies for two (a) or five (b) iterations.

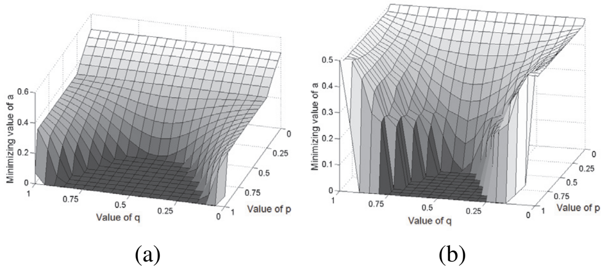

Figure 3: The a value in the most harmful strategies for two (a) or five (b) iterations.

Proof. To get information leakage be 0, refer to formula (7), $H(O_i) = H(O_i|\overline{R_i}, \overline{Q_{i-1}})$ must hold for all i. The equalities are achieved iff O_i is independent of $\overline{R_i}, \overline{Q_{i-1}}$, which means $p(o_i|\overline{r_i}, \overline{q_{i-1}}){=}p(o_i){=}{}^1\!/\!{}_2$. This implies $p(o_i{=}r_i, \overline{r_i}, \overline{q_{i-1}}){=}p(o_i{\neq}r_i, \overline{r_i}, \overline{q_{i-1}})$, which instantiate $\gamma_c^1 + p\beta_{\hat{b}_i,c}$ and γ_c^0. □

Numerical results for n iterations Using numerical approximation, we found strategies that cause close to maximal harm, given specific p, q and n. Figure 2 plots the information leakage w.r.t. p and q, for n=2 in Figure 2a, and n=5 in Figure 2b. Similarly, Figure 3 plots the initial probability of telling the truth (a) w.r.t. p and q, for n=2 and n=5.

Observe that the graphs are symmetrical over $q = {}^1\!/\!{}_2$, since when $q < {}^1\!/\!{}_2$ we can simply swap the meaning of $Q = 0$ and $Q = 1$. Moreover, note that a is bound by ${}^1\!/\!{}_2$ and information leakage by n.

In Figure 2, note that if $p = 1$, all information is leaked – since all advisors are honest. Secondly, for two iterations, the information leakage is 0 when $0 \leqslant p \leqslant {}^1\!/\!{}_4$, and for five, when $0 \leqslant p \leqslant {}^1\!/\!{}_{32}$. This pattern holds for all q, except $q = {}^1\!/\!{}_2$, in which case information leakage is 0 when $0 \leqslant p \leqslant {}^1\!/\!{}_2$. We find that (barring $q = {}^1\!/\!{}_2$) obtaining 0 information leakage requires exponentially more attackers relative to honest users, as the number of iterations increases. This extends the theoretical result in Theorem 4, where we prove this pattern for aware users ($q = 1$).

Regarding Figure 3, first observe that there appears to be a phase transition at $p = {}^1\!/\!{}_2$ and $p = {}^1\!/\!{}_{32}$, in Figures 3a and Figures 3b, respectively. The choice of optimal a is independent of q, for smaller p. Since the same strategy is optimal for blind users and aware users, when there are many attackers, and there is no point pretending to be honest to blind users, there is no point pretending to be honest to any users (for small p).

Second, observe that for larger p and q close to ${}^1\!/\!{}_2$, $a = 0$. Thus, the attacker initially always lies (Corollary 1 proves this for $q = {}^1\!/\!{}_2$). As n increases, the area where the attacker always lies appears to shrink. The strategy of always lying is less often the most harmful, when the number of ratings increases.

Finally, for larger p and q closer to 1 (or 0), $a > 0$. Thus, the attacker sometimes tells the truth, despite the fact that for static scenarios it should always lie. This indicates that pretending to be honest initially (like probabilistic camouflaging) is (the most) harmful. The probability of telling the truth initially should, however, not be too high – it appears to be bounded by ${}^1\!/\!{}_2$.

Camouflage attacks Generally, the analysis of camouflage attacks is the same as for aware users, with the major exception of an area around $q = {}^1\!/\!{}_2$. The area in the graphs in Figure 3 where $a = 0$ denotes combinations of p, q and n where pretending to be honest definitely causes no harm. We see that the area increases in size as p grows, but decreases in size as n grows. Furthermore, as n and p grow, the cutoff (between the cases where pretending to be honest causes harm and cases where it does not) becomes sharper.

Quantitatively, letting $\dot{b}_i$ be k 1's followed by $i{-}k$ 0's, the information leakage of k-camouflage is $n{+}\sum_{k<i\leq n} \sum_{c\in B_i} \big(\mathbf{f}(p\beta_{\hat{b}_i,c}){+}\mathbf{f}((1{-}p)\beta_{\dot{b}_i,c}){-}\mathbf{f}(p\beta_{\hat{b}_i,c}{+}(1{-}p)\beta_{\dot{b}_i,c})\big)$. Again, this is minimised when $k = 0$, meaning that even always lying is worse than camouflage.

Conclusion

In the camouflage attack, users initially pretend to be honest, to increase the impact of later dishonest ratings. The motivating question for this paper, is whether such behaviour is harmful. It turns out that the answer to that question depends on the hindsight of the users. If users have no hindsight (blind users), then attackers pretending to be honest can only be beneficial to them. If users have perfect or limited hindsight (aware/general users), then it may cause harm when attackers initially pretend to be honest with a (small) probability. It never causes harm to (initially) pretend to be honest with probability over ${}^1\!/\!{}_2$, let alone probability 1. Therefore, the camouflage attack, where attackers are always honest until a certain point, is not very harmful.

The results are obtained by using a method to measure the strength of unfair rating attacks applied in the literature (Wang et al. 2015a; 2015b). The method is not suitable for multiple ratings, so we had to generalise the methodology. Moreover, we needed to construct a formal model of all dynamic rating behaviour. We let the advisor be a stochastic process that generates the ratings according to some strategy. We are able to derive explicit formulas expressing the harm of any dynamic attack against any of the three users.

The theory not only allowed us to answer the main question, but also to prove interesting properties about the harm of dynamic attacks. For example, attackers can render ratings completely useless to blind users, whenever they are in the majority. But for all other users the attackers need to greatly outnumber the honest users (exponential in the number of ratings) to render ratings useless. Another interesting result is that against aware users, the most harmful attacks do not depend on how many attackers there are (unless they greatly outnumber the honest users).

Now we know that it is typically not very harmful if users merely pretend to be honest. An interesting future direction would be to exploit the information present in ratings made by those pretending to be honest. Moreover, we have characterised different kinds of attacks and computed the most harmful ones. Therefore, it may be possible to strengthen a trust system specifically against the most harmful dynamic attacks.

Acknowledgments

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.

References

Bao, F.; Chen, I.-R.; Chang, M.; and Cho, J.-H. 2012. Hierarchical trust management for wireless sensor networks and its applications to trust-based routing and intrusion detection. *IEEE Transactions on Network and Service Management* 9(2):169–183.

Cover, T. M., and Thomas, J. A. 1991. Entropy, relative entropy and mutual information. *Elements of Information Theory* 12–49.

Glynos, D.; Argyroudis, P.; Douligeris, C.; and Mahony, D. O. 2008. Twohop: metric-based trust evaluation for peer-to-peer collaboration environments. In *Global Telecommunications Conference*, 1–6. IEEE.

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

Jensen, J. L. W. V. 1906. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica* 30(1):175–193.

Jiang, S.; Zhang, J.; and Ong, Y.-S. 2013. An evolutionary model for constructing robust trust networks. In *Proceedings of the 12th international conference on Autonomous agents and multi-agent systems*, 813–820. IFAAMAS.

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

Jøsang, A. 2010. Robustness of trust and reputation systems. In *Proceedings of the 4th International Conference on Self-Adaptive and Self-Organizing Systems Workshop*, 159–159. IEEE.

Kamvar, S. D.; Schlosser, M. T.; and Garcia-Molina, H. 2003. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, 640–651. ACM.

Kerr, R., and Cohen, R. 2009. Smart cheaters do prosper: defeating trust and reputation systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, 993–1000. IFAAMAS.

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

Oram, A. 2001. *Peer-to-peer: harnessing the benefits of a disruptive technology*. O'Reilly Media, Inc.

Regan, K.; Poupart, P.; and Cohen, R. 2006. Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In *Proceedings of the 21st National Conference on Artificial Intelligence*, volume 21, 1206. AAAI.

Şensoy, M.; Zhang, J.; Yolum, P.; and Cohen, R. 2009. Poyraz: Context-aware service selection under deception. *Computational Intelligence* 25(4):335–366.

Shaikh, R. A.; Jameel, H.; d'Auriol, B. J.; Lee, H.; Lee, S.; and Song, Y.-J. 2009. Group-based trust management scheme for clustered wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems* 20(11):1698–1712.

Sun, Y. L.; Han, Z.; Yu, W.; and Liu, K. R. 2006. A trust evaluation framework in distributed networks: Vulnerability analysis and defense against attacks. In *Proceedings of the 25th International Conference on Computer Communications*, 1–13. IEEE.

Vogiatzis, G.; MacGillivray, I.; and Chli, M. 2010. A probabilistic model for trust and reputation. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, 225–232. IFAAMAS.

Wang, D.; Muller, T.; Irissappane, A. A.; Zhang, J.; and Liu, Y. 2015a. Using information theory to improve the robustness of trust systems. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems*, 791–799. IFAAMAS.

Wang, D.; Muller, T.; Zhang, J.; and Liu, Y. 2015b. Quantifying robustness of trust systems against collusive unfair rating attacks using information theorys. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, 111–117.

Xiong, L., and Liu, L. 2004. Peertrust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE Transactions on Knowledge and Data Engineering* 16(7):843–857.

Yu, B., and Singh, M. P. 2003. Detecting deception in reputation management. In *Proceedings of 4th International Autonomous Agents and Multi Agent Systems*, 73–80. IFAAMAS.

Yu, Y.; Li, K.; Zhou, W.; and Li, P. 2012. Trust mechanisms in wireless sensor networks: Attack analysis and countermeasures. *Journal of Network and Computer Applications* 35(3):867–880.