How to Use Information Theory to Mitigate Unfair Rating Attacks

Tim Muller, Dongxia Wang, Yang Liu, Jie Zhang

How to Use Information Theory to Mitigate Unfair Rating Attacks

Tim Muller, Dongxia Wang, Yang Liu, and Jie Zhang

Nanyang Technological University

Abstract. In rating systems, users want to construct accurate opinions based on ratings. However, the accuracy is bounded by the amount of information transmitted (leaked) by ratings. Rating systems are susceptible to unfair rating attacks. These attacks may decrease the amount of leaked information, by introducing noise. A robust trust system attempts to mitigate the effects of these attacks on the information leakage. Defenders cannot influence the actual ratings: being honest or from attackers. There are other ways for the defenders to keep the information leakage high: blocking/selecting the right advisors, observing transactions and offering more choices. Blocking suspicious advisors can only decrease robustness. If only a limited number of ratings can be used, however, then less suspicious advisors are better, and in case of a tie, newer advisors are better. Observing transactions increases robustness. Offering more choices may increase robustness.

1 Introduction

Online systems nowadays are typically too large for a single user to oversee. A user must rely on recommendations, reviews, feedback or *ratings* from other users (i.e., advisors), to be able to use a system to its fullest extent. In practice, we see that ratings are ubiquitous in large online systems. The exact design challenges introduced by supporting ratings depend on context (e.g. rating format, distributing ratings, subjectivity). One major challenge for all systems is how to deal with unfair ratings.

Typical approaches perform some or all of the following: incentivise honest ratings, detect and filter unfair ratings, update advisors' trustworthiness. More involved approaches may attempt to use possibly unfair ratings, and correct for the possible error, e.g. using machine learning or statistical methods. We call such methods aggregation mechanisms. The power of aggregation mechanisms is limited. Specifically, given a set of ratings, the amount of information that can be extracted is bounded upwards by a certain quantity. We call this quantity the *information leakage* of the ratings. No aggregation mechanism can be expected to do better than that.

Fortunately, the set of ratings that an aggregation mechanism operates on is not a universal given. One may control factors such as the number of advisors, which advisors to ask and the rating format. Changing these factors will change the information leakage of the ratings, and thus the limits of the power of aggregation mechanisms. Ideally, we want to increase the limit.

We formalise an abstract model of rating systems that assumes the bare minimum. Its only assumption is that there exist honest advisors, and their ratings correlate somehow with the decision that a user should make. Furthermore, we show that information theory offers effective tools to measure the quality of ratings. Specifically, we prove that information leakage of a rating about a decision puts a hard bound on the accuracy of a decision. The remaining results are a set of design guidelines. These guidelines are valid for any rating system that has ratings that somehow correlate with good decisions. Specifically, 1) blocking suspicious advisors is not helpful and can decrease robustness, 2) when receiving ratings is costly, less suspicious advisors should be preferred, 3) and if advisors are equally suspicious, newer ones are preferable, 4) if possible, keep track of who has direct experience, and 5) changing the rating format and the options in a decision may increase robustness.

The paper is organised as follows: we discuss related idea and approaches in Section 2. We discuss the problem at hand – unfair rating attacks, in Section 3. Then we introduce the abstract notion of trust systems, in Section 4, formalise them (Section 4.1) and discuss what defenders can alter (Section 4.2). Then we show that limiting the information leakage is limiting the accuracy, in Section 5. In Section 6, we prove the five aforementioned guidelines.

2 Related Work

Multiple types of approaches exist to deal with unfair rating attacks. Some approaches provide incentives to promote honest rating behaviour [**?**,**?**,**?**]. Jurca and Faltings design a payment-based incentive scheme, which explicitly rewards honest feedback by an amount that offsets both the cost and the benefit of lying [**?**]. The payment schemes can be based on proper scoring rules, or correlation between the ratings of different advisors. In [**?**], they study how to resist against collusive advisors: colluders that share a lying strategy have to suffer monetary losses. Some other approaches aim to detect and filter out unfair ratings [**?**,**?**]. For product-rating based online rating systems, Yafei et al. propose to detect collaborative biased ratings by observing time intervals where they are highly likely [**?**]. Reporting ratings is treated as a random process, and signal-processing techniques are applied to detect changes in rating values (e.g., detecting mean change). Most approaches evaluate advisors' trustworthiness, based on which reliable advisors are selected or ratings get discounted [**?**,**?**]. Yu et al., propose a reinforcement learning based framework to filter unfair ratings and make more accurate decisions in selecting trustees [**?**]. Both direct experiences and indirect evidences from advisors are aggregated to select highly reputable trustees. The reward derived from the interaction with a trustee is used to update advisors' credibility, and ratings from less credible advisors are discarded. Meanwhile, weights assigned to direct and indirect trust evidences are also updated in trust evidence aggregation. We call these defense approaches as aggregation mechanisms.

The classification for different aggregation mechanisms is not absolute. Different types of approaches may be aggregated. For example, in [?], statistical methods are used to detect unfair ratings, of which the results are used to evaluate trustworthiness of advisors. The trustworthiness of advisors is then used to aggregate ratings, and also detect future suspicious ratings.

Despite of deviating from the truth, unfair ratings may still contain useful information (e.g., if they are correlated with a user's direct experiences). There are approaches which exploit such correlation to make use of unfair ratings [?,?,?]. BLADE [?] and HABIT [?] learn from statistical correlations between a user's direct experiences and an advisor' ratings to adjust his ratings. For example, if an advisor always report bad ratings about a trustee, of which the user has good trust opinion, then his ratings get reversed. In this paper, we proved that suspicious advisors may still provide useful ratings (Proposition 4). Ratings from honest advisors may be subjectively different from a user's direct experiences, but they differentiate from unfair ratings from attackers. Subjective ratings may provide useful information as they are relevant for a user. By directly discarding or filtering ratings that deviate from direct experiences, subjective ratings from honest advisors may also get excluded.

The quantification of the amount of information in ratings (i.e., *information leakage*) is already well studied in [?]. The defense approaches above cannot change the information leakage of ratings in a system, and they only differ in the way of exploiting it. Hence, their effectiveness is limited. From [?], we know that different attacks make information leakage in a system different. In the worst-case attacks where there is little information leakage, these approaches may not help at all. A robust rating system should not let its limitation to be controlled by attacks. Hence, it is vital to increase the limit of information leakage under attacks.

We found that some properties of a system, like the format of ratings, can affect the information leakage. Also, the conditions to achieve the minimal information leakage may also change based on these properties. By proper design, the power of defense approaches can be limited less, and the power of attacks can be decreased.

3 Unfair Rating Attacks

Unfair rating attacks are known to exist. They have been detected on existing trust systems [?,?], and they are well-studied in the literature [?]. It seems straightforward what an unfair rating attack is (unfair ratings are provided to mislead a user). But in reality, only 'rating' is unambiguous. For example, subjectivity may blur the truth and lies, meaning ratings deviating from the truth may not be from attackers, but subjective honest advisors. Moreover, with some probability, an honest user may perform the same sequence of actions (trace) as a user that intends to attack the system [?]; is that trace an attack? The issues lies in considering only the actual ratings.

Ratings cannot be fair or unfair by themselves. Not only may subjectivity lead to false ratings that are not unfair, but unfair ratings can be (objectively or subjectively) true. Advisors may tell the truth to mislead users that believe the advisor is more likely to lie [?]. Advisors may tell the truth because they are colluding, but want to remain undetected [?]. Or advisors may tell the truth because they do not want to lose the user's trust [?]. In each case, the unfairness lays in the fact that the advisor merely acts out some malicious strategy, rather than respecting the truth.

We want to have a pragmatic definition of unfair rating attacks. Our goal is to make rating systems robust against unfair rating attacks. In other words, ratings must be useful, even if some sources are malicious. However, how useful ratings are to a user, depends on what the user chooses to do with these ratings. The aim of this paper is not to provide the right aggregation mechanism or dictate user's decisions, so – pragmatically – we take a measure of how much a user can do with the ratings: information leakage. We prove, in Theorem 1, that the information leakage measures the potential usefulness of ratings.

Attackers have an underlaying strategy. Attacks are considered successful, when they achieve some goal. The goal is not known in advance. Since we are considering robustness, we primarily care about the worst-case for a user – the information leakage is minimal. Hence, we pragmatically assert that the goal of an attack is to minimise information leakage. We assume attackers select the strategy that minimises information leakage. If we are wrong, then the information leakage increases by definition. Section 5.1 provides detailed formal analysis.

4 Rating System

In this paper, we are not necessarily interested in rating systems themselves, but rather in which aspects we can control in our advantage. Particularly, we study how to set the parameters that we control to harden a rating system – maximising the minimal information leakage (Section 3). In this section, we present an abstract representation of a rating system, that allows us to analyse the relevant aspects without dissolving in details.

Users make decisions based on ratings. Some decisions are better than others. We take a simple notion of correctness of decisions. Users are given n choices to make a decision, of which 1 choice is the best option. The relevant part is that we have a ground truth, that the user wants to deduce.

We exemplify the theory with a simple example. The results that we present in this paper are general results for all rating systems that follow the notions from this section. However, the results are easier to interpret on a simple example (e-commerce) system:

Example 1. On an e-commerce system, two similar products are offered for sale. The user has three options: buy product x, buy product y, or buy neither. In the e-commerce system, another buyer can fulfill the role of advisor, and assign scores of $1 - 5$ stars to x, y, neither or both. Each of these (combinations of) scores may imply something about the user's decision, hence the abstract rating must contain each. Thus, there are 1 (neither) plus 5 (just x) plus 5 (just y) plus $5 \cdot 5$ (both), which is 36, possible ratings. An honest advisor provides a rating that – at the very least – correlates with the correct decision. Here, if buying x is the best option for the user, then an honest advisor is more likely to assign 5 stars than 1 star to x. Some participants may have a hidden agenda, for example, to boost sales of product y. These participants provide ratings strategically, and we call them attackers.

The example shows that the actual details of the decision are not very important. The relationship between the abstract honest ratings and the decision is crucial for users. In this paper, we assert an arbitrary non-independent relationship between honest ratings and the decision. We do not concretely model this relationship (contrary to e.g. [**?**]).

4.1 Formal Model

A user makes a decision by picking one option from the ratings. And he tries to select the best option. We simply model the decision as a random variable, with its (unknown) outcome representing the (unknown) best option. The outcomes of a *decision* consists of a set of *options* $\mathcal{O} = \{0, \dots, n-1\}$). We use the random variable Θ over the options $\mathcal{O}$ to denote the best option. Thus, $P(\Theta = \theta | \phi)$ is the probability that θ is the best option, when ϕ is given.

A rating has a certain format, it could be a number of stars (i.e. discrete and ordered), a list of tags (i.e. discrete and not ordered) or a real value in some range. On an abstract level, the structure is actually not that relevant. Specifically, it is only relevant when constructing an aggregation mechanism – which is not the purpose of this paper. We consider a *rating format* to be a set of scores $\mathcal{R}$, and a *rating* to be a random variable R, which has the property that it says something about Θ when the advisor is honest. To accommodate for multiple advisors giving ratings, let $\mathcal{A} = 0, \dots, m-1$ be the set of advisors, and let R_j be the rating provided by $j \in \mathcal{A}$. We use $\mathbf{R}_A$ to mean $R_{a_0}, \dots, R_{a_k}$ for $\{a_0, \dots, a_k\} = A \subseteq \mathcal{A}$.

Advisors can be honest or malicious. We introduce the status of advisor a, which is honest ($\top$) or malicious ($\bot$), as a random variable S_a. An honest advisor would typically not give the same rating as a malicious advisor. We introduce $\widehat{R}$ and $\widetilde{R}$, both over $\mathcal{R}$, to represent the rating an honest or malicious advisor would give. Thus, $R_a = \widehat{R}_a$ whenever $S_a = \top$, and $R_a = \widetilde{R}_a$ whenever $S_a = \bot$. We shorthand the prior probability that a is honest as $P(S_a) = s_a$. As we reason about a trust system, we assert $0 < s_a < 1$.

In the running example, we mentioned that the honest advisors' ratings say something about the best option (decision). We distil that notion by saying Θ is not independent from honest advisors' ratings $A \subseteq \mathcal{A}$: $P(\Theta) \neq P(\Theta | \widehat{\mathbf{R}}_A)$. Another way to phrase this, is to say that honest advisors' ratings *leak* information about the decision. We need to use information theory to encode this notion [**?**]:

Definition 1. *Let X, Y, Z be discrete random variables.*
The surprisal *of an outcome x of X is $-\log(P(x))$.*
The entropy *of X is*

$$H(X) = \mathbb{E}_X(-\log(P(x)) = \sum_i P(x_i) \cdot -\log(P(x_i))$$

The conditional entropy *of X given Y is*

$$H(X|Y) = \mathbb{E}_X(-\log(P(x|y))) = \sum_{i,j} P(x_i, y_j) \cdot -\log(P(x_i|y_j))$$

The mutual information *of X and Y is*

$$I(X;Y) = \mathbb{E}_{X,Y}(\log(\frac{P(x,y)}{P(x)P(y)})) = \sum_{i,j} P(x,y)\log(\frac{P(x,y)}{P(x)P(y)})$$

The conditional mutual information *of X and Y given Z is*

$$I(X;Y|Z) = \mathbb{E}_Z(I(X;Y)|Z) = \sum_{i,j,k} P(x,y,z)\log(\frac{P(x,y|z)}{P(x|z)P(y|z)})$$

Information leakage of Y about X (given Z) is the (conditional) mutual information of X and Y (given Z). Information leakage is the difference in the information about X when Y is given and not given: $I(X;Y) = H(X) - H(X|Y)$, or $I(X;Y|Z) = H(X|Z) - H(X|Y,Z)$ [**?**]. Information leakage is non-negative.

Information theory allows us to rewrite the link between honest ratings and correct options as conditional information leakage:

$$\begin{aligned} & I(\Theta; \mathbf{R}_A | \mathbf{S}_A{=}\top) \\ =& H(\Theta|\mathbf{S}_A{=}\top) - H(\Theta|\mathbf{R}_A, \mathbf{S}_A{=}\top) \\ =& \sum_\theta p(\theta|\mathbf{S}_A = \top) - \sum_{\mathbf{r}_A} p(\mathbf{r}_A) \sum_\theta p(\theta|\mathbf{r}_A, \mathbf{S}_A = \top) \end{aligned}$$

We assume honest ratings are always correlated with correct options, hence there is always conditional information leakage: $I(\Theta; \mathbf{R}_A | \mathbf{S}_A{=}\top) > 0$.

Now, the dishonest advisors select R_j such that $I(\Theta; R_{\mathcal{A}}|\Phi)$ is minimised for given Φ (such as $\Phi = \mathbf{S}_A$). The term, $I(\Theta; R_{\mathcal{A}}|\Phi)$, quantifies how much the ratings say about the optimal option.

$$\begin{aligned} & I(\Theta; \mathbf{R}_{\mathcal{A}}|\Phi) \\ =& H(\Theta|\Phi) - H(\Theta|\mathbf{R}_{\mathcal{A}}, \Phi) \\ =& \sum_\theta p(\theta|\phi) - \sum_{\mathbf{r}_{\mathcal{A}},\phi} p(\mathbf{r}_{\mathcal{A}}, \phi) \sum_\theta p(\theta|\mathbf{r}_{\mathcal{A}}, \phi) \end{aligned}$$

We can use Example 1 to showcase parts of the formalisation:

Example 2. In Example 1, option "buy x" becomes 0, "buy y" 1 and "nothing" 2. The decision outcomes are $\{0, 1, 2\}$. We have 4 advisors, $\{0, 1, 2, 3\}$, and $0, 1$ are suspicious with $P(S_0) = 0.2, P(S_1 = 0.3)$, and $2, 3$ are not with $P(S_2) = 0.8, P(S_3) = 0.9$. Ratings can be any of $\{(r, s)|r, s \in \{\emptyset, 1, 2, 3, 4, 5\}\}$.

4.2 Controlled Parameters

In the case of a centralised system, the designer himself can make these decisions. For decentralised systems, it may be users themselves to make decisions. In the latter case, the designer of the system should try to encourage users to make the right decisions. Either way, it is important to have theoretically rigourous guidelines to make robust decisions. Here, we look at the parameters of a system that can or cannot be controlled.

In this paper, we take the viewpoint of any party that wants to increase the robustness of the system. We refer to the parties that want to increase the minimal information leakage as the *defender*. For example, when we say "under the defender's control", we mean that the user, the designer or any party that strives for robustness controls it.

The set of advisors $\mathcal{A}$ is not under the defender's control. Moreover, for any advisor $a \in \mathcal{A}$, the random variables S_a, $\widehat{R}_a$, $\widetilde{R}_a$ and R_a cannot be controlled. However, the defender can blacklist/whitelist a subset of the advisors. Formally, the defender can choose $A \subseteq \mathcal{A}$ in $I(\Theta; \mathbf{R}_A)$. Moreover, in some systems, the defender can monitor which advisors potentially have information (e.g. which advisors have performed relevant transactions). If random variable K_a captures this fact for advisor a, then the defender may choose to have $\mathbf{K}_A$ as a condition: $I(\Theta; \mathbf{R}_A | \mathbf{K}_A)$. Finally, the advisor may be able to change the actual decision, thus changing (the size of) the random variable Θ.

5 Limited Information Implies Inaccuracy

For the conclusions in this paper to hold relevance, we need to show that limited information leakage leads to limited accuracy. We do so constructively. In other words, we construct the best possible opinions that can result from given information leakage and some aggregation mechanism, and show that the accuracy of these opinions is limited by the amount of information leakage.

An opinion is an assignment of probability to each of the options in a decision. An opinion is said to be accurate, when the probability assigned to the right option is high. One way of measuring this is to take the cross entropy:

Definition 2. *For discrete random variables X, Y with the same support, the cross entropy is*

$$H_{cross}(X, Y) = \mathbb{E}_{x_i}(\log(P(y_i)) = -\sum_i P(x_i) \log(P(y_i))$$

The Kullback-Leibler divergence is

$$D_{KL}(X||Y) = H_{cross}(X, Y) - H(X) = \sum_i P(x_i) \log \frac{P(x_i)}{P(y_i)}$$

The cross entropy (and Kullback-Leibler divergence) is a standard tool to measure the quality of an approximation Y of the true distribution X. Specifically, the cross entropy takes the expectation of the surprisal one has under the approximation Y. An advantage of Kullback-Leibler divergence is that the term $-H(X)$ translates the values, such that 0 divergence occurs when $X \sim Y$. Moreover, Kullback-Leibler divergence must be non-negative.

We use cross-entropy to measure accuracy. Let $O : \mathcal{O} \to [0,1]$ such that $\sum_{\theta \in \mathcal{O}} O(\theta) = 1$ be an opinion. The accuracy of an opinion O is $-\sum_i P(\Theta = i)\log(O(i))$. The accuracy of O is limited by the information leakage of Θ. Specifically, given ratings $\mathbf{R}$, no matter how we select O, its accuracy cannot exceed a certain value, namely $H(\Theta|\mathbf{R})$. The theorem must state that O's accuracy cannot exceed a threshold determined by the information leakage.

Theorem 1. *There is no opinion O, such that $-\sum_i P(\Theta = i|\mathbf{R})\log(O(i))$ exceeds the threshold $H(\Theta) - I(\Theta; \mathbf{R})$.*

Proof. Using only standard notions from information theory: First, note $H(\Theta) - I(\Theta; \mathbf{R}) = H(\Theta|\mathbf{R})$. Second, $-\sum_i P(\Theta = i|\mathbf{R})\log(O(i)) = H(\Theta|\mathbf{R}) - D_{KL}(\Theta||O)$, which suffices, since $D_{KL}(\Theta||O) \geq 0$. □

5.1 Minimising Information Leakage

By definition, when all users are honest, there is non-zero information leakage. After all, if all users are honest $I(\Theta; \mathbf{R}_A) = I(\Theta; \widehat{\mathbf{R}}_A) > 0$. However, if some users are malicious, then there may not be information leakage. Formally:

Proposition 1. *There exist A, Θ, $\widehat{R}$, $\widetilde{R}$, S, such that $I(\Theta; \mathbf{R}_A) = 0$.*

Proof. Take $A = \{a, b\}$, $P(\Theta{=}0) = 1/2 = P(\Theta = 1)$, $P(\widehat{R}{=}\Theta) = 1$, $P(\widetilde{R}{=}1 - \Theta) = 1$ and $P(S_a{=}h) = P(S_b{=}h) = 1/2$. Obviously, honest ratings leak (1 bit of) information, however, the actual ratings leak no information (about Θ). □

On the other hand, it is not guaranteed for all A, Θ, $\widehat{R}$ and S, a malicious strategy exists that achieves zero information leakage:

Proposition 2. *There exist A, Θ, $\widehat{R}$, S, such that for all $\widetilde{R}$, $H(\Theta) - H(\Theta|\mathbf{R}_A) > 0$.*

Proof. Take A, Θ, $\widehat{R}$ as in Proposition 1, but $P(S_a{=}h) = P(S_b{=}h) = 0.51$. Now $P(\Theta = 1|R_a = 1, R_b = 1) \geq P(\Theta = 1, S_a = h, S_b = h|R_a = 1, R_b = 1) \geq 0.51$. □

Under certain specific circumstances, it is even possible to deduce exactly when it is possible for malicious advisors to block information leakage. The quantities depend on the exact assumptions. In [?,?,?], we looked at cases where the ratings perfectly match the option (i.e. full information leakage for honest users). For example, if malicious advisors are static and independent, the average probability of honesty must be below $1/n$, for n options [?]. In this paper, we do not quantify values, but study their relationships.

It may be possible for attackers to block information leakage (Proposition 1), but it may also be impossible (Proposition 2). Does the latter imply that there is no harmful attack? To answer that, we must determine the existence of an attack, such that the information leakage with the attack is lower than without. In fact, such an attack must always exist, provided that there is at least one user that has non-zero probability of being malicious.

Theorem 2. *For all A, Θ, $\widehat{R}$, S, there exists $\widetilde{R}$ such that $I(\Theta;\mathbf{R}_A) < I(\Theta;\widehat{\mathbf{R}}_A)$.*

Proof. Since $I(\Theta;\widehat{\mathbf{R}}_A)$, Θ and $\widehat{\mathbf{R}}_A$ are not independent, and there exists $\theta, \widehat{\mathbf{r}}_A$, such that $P(\theta|\widehat{r}) > P(\theta) + \epsilon$ and $P(\theta|\widehat{r'}) < P(\theta) - \epsilon$ for some other rating $\widehat{r'}$. Take $P(\widetilde{R} = \widehat{r}|\theta) = P(\widehat{R} = \widehat{r}|\theta) - \epsilon$, and $P(\widetilde{R} = \widehat{r'}|\theta) = P(\widehat{R} = \widehat{r'}|\theta) + \epsilon$. All summands but two remain the same: $P(\theta, \widehat{r}) \log P(\theta|\widehat{r}) + P(\theta, \widehat{r'}) \log P(\theta|\widehat{r'})$ are closer to their average, we can apply Jensen's inequality to get the theorem. □

So far, we have proven that some attacks may block all information leakage, but that such an attack may not exist, and that, nevertheless, a harmful attack must exist, except in trivial cases. These results suggest the possibility that all attacks reduce information leakage. However, this is not the case. There exist attacks that increase the information leakage:

Proposition 3. *There exist A, Θ, $\widehat{R}$, $\widetilde{R}$, S, such that $I(\Theta;\mathbf{R}_A) > I(\Theta;\widehat{\mathbf{R}}_A)$.*

Proof. Take A and Θ as in Proposition 1. Take $P(\widehat{R} = \Theta) = 0.6$, $P(\widetilde{R} = \Theta) = 0.7$ and $0 < P(\mathbf{S}_A) < 1$. The inequality is satisfied with these values. □

Notice that in the proof of Proposition 3, the information leakage of $\widetilde{R}$ is (strictly) greater than $\widehat{R}$. This is a necessary condition for an attack not to be harmful. However, it is not a sufficient condition. If we had taken $P(\widetilde{R} = \Theta) = 0.3$, then the information leakage of $\widetilde{R}$ remains the same, but the information leakage of R decreases (as long as $P(S)$ is not close to 0).

A realistic scenario where Proposition 3 could apply, is a camouflage attack. In the camouflage attack, an advisor provides high quality ratings (i.e. high information leakage), to gain trust, and later abuses the trust for a specific goal. In [**?**], we have studied these camouflage attacks, and identified that an existing attack is actually not harmful. Furthermore, we found that a probabilistic version of the camouflage attack can minimise information leakage.

If we want a trust system to be robust, then it must be able to deal graciously with the all malicious ratings, including the ones that minimise information leakage. The ratings that minimise information leakage are referred to as the *minimal* $\widetilde{\mathbf{R}}_{\mathbf{A}}$. Which ratings minimise information leakage depends on A, Θ, $\widehat{\mathbf{R}}_A$ and $\mathbf{S}_A$.

6 Design Guidelines

This section is the core of the paper. Here, we study choices that one can make to mitigate unfair rating attacks. This section is divided into subsections, each of which considers an individual choice. To give a quick overview of the results:

1. It is not helpful to block seemingly malicious advisors, and often counterproductive.
2. When the number of advisors is limited, seemingly honest advisors should be preferred.
3. Disregard advisors that should not have information about the decision; e.g. buyers that never bought from a seller.
4. When forced to choose between two seemingly equally honest advisors, the better-known advisor should be preferred.
5. Different groups of honest advisors whose ratings may have the same information leakage, but different robustness towards attackers. We find a property that characterises the robustness of ratings from equally informative groups of honest advisors. This shows that for a reasonable way to increase the size of a decision, information leakage increases.

In the relevant sections, we not merely show the results, but, more importantly, we analyse and interpret them. Some our suggestions are already widely adopted. However, they are adopted for reasons other than robustness. Moreover, our guidelines are based on a solid information-theoretic foundation.

6.1 Blocking Malicious Advisors

Theorem 2 states that the minimum information leakage of ratings is strictly smaller than the honest ratings. So problematic attacks reduce information leakage. Perhaps, we can robustly increase the information leakage, by blocking suspicious advisors. On the other hand, blocking suspicious advisors may decrease the information leakage, as even a suspicious advisor may provide honest ratings. We show in this section, that the latter holds: Blocking malicious advisors does not increase the robustness against unfair rating attacks.

First, we start with a weak version of the theorem, that directly refutes the intuition that sufficiently suspicious advisors must be blocked. We introduce a threshold of suspicion c, such that only those ratings from advisors at or above the threshold are considered. If indeed it helps to block sufficiently suspicious advisors, then such a c must exist. However, this is not the case:

Proposition 4. *For all A, Θ, $\widehat{R}$, $\widetilde{R}$ and S, there is no threshold $c \in [0,1]$, with $A^{\geq c} \subseteq A$ as the set of advisors such that $s_a \geq c$, such that $I(\Theta; \mathbf{R}_{A^{\geq c}}) > I(\Theta; \mathbf{R}_A)$.*

Proof. Since $H(X|Y,Z) \leq H(X|Y)$, $-H(\Theta|\mathbf{R}_{A^{\geq c}}) \leq -H(\Theta|\mathbf{R}_A)$. □

For a pair $c < d$, we can let $A' = A^{\geq d}$ and automatically $A'^{\geq c} = A^{\geq c}$, and the proposition applied to A' proves that $I(\Theta; \mathbf{R}_{A^{\geq c}}) > I(\Theta; \mathbf{R}_A^{\geq d})$. Therefore, Proposition 4 proves *monotonicity* of $I(\Theta; \mathbf{R}_{A^{\geq c}})$ over c.

For the vast majority of thresholds, however, blocking suspicious advisors is not just ineffective, but actually harmful. Thus, for some (small but non-zero) blocking thresholds, blocking does not alter information leakage, but for most thresholds – including all thresholds over $1/2$ – blocking strictly decreases information leakage:

Theorem 3. *For all A, Θ, $\widehat{R}$, S and minimal $\widetilde{R}$, there exists a threshold $d \in [1/n, 1)$ such that $I(\Theta; \mathbf{R}_{A^{\geq c}}) = I(\Theta; \mathbf{R}_A)$ iff $c \leq d$.*

Proof. Note that if $c = 0$ then trivially the equality holds, and if $c = 1$, then the equality trivially does not hold, since $s_a < 1$, $A^{\geq c} = \emptyset$ and thus $I(\Theta; \mathbf{R}_{A^{\geq c}}) = 0 < I(\Theta; \mathbf{R}_A)$. Using the monotonicity proved in Proposition 4, it suffices to prove the equality for $c = \frac{1}{n}$: Now $P(R_b|\Theta) = s_a P(\widehat{R}_b|\Theta) + (1 - s_a) P(\widetilde{R}_b|\Theta)$, and if $s_a \leq \frac{1}{n}$, there exists $\widetilde{R}_b$ such that $P(R_b|\Theta) = \frac{1}{n}$, meaning the minimum $\widetilde{R}_b$ can achieve zero information leakage for $c \leq \frac{1}{n}$. □

Arguably, Proposition 4 is not particularly interesting from an information-theoretic perspective – although it may be somewhat surprising superficially. After all, meaningless additional information does not decrease the information leakage. However, Theorem 3 strengthens the result to say that there exists a level of suspicion below which blocking is harming robustness. In [**?**,**?**], we show that for typical systems, this threshold is very low.

Design Hint 1 *Unless the suspicion that an advisor is malicious is extremely high, blocking a suspicious advisor is either useless or counterproductive for robustness.*

6.2 More Honest Advisors

The reason that blocking advisors does not help is the simple theorem that more random variables in the condition cannot decrease information leakage. However, clearly, a seemingly honest advisor contributes more information than a suspicious one. There may be a cost associated to requesting/receiving too many ratings, or other reasons why the number of advisors is limited. In these cases, we expect that preferring those advisors that are more likely to be honest is better for gaining information.

Take two near-identical trust systems, that only differ in the degree of honesty of the advisors. The decisions and honest ratings remain the same, and the attacker minimises information leakage on both sides. Then, the more honest system has higher information leakage:

Theorem 4. *For all $A, A', \Theta, \widehat{R}, S$, such that $|A| = |A'|$ and $P(S_{a_i}) \geq P(S_{a'_i})$, if $\widetilde{R}$ and $\widetilde{R}'$ are minimising, then $I(\Theta; \mathbf{R}_A) \geq I(\Theta; \mathbf{R}_{A'})$.*

Proof. The attacker can select $\widetilde{\mathbf{R}}'_A$ such that $P(\mathbf{R}_A = x|\varphi) = P(\mathbf{R}_{A'} = x|\varphi)$, for any condition φ. The minimum information leakage is at most equal to the construction's information leakage. □

Design Hint 2 *When there is a cost to gathering too many ratings, then seemingly more honest advisors should be asked before more suspicious ones.*

6.3 Unknowing Advisors

The first question is, whether it is useful to be aware of whether advisors could have knowledge. We introduce a random variable K_a, such that $K_a = 0$ if a does not have knowledge, and $K_a = 1$ if a may have some knowledge. Formally, $I(\Theta; \widehat{R}_a|K_a = 0) = 0$ and $I(\Theta; \widehat{R}_a|K_a = 1) > 0$. Now we can reason about the difference in information with and without K_a as a condition: $I(\Theta; \mathbf{R}_A)$ is the information leakage without knowing whether advisors could have knowledge, and $I(\Theta; \mathbf{R}_a|\mathbf{K})$ is the expected information leakage when we know whether advisors could have knowledge. In fact, in the latter case, we have at least as much information leakage:

Theorem 5. *For all $A, \Theta, \mathbf{R}_A$ and $\mathbf{K}$, $I(\Theta; \mathbf{R}_A) \leq I(\Theta; \mathbf{R}_A|\mathbf{K})$*

Proof. Since $H(\Theta|\mathbf{K}) = H(\Theta)$, $I(\Theta; \mathbf{R}_A|\mathbf{K}) = I(\Theta; \mathbf{R}_a, \mathbf{K})$. And additional conditions do not increase entropy, hence $I(\Theta; \mathbf{R}_A, \mathbf{K}) \geq I(\Theta; \mathbf{R}_a)$. □

Based on that result, we can revisit the notion of blocking users. Should we block unknowing advisors? It turns out that blocking unknowing advisors never changes information leakage:

Corollary 1. *For arbitrary conditions ψ, φ, and $a \in \mathcal{A}, \Theta, R_a, K_a$, we have $I(\Theta; R_a, \psi|K_a = 0, \varphi) = I(\Theta; \psi|\varphi)$*

Proof. Since $I(\Theta; \widehat{R}_a|K_a = 0, \varphi, \psi) = 0$, the optimal strategy for the malicious advisor is to set $\widetilde{R}_a$ to satisfy $I(\Theta; \widetilde{R}_a|K_a = 0, \varphi, \psi) = 0$, which implies $I(\Theta; R_a|K_a = 0, \varphi, \psi) = 0$. Then, R_a can be eliminated without loss of generality, and then also K_a. □

Design Hint 3 *When possible, keep track of whether an advisor can provide useful advise; e.g. because he actually performed a transaction. Advisors that cannot provide useful advise can be filtered out without loss of generality.*

6.4 Newer Advisors

Throughout the paper, we have ignored previous ratings. However, as we show in [**?**], if we learn about the correctness of an advisor's rating in hindsight, then we learn about the status of that advisor. Concretely, for some random variable Q, if $P(Q = 1|R_a = \widehat{R}_a) > P(Q = 1|R_a = \widetilde{R}_a)$, then $P(S_a|Q) \neq P(S_a)$. Here Q corresponds to the probability that an advisor's rating is correct in hindsight.

To keep the notation simple, we did not introduce Q previously. Here, we need it to prove a theorem.

For an honest user, R_a is always equal to $\widehat{R}_a$, and an attacker may or may not set R_a equal to $\widehat{R}_a$ (by selecting $\widehat{R}_a = \widetilde{R}_a$). Often, there may be multiple Q's per advisor, so we use $\mathbf{Q}_a^{\leq i}$ to denote $Q_1, Q_2, \ldots, Q_i$. Let a, a' be advisors, such that $P(S_a|\mathbf{Q}_a^{\leq i}) = P(S_{a'}|\mathbf{Q}_{a'}^{\leq j})$ and $i < j$. Then, a and a' are equally suspicious (their statuses are equiprobable with the given conditions), but a is a newer advisor than a'. For the current rating, a and a' are equally useful (as they are equally likely to be honest), but if the current rating will be correct in hindsight, then a loses suspicion quicker than a'.

For example, let a be completely new – its suspicion is the prior $P(S_a)$ – and a' be older, but have both seemingly correct and seemingly wrong ratings in hindsight, such that $P(S_{a'}|\mathbf{Q}) = P(S_a)$. Clearly, if we got a seemingly correct rating from a in hindsight, its suspicion level changes much more radically than when the rating of a' was correct in hindsight. The same holds if the rating was seemingly wrong. However, if we only select a or a', then the other advisor's suspicion level remains unchanged. Therefore, if we must select either a or a' and both are equally suspicious, then we should prefer the newer advisor.

To formalise this concept, measure the information leakage over two ratings, where we switch advisor if their ratings seemed wrong in hindsight:

Theorem 6. *For a, a', such that $P(S_a|\mathbf{Q}_a^{\leq i}) = P(S_{a'}|\mathbf{Q}_{a'}^{\leq j})$ and $i < j$:*

$$I(\Theta; R_a|\mathbf{Q}_a^{\leq i}) + P(Q{=}0|R_a, \Theta) \cdot I(\Theta; R_{a'}|\mathbf{Q}_{a'}^{\leq j}) + P(Q{=}1|R_a, \Theta) \cdot I(\Theta; R_a|\mathbf{Q}_a^{\leq i}, Q{=}1) \geq$$
$$I(\Theta; R_{a'}|\mathbf{Q}_{a'}^{\leq j}) + P(Q{=}0|R_a, \Theta) \cdot I(\Theta; R_a|\mathbf{Q}_a^{\leq i}) + P(Q{=}1|R_a, \Theta) \cdot I(\Theta; R_{a'}|\mathbf{Q}_{a'}^{\leq j}, Q{=}1).$$

Proof. Note that since $I(\Theta; R_a|\mathbf{Q}_a^{\leq i}) = I(\Theta; R_{a'}|\mathbf{Q}_{a'}^{\leq j})$, all terms on both sides are equal, except $I(\Theta; R_a|\mathbf{Q}_a^{\leq i}, Q)$ and $I(\Theta; R_{a'}|\mathbf{Q}_{a'}^{\leq j}, Q)$. Hence it suffices to prove $I(\Theta; R_a|\mathbf{Q}_a^{\leq i}, Q{=}1) \geq I(\Theta; R_{a'}|\mathbf{Q}_{a'}^{\leq j}, Q{=}1)$.

The Q's are independent of Θ, meaning that it suffices to prove $H(\Theta|R_a, \mathbf{Q}_a^{\leq i}, Q{=}1) \leq H(\Theta|R_{a'}, \mathbf{Q}_{a'}^{\leq i}, Q{=}1)$. Now, $P(\Theta|R_a, \mathbf{Q}_a^{\leq i}, Q{=}1) = P(\Theta|R_a, S^*)$, with $S_a^* \; S_a|\mathbf{Q}_a^{\leq i}, Q{=}1$, and similarly on the other side. Thus, if $\mathbb{E}(S_a^*) \geq \mathbb{E}(S_{a'}^*)$, then the theorem follows. Since the Q's are Bayesian updates, we can model S^* as a Beta distribution times an arbitrary prior. The expectation of the Beta distribution is $\frac{p+1}{p+n+2}$, which is more sensitive to increasing p when $p + n$ is small. □

Design Hint 4 *When two advisors appear equally likely to be honest, but only one may provide a rating, then the advisor with whom we have the shortest history should be preferred.*

6.5 More Options

The defender may want to increase the robustness of the rating system, by changing the rating format or the number of choices in a decision. There is an immediate problem when formalising this idea, which is that we have not formalised how honest users respond when either the domain of Θ or R changes.

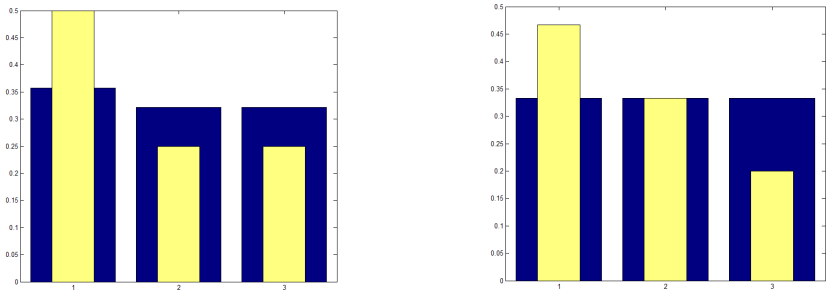

Fig. 1. Two ratings (blue), the corresponding honest ratings (yellow), with the values of θ on the x-axis, and the corresponding probability on the y-axis.

Naively, we may simply assert that the information leakage of honest ratings remains unchanged. Thus $I(\Theta;\widehat{\mathbf{R}}_A) = I(\Theta';\widehat{\mathbf{R}}'_A)$. However, the robustness of the system on the left is not equal to that on the right.

Theorem 7. *For given $A, \Theta, \Theta', \widehat{\mathbf{R}}, \widehat{\mathbf{R}}'$ with minimising $\widetilde{\mathbf{R}}$ and $\widetilde{\mathbf{R}}'$, $I(\Theta;\widehat{\mathbf{R}}_A) = I(\Theta';\widehat{\mathbf{R}}'_A)$ does not imply $I(\Theta;\mathbf{R}_A) = I(\Theta';\mathbf{R}'_A)$.*

Proof. Let the prior $P(\Theta) = P(\Theta') = 1/3$. Moreover let $P(\Theta = i|\widehat{\mathbf{R}}_A = i) = 1/2$ and $P(\Theta = i|\widehat{\mathbf{R}}_A \neq i) = 1/4$; meaning $I(\Theta;\widehat{\mathbf{R}}_A) = 1.5$. Then, we let $P(\Theta' = i|\widehat{\mathbf{R}}'_A = i) = 7/15$, then it follows from the fact that the information leakage is 1.5 that $P(\Theta' = i|\widehat{\mathbf{R}}'_A \equiv_3 i+1) \approx 0.195 \approx 1/5$ and $P(\Theta' = i|\widehat{\mathbf{R}}'_A \equiv_3 i+2) \approx 0.338 \approx 1/3$ (or vice versa). See Figure 1. When $P(\mathbf{S}_A) \approx 5/7$, we can achieve 0 information leakage by setting $P(\Theta' = i|\widehat{\mathbf{R}}'_A \equiv_3 i+1) = 1/3$, and $P(\Theta' = i|\widehat{\mathbf{R}}'_A \equiv_3 i+1) = 2/3$, then it follows that $P(\Theta'|\mathbf{R}'_A) = 1/3 = P(\Theta' = i)$. Thus, for $P(\mathbf{S}_A) \approx 5/7$ we have 0 information leakage for Θ', but since $P(\Theta = i|\mathbf{R}_A = i) \geq P(\Theta = i|\widehat{\mathbf{R}}_A = i) \cdot 5/7 = 5/14$, $P(\Theta = i|\mathbf{R}_A = i) \neq 1/3 = P(\Theta = i)$, there is non-zero information leakage for Θ. Hence their robustness is different. □

The proof is visualised in Figure 1, where the yellow/light bars are the honest ratings, and the blue/dark bars are the overall ratings. The honest ratings have the same information leakage in both graphs, whereas the overall ratings clearly do not.

More choices generally mean more information leakage. However, as proven in Theorem 7, special circumstances must be met when expanding Θ. In particular, Θ is expanded together with R in an orthogonal way – the additional options have their own corresponding ratings:

Theorem 8. *Given two rating systems that one has more options for ratings and choices: $|\Theta| = |\Theta'| + 1, |\mathbf{R}_A| = |\mathbf{R}'_A| + 1$, if $p(\theta|r) = p(\theta'|r')$, then $I(\Theta; R_a)$ can be equal, or larger than $I(\Theta'; R'_a)$*

Proof. Let $p(\theta|r) = 0$ for either $\theta{>}|\Theta'|$ or $r{>}|\mathbf{R}'_A|$, except that there is a $\dot{r} > |\mathbf{R}'_A|$ and a $\dot{\theta} > |\Theta'|$, for which $p(\dot{\theta}|\dot{r}) = 1$. We get $H(\Theta) = H(\Theta') - \mathbf{f}(p(\dot{r}))$, and

$H(\Theta|R_a) = H(\Theta'|R'_a) - p(\dot{r}) \cdot \mathbf{f}(p(\dot{\theta}|\dot{r})) = H(\Theta|R'_a)$. Given $p(\dot{r}) \geq 0$, $H(\Theta) - H(\Theta|R_a) \geq H(\Theta') - H(\Theta'|R_a)$. Hence, $I(\Theta; R_a) \geq I(\Theta'; R'_a)$. □

Intuitively, one of two ratings can happen, if the additional rating occurs, the user gains a lot of information – namely that the additional option occurred. The remaining cases do not involve the new rating or choice, and remains essentially unchanged (only linear weights change proportionally).

Design Hint 5 *Increasing the number of options in a decision is good, assuming that the additional options are sufficiently distinctive for the advisors. Care should be taken, because additional options may harm robustness in some cases.*

7 Conclusion

Users form opinions based on ratings. Systems that allow more accurate opinions are better. We use cross entropy (or Kullback-Leibler divergence) to measure the accuracy of opinions. The maximum accuracy is limited by a quantity called information leakage. Information leakage measures how much a rating tells about the decision a user wants to make.

The amount of information leakage of honest ratings has a certain non-zero quantity. Thus, we assume that there is some correlation between honest ratings and what the best decision is. We cannot make such assumptions about ratings from attackers. Attackers have a hidden agenda that they base their ratings upon. We want to reduce the negative effect that attackers have on the information leakage. To be on the safe side, we assume that attackers rate in the way that minimises the information leakage – any other behaviour results in at least as much information leakage.

Our model of a rating system is abstract. Decisions and ratings are abstract entities. Ratings are not under the control of the defender. However, the defender can select which advisors to use, potentially monitor whether an advisor may be knowledgable, and consider more options in his decision. Our main contribution is a set of guidelines for the defender to use these factors in his advantage.

Our main guidelines are: Blocking suspicious advisors can only decrease robustness (1). If only a limited number of ratings can be used, however, then less suspicious advisors are better (2), and in case of a tie, newer advisors are better (3). Observing transactions increases robustness (4). Offering more choices may increase robustness (5).

References

1. Thomas M Cover and Joy A Thomas. Entropy, relative entropy and mutual information. *Elements of Information Theory*, pages 12–49, 1991.
2. Hui Fang, Yang Bao, and Jie Zhang. Misleading opinions provided by advisors: Dishonesty or subjectivity. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (IJCAI)*, pages 1983–1989, 2013.

3. Nan Hu, Ling Liu, and Vallabh Sambamurthy. Fraud detection in online consumer reviews. *Decision Support Systems*, 50(3):614–626, 2011.
4. Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.
5. Radu Jurca and Boi Faltings. Minimum payments that reward honest reputation feedback. In *Proceedings of the 7th ACM conference on Electronic commerce*, pages 190–199. ACM, 2006.
6. Radu Jurca and Boi Faltings. Collusion-resistant, incentive-compatible feedback payments. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 200–209. ACM, 2007.
7. Reid Kerr and Robin 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. IFAAMAS, 2009.
8. Robert J. McEliece. *Theory of Information and Coding.* Cambridge University Press New York, USA, 2nd edition, 2001.
9. Kevin Regan, Pascal Poupart, and Robin Cohen. Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 1206–1212, 2006.
10. W. T. Luke Teacy, Michael Luck, Alex Rogers, and Nicholas R. Jennings. An efficient and versatile approach to trust and reputation using hierarchical bayesian modelling. *Artificial Intelligence*, 193:149–185, 2012.
11. Dongxia Wang, Tim Muller, Athirai A Irissappane, Jie Zhang, and Yang Liu. Using information theory to improve the robustness of trust systems. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 791–799, 2015.
12. Dongxia Wang, Tim Muller, Jie Zhang, and Yang Liu. Quantifying robustness of trust systems against collusive unfair rating attacks using information theorys. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 111–117, 2015.
13. Dongxia Wang, Tim Muller, Jie Zhang, and Yang Liu. Is it harmful when advisors only pretend to be honest? In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016.
14. Jianshu Weng, Zhiqi Shen, Chunyan Miao, Angela Goh, and Cyril Leung. Credibility: How agents can handle unfair third-party testimonies in computational trust models. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(9):1286–1298, 2010.
15. Andrew Whitby, Audun Jøsang, and Jadwiga Indulska. Filtering out unfair ratings in bayesian reputation systems. In *Proceedings of the AAMAS Workshop on Trust in Agent Societies (TRUST)*, pages 106–117, 2004.
16. Yafei Yang, Yan Sun, Steven Kay, , and Qing Yang. Securing rating aggregation systems using statistical detectors and trust. *IEEE Transactions on Information Forensics and Security*, 4(4):883–898, 2009.
17. Han Yu, Zhiqi Shen, Chunyan Miao, Bo An, and Cyril Leung. Filtering trust opinions through reinforcement learning. *Decision Support Systems*, 66:102–113, 2014.
18. Jie Zhang and Robin Cohen. Design of a mechanism for promoting honesty in e-marketplaces. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI)*, pages 1495–1500, 2007.