Information Theoretical Analysis of Unfair Rating Attacks under Subjectivity

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

***Abstract*—Ratings provided by advisors can help an advisee to make decisions, e.g., which seller to select in e-commerce. Unfair rating attacks – where dishonest ratings are provided to mislead the advisee – impact the accuracy of decision making. Current literature focuses on specific classes of unfair rating attacks, but this does not provide a complete picture of the attacks. We provide the first formal study that addresses all attack behaviour that is possible within a given system. We propose a probabilistic modelling of rating behaviour, and apply information theory to quantitatively measure the impact of attacks. In particular, we can identify the attack with the worst impact. In the simple case, honest advisors report the truth straightforwardly, and attackers rate strategically. In real systems, the truth (or an advisor's view on it) may be subjective, making even honest ratings inaccurate. Although there exist methods to deal with subjective ratings, whether subjectivity influences the effect of unfair rating attacks was an open question. We discover that subjectivity decreases the robustness against attacks.**

***Index Terms*—Unfair Rating Attacks, Worst-Case Attacks, Robustness, Subjective Rating, Trust Systems**

I. Introduction

Users can help each other make decisions by sharing their opinions, especially when direct experience or evidence is insufficient. Ratings are a discrete form of such shared information. Rating mechanisms are popularly applied in existing online systems, such as trust systems, recommender systems, e-commerce systems and security systems [1], [2], [3], [4].

Not all ratings accurately reflect reality. Malicious advisors (attackers) may deliberately provide fake or unreliable ratings to impact the decisions of some other users (advisees). This is known as an unfair rating attack. Unfair rating attacks reduce the accuracy of decision making based on ratings.

Many approaches have been proposed in the literature to improve the robustness of trust systems against unfair rating attacks. What is typically proposed is to estimate trustworthiness of advisors, based on which ratings are discounted or filtered. We argue that being aware of advisors' honesty is not sufficient to have a complete picture of unfair rating attacks. How advisors behave when they are dishonest also needs to be studied. There are also some approaches which aim to propose countermeasures for the existing types of attacks. However, attackers can always adapt their behaviour or strategies to the countermeasures. In such an arms-race, the system designer will be one step behind the attackers.

We propose a probabilistic modelling of a rating by an arbitrary advisor, which allows us to consider various possible strategies of a random attacker. Given the uncertainty in the behaviour of the attacker, we propose to investigate the worst-case scenario that he can cause to an advisee. From a security perspective, a secure system should be prepared for the worst-case attacks. Measuring which unfair rating attack is worse requires the right metric.

An advisee aims to get information about the observed facts leading to a rating. From the advisee's perspective, an attack hinders his learning about the facts. How much information about the facts a rating provides can be quantified by a measurement from information theory, namely, information leakage [5]. The less information ratings leak about the facts, the greater the impact of the attack, since the advisee is more hindered in his learning. The worst-case scenario for the advisee is that the attack is the one that minimises the information leakage about the facts. We say that a system is more robust than another system, when in all situations, the maximum impact of an attack is lower in the former system than in the latter.

Dishonesty is not the only element that reduces the quality of ratings, in realistic scenarios. Honest advisors can be subjective in rating, or have different preferences from an advisee. The same observed fact may cause different honest advisors to provide different ratings. More importantly, an honest advisor may come to a different conclusion than an advisee would have, given the same facts.

In the literature, dishonesty and subjectivity are usually treated separately, or with one as the special case of the other. It was an open question whether subjectivity changes the effects of dishonesty or unfair rating attacks. We compare the difficulty of achieving strong unfair rating attacks in rating with and without subjectivity, and find that the existence of subjectivity makes attacks easier. We then introduce an ordering of subjectivity, based on which we prove that higher degree of subjectivity means being less robust against unfair rating attacks.

Methods to mitigate the negative effect of subjectivity on rating's accuracy have been proposed. Since subjectivity decreases robustness of rating systems, we study whether these methods improve the robustness. The first method is for advisors to rate individual features of a target (feature-based rating) instead of only providing an overall rating. We compare the restrictions that feature-based rating impose on achieving ultimate attack to that of overall rating. We find feature-based rating does not necessarily improve robustness compared to overall rating, and may even worsen it. Clustering advisors based on their behaviour is another way to discern subjectivity difference. We alter the rating model to allow clustering, and find that clustering increases expected information leakage regardless of attackers' strategies (and hence robustness). There exist different ways to deal with clusters, e.g., excluding seemly dishonest clusters or exploiting all clusters. We find that clusters should only be excluded in extreme cases.

Our main contribution is a formal way to measure the amount of information a rating carries. Attackers can rate in different ways, which affects how informative a rating is. Our measure allows us to 1) compare the impact of different attacks, 2) identify the circumstances under which an attacker can eliminate all information, and 3) find the behaviour that minimises the information. We first present a measure for objective rating, and then a more involved version for subjective rating. Our measure allows us to formally reason about the interplay between subjectivity and dishonesty, which has not been done before. Furthermore, it also allows us to formally reason about approaches that deal with both subjectivity and dishonesty. In particular, we look at feature-based rating and at clustering.

The work in this paper mainly consists of two parts. In the first part, we study attacks where honest advisors are assumed to be objective in rating – an essential scenario[1]. We propose a probabilistic rating model and an information-leakage based quantification method, as a basis of the study on unfair rating attacks throughout the paper. We find the worst-case attack strategies. In the second part, we study the effects of attacks when honest advisors can be subjective in any ways, emphasizing a comparison with the earlier results. We also study whether the existing methods of dealing with subjectivity would influence robustness against attacks.

II. Related Work

We survey related approaches that solely deal with unfair rating attacks, and also those that consider both dishonest ratings and subjective ratings.

A. Dealing with Unfair Rating Attacks

Unfair rating attacks, also known as misleading feedback attacks, reduce the accuracy of rating-based decision making. They are among one of the most popular types of attacks in trust and reputation systems [7]. Various approaches have been proposed to diminish the effect of unfair rating attacks. Most of them rely on estimation of the trustworthiness of advisors (called recommender trust or feedback reputation) to judge the quality of ratings. Typically, ratings are discounted/filtered based on trustworthiness of advisors, before being aggregated. Trustworthiness of advisors can be evaluated using different considerations, e.g., advisor's similarity with an advisee [8], [9], [10], [11], the time of rating [12], [13], and the consistency between previous ratings and the observed outcomes [14].

Similarity is a popular criterion for evaluation advisors and their ratings. Weng et al. determine the credibility of an advisor by measuring the statistical correlation between its ratings and an advisee's own experiences regarding the same targets [9]. A local table is built for each advisor, to store their past ratings and the advisee's experiences. Ratings are weighted with the advisors' credibility values, which need to exceed a threshold set by the advisee's own confidence. In [8], Zhang and Cohen propose to use both local and global ratings to estimate the trustworthiness of an advisor. Whether a rating is local or global is determined by whether it refers to the target under evaluation or other targets in the system. For example, ratings about other sellers from an advisor are useful when there are few sellers with whom both the buyer and the advisor have interacted. Liu et al. propose an approach called iCLUB [11], where ratings are clustered based on their similarity. Advisors whose ratings are in the same clusters with an advisee's ratings are considered reliable by the advisee. Ratings from other advisors are considered as unfair and would be filtered. Note that the clustering does not distinguish whether filtered ratings are dishonest or subjective. Liu et al. also propose to use Dempster-Shafer theory to combine information from both local ratings and global ratings to identify trustworthy advisors [10].

Alternatively, the time domain of rating can be employed. In [12], Yang et al., propose to detect suspicious ratings and also time intervals where attacks are more likely. The detection results help decide in what degree advisors can be trusted. Highly suspicious ratings are removed. In [15], a technique called CUSUM [16] is employed to detect suspicious time intervals where attacks very likely happen. To avoid mistakenly selecting normal but deviating ratings as suspicious, the correlations among advisors are then learned to identify which ratings are from colluding malicious advisors (which are assumed to have large correlation). The Expectation Maximization algorithm and hypothesis test method are applied to resist random and coordinated malicious rating attacks in [17].

Additionally, in [13], three aspects of rating behaviour are considered to evaluate advisors: the time when ratings are provided, similarity between an advisor and the advisee, and also confidence of the advisor. For example, from the time aspect, ratings provided more recently are considered to be more reliable. From confidence aspect, ratings from a more experienced advisor are considered to be more convincing. Fuzzy logic is applied to fuse these three aspects. And finally, in [14], Yu et al. propose a reinforcement-learning based approach. An advisor's trustworthiness is updated after each interaction based on whether its ratings are consistent with the actual observed behaviour of the rated target.

Accurate evaluation of advisors' honesty is crucial when it is the only criterion to assess ratings, but it is difficult to achieve. If the evaluation mechanism is majority rule[2], malicious advisors can choose targets that are rated by only a few honest advisors, and make their false ratings be the majority. In this way, they can reduce the reputation of those honest advisors and increase their own reputation, deceiving advisees. This kind of behaviour is known as Reptrap attack [18]. In fact, as we presented in our previous work [6], even if accurate degrees of advisors' honesty are given, trust models may still perform poorly under strong attacks. Dishonest advisors can pick strategies, some of which can be much more serious than the others. It is insufficient to focus on only on whether an advisor is honest, but to ignore their strategies. We aim to find and study the most serious attack strategies.

[1] The work in this part has been published in [6].

[2] Ratings that belong to minority would be treated as unreliable.

In approaches based on advisor reputation, ratings from dishonest advisors are usually discounted or filtered, possibly resulting in a loss of useful information. Some approaches try to make use of dishonest ratings. BLADE [19] and HABIT [20] aim to extract useful information from (dishonest) ratings, as long as there is statistical correlation between the ratings of an advisor and the advisee's own experience. For example, if an advisor always rates with a negative bias for an advisee, HABIT correct this bias when using his ratings. BLADE records the correlation by building behaviour functions for advisors, which serve to interpret their ratings. For instance, if an advisor is detected to always badmouth a reputable seller, his ratings would be reversed in the future. These approaches indicate the importance of correlation between ratings and the underlying facts. Regardless of the forms of attacks, if ratings are sufficiently correlated with the facts, there can be a way to uncover the truth. However, if there is little correlation, the truth can hardly be learned. Based on this reasoning, we propose to measure the impact of an attack by quantifying how much information it provides about the truth. That means, we do not use criteria such as some heuristic perceptions of attacks[3] or the direct effects of attacks on a specific system. Our measurement is general and would not be confined to specific systems.

Instead of judging the honesty of advisors before dealing with their ratings, some approaches directly detect and filter unfair ratings, using statistical methods for example. Weng et al. propose an entropy-based method to measure the deviation of ratings from an advisee's own experience [21]. Ratings that deviate too much are removed. This methodology is sometimes called endogenous filtering, but it is a highly problematic approach [22] Alternatively, contextual information can be used for filtering. For example, Wang et al. propose a detection-based method for web service recommendation system [4]. They aim to identify malicious ratings and also find the corresponding advisors' IP addresses. These advisors would then be refused to rate by the server.

Many defense mechanisms have assumptions about attackers' rating behaviour. For example, bad-mouthing[4] and ballot-stuffing attacks[5] are the most popularly studied unfair rating attacks. Sometimes, more complex attacks are analysed. Feng et al. study three types of attacks, namely RepBad, RepSelf and RepTrap [23]. Jiang et al. propose a trust model based on evolutionary computation (named MET) to cope with four types of attacks and their combinations [24]. Liu et al. study attacks that come from a cyber competition where human participants compete to break down a trust system [13]. To be able to resist well-known attacks is useful, however, it cannot ensure robustness faced with future attack strategies. In fact, to assume attackers' behavior makes defense passive, as attackers can adapt their strategies, especially when they are aware of the system design. With this in mind, we study from an active perspective – we want to figure out what would be the worst case that attackers can cause. From a security view, a secure (robust) system should be prepared for the worst-case attacks.

Finally, we note that instead of dealing with dishonest rating, some approaches aim to disincentivise advisors to rate dishonestly. Zhang et al., propose to reward reputable advisors by making sellers provide products with lower prices but increased quality [25] to them. In [26], a limited inventory of each seller is considered, where buyers compete with each other to get the purchase. Buyers that report truthful ratings are assigned higher scores, making them have more opportunities to transact with reputable sellers.

B. *Dealing with Attacks under Subjective Rating*

So far, when we use the term "unfair ratings", we mean ratings that are deliberately provided by strategic advisors. In some works, "unfair ratings" refer to any ratings that indicate divergent opinions with an advisee, even if the divergence comes from the conflict interests or views between an honest advisor and the advisee, e.g., subjective ratings [27]. Here, we distinguish two kinds of ratings using different terms: "subjective ratings" are from honest advisors with different opinions, while "unfair ratings" still denotes ratings from attackers.

Subjectivity is typically unavoidable in realistic rating systems. Both subjectivity and dishonesty may cause biased ratings, impacting rating-based decision making. Considering their analogous negative effect on the accuracy of ratings, some researchers treat them equally without distinguishing the motivations of advisors [27], [13], [9]. Some others propose to differentiate subjective but honest advisors from dishonest ones. In [28], Fang et al. propose to use a clustering scheme for each advisee to identify his advisors as subjective or dishonest. Advisors with similar subjectivity are clustered in a same group. An advisee can make use of ratings from both its subjective groups of advisors and also dishonest advisors' groups, if the dishonest advisors have fixed behaviour pattern. Noorian et al. propose a two-layer filtering approach where the first layer excludes malicious advisors, and the second layer discerns the dispositions of the remaining advisors [29].

Interestingly, subjectivity may change the nature of unfair rating attacks. For example, Noorian et al. [30] consider an attack based on using subjectivity, where dishonest advisors disguise as honest-but-subjective. The effects of subjectivity and dishonesty are not additive. Hence, we formally study how subjectivity influences the robustness against dishonest behaviour.

III. Preliminaries

We briefly introduce some concepts from information theory, which support our work throughout this paper.

Shannon entropy is used to measure the expected amount of information carried in a random variable, which is determined by the uncertainty of the random variable [31]:

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

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

[3] For example, some people may naturally think that it is the worst case if advisors always lie.

[4] Dishonest advisors slander reputable targets, e.g., in collusion with other targets to defame their competitors.

[5] Dishonest advisors provide untruthful positive ratings for a target, e.g., who bribed them to promote his reputation.

The Shannon entropy is maximal when all possible outcomes are equiprobable. The base of the logarithm is set to 2, wlog.

Conditional entropy measures the expected amount of information in one random variable when another random variable is known [31]:

Definition 2. *(Conditional entropy) The conditional entropy of a discrete random variable X under Y is given as:*

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

$H(X|Y) = H(X)$ iff X and Y are independent. For brevity, we leave out the cases where only one of X and Y is continuous. Note that $0 \leq H(X|Y) \leq H(X)$.

Information leakage measures the gain of information about one random variable when another random variable is known. This definition coincides with mutual information [5]:

Definition 3. *(Information leakage) The information leakage of X under Y is given as:*

$$H(X) - H(X|Y) = \sum_{x,y} p(x,y) \cdot \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$

Only independent random variables do not leak information about each other and vice versa:

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

A crucial theorem for proving inequalities, is Jensen's inequality [32]. Applied to probabilities, it states that the uniform distribution has the lowest entropy, and that distributions closer to the uniform distribution have lower entropy:

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

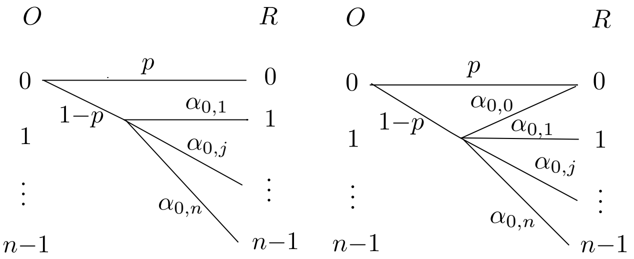
$$f\left(\frac{\sum_i a_i \cdot x_i}{\sum_i a_i}\right) \leq \frac{\sum_i a_i f(x_i)}{\sum_i a_i}$$

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

We introduce some shorthand which will be used throughout the paper. Given variable X, the lower-case x denotes one of its outcomes, and moreover $P(x)$ means $P(X{=}x)$. $\forall x$ means for any x in X's outcome set. We typically omit the domain of such variables and, for example, write $\sum_x$ to denote the summation over all outcomes of X. Since $x\log(x)$ is a common term, we introduce the shortcut $\mathbf{f}(x) = x\log(x)$. For practical reasons, we let $\mathbf{f}(0) = 0\log(0) = 0$.

IV. Quantifying Attacks under Objective Rating

In this section, we quantitatively study the effects (impact) of unfair rating attacks. Specifically, we consider the scenario where an arbitrarily selected advisor is rating a given subject. One important aim is to find the worst-case attack strategies that the advisor can undertake, for this specific rating. For now, we assume that honest advisors' ratings are objective, meaning that they are equal to the observed facts. We do not assume specific behaviour for attackers. The probabilistic rating model we propose considers any possible degrees of honesty and behavior of an arbitrary advisor, within the restriction of our assumptions. The work in this section has been published in [6].



(a) Naive objective rating model (b) Proper objective rating model

Fig. 1. Objective rating models with n options of observable facts and ratings.

A. Modeling Objective Rating

A rating process consists of an advisor who reports a rating, based on his observation of a fact about the target under evaluation, and an advisee who wants to use ratings to make decisions regarding the target. Take an e-commerce system as an example, a buyer can use ratings from other buyers about, e.g., the quality of a product or the reliability of sellers to decide whether to buy the product, or which seller to choose. We consider a set-up with a single advisor in a single rating. Random variables R and O represent the rating and the observed fact behind the rating. The exact meaning of O depends on the purpose of a system.

Consider a simple example, in rating whether an app is malware or not, O has two outcomes "Yes/No". Outcomes of O are discrete and finite, which w.l.o.g. are labeled as $0, \cdots, n{-}1$ ($n{>}1$). For now, we assume that options of rating are the same as the possible outcomes of O[6], which also belong to the list $0, \cdots, n{-}1$. For an advisee, without R, O is assumed to have maximum uncertainty, meaning its prior distribution is uniform and $H(O) = \log(n)$.

We characterize an arbitrary advisor's behaviour, considering both him being honest (with probability p, denoted as H), and being dishonest (or strategic, with probability $1-p$, denoted as $\neg H$). The probability that an advisor is honest can be interpreted in a Bayesian framework as representing the knowledge of an advisee about the honesty of the advisor in question. In a frequentist interpretation, the value p could indicate that there is a population where a fraction of size p of the advisors is honest, and $1{-}p$ is malicious, and we randomly select an advisor from this population.

Given an observation $O{=}i$, an honest advisor always reports the truth, i.e., $P(R{=}i|O{=}i, H) = 1$. How a dishonest advisor rates can be characterized by the conditional probabilities of all rating options, i.e., $P(R{=}j|O{=}i)$, $j{\in}\{0, \cdots, n{-}1\}$. All these conditional probabilities form an $n{\times}n$ matrix denoted as α, with $\alpha_{i,j}{=}P(R{=}j|O{=}i)$ and $\sum_j \alpha_{i,j} = 1$. Subscript i, j also denote the row and column index of the entry $\alpha_{i,j}$, with both starting from 0. Whatever behaviour an attacker exhibits, there exists a matrix α that describes it. Initially, we assume that attackers will not tell the truth, i.e., $\forall_i \alpha_{i,i} = 0$. This rating set-up is the *naive rating model*, shown in Figure 1(a).

[6] There are a lot of rating systems that define fixed options of ratings, e.g., five-star rating systems provided by Tripadvisor and Amazon, scoring system of Booking and Taobao.

B. Ultimate Attacks

For an advisee, to deduce the truth behind a rating, the rating needs to be correlated in some way to the observation. If the rating is completely independent of the observation, then there is no way to learn the truth. We name attacks which cause this extreme case as ultimate attacks. The strategy to achieve ultimate attacks in the naive rating model is provided in Theorem 2:

Theorem 2. *In the naive rating model, rating R is independent of observation O iff $p = \frac{1}{n}$ and $\alpha_{i,j} = \frac{1}{n-1}(i{\neq}j)$.*

Proof. If variables O and R are independent, then $P(R{=}j|O{=}j) = P(R{=}j|O{=}i)$, for all $j,i{\in}\{0,\cdots,n{-}1\}$ and $i{\neq}j$. The equation can be rewritten as $p = (1{-}p)\alpha_{i,j}$. Since $\sum_j \alpha_{i,j}{=}1$, namely $\frac{(n-1)p}{1-p}{=}1$, we get $p{=}\frac{1}{n}$ and $\alpha_{i,j}{=}\frac{1}{n-1}$. On the other hand, when $p{=}\frac{1}{n}$ and $\alpha_{i,j}{=}\frac{1}{n-1}$, $P(R{=}j|O{=}i){=}\frac{1}{n}$ holds for all j,i, which implies the independence between O and R. □

Intuitively, we expect that lower values of p (more probably an attacker) should make it easier to hide O. However, Theorem 2 implies that when $p{\leq}\frac{1}{n}$, the observation cannot be perfectly hidden, whereas for $p{=}\frac{1}{n}$, it can. Therefore, we need to alter the naive rating model to accommodate for the case $p{<}\frac{1}{n}$.

When $p{<}\frac{1}{n}$, the independence of O and R implies that $\sum_{j\neq i}\alpha_{i,j}{<}1$, which is impossible in the naive rating model. This is caused by the fact that the advisor is forced to lie if he is strategical. 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 report the truth, as part of the deceit. Consider a real-world scenario: in 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 when he states Queen, and Alice has the King, Alice would know that Bob actually has the Ace. Thus, as a strategical player, Bob should sometimes report the truth to deceive Alice.

It is sometimes assumed that dishonesty implies not telling the truth. The above argument establishes that not allowing attackers to tell the truth would be a modelling error. Therefore, we introduce an alternative rating option $\alpha_{j,j}$ (e.g., $\alpha_{0,0}$ when $j = 0$) for a dishonest advisor, as depicted in the proper rating model in Figure 1(b). Now the strategy to achieve ultimate attacks changes as follows:

Theorem 3. *In the proper rating model in Figure 1(b), rating R is independent of observation O iff $p{\leq}\frac{1}{n}$ and $\alpha_{i,j}{=}\frac{p}{1-p}{+}\alpha_{j,j}$.*

Proof. If O and R are independent, then $\forall\{i,j\}, i{\neq}j, P(R{=}j|O{=}j){=}P(R{=}j|O{=}i)$. The equation can be rewritten as $p{+}(1{-}p)\alpha_{j,j}{=}(1{-}p)\alpha_{i,j}$ or $\alpha_{i,j}{=}\frac{p}{1-p}{+}\alpha_{j,j}$. Take i fixed, and sum over $j,j{\neq}i$ on both sides, we get $1{-}\alpha_{i,i}{=}\frac{(n-1)p}{1-p}{+}\sum_{j\neq i}\alpha_{j,j}$. Since $\sum_j \alpha_{j,j}{\geq}0$, we get $\frac{1-np}{1-p}{\geq}0$ and $p{\leq}\frac{1}{n}$. On the other hand, if $p{\leq}\frac{1}{n}$ and $\alpha_{i,j}{=}\frac{p}{1-p}{+}\alpha_{j,j}$, $P(R{=}j|O{=}i){=}p{+}(1{-}p)\alpha_{j,j}$ holds for all j,i, which means O and R are independent. □

When $\sum_j \alpha_{j,j}{=}0$, we get $\alpha_{i,j}{=}\frac{p}{1-p}$. As $\sum_j \alpha_{i,j}{=}1$, we get $p{=}\frac{1}{n}$ and $\alpha_{i,j}{=}\frac{1}{n-1}$, in which way Theorem 3 equals Theorem 2. Note that $\sum_j \alpha_{j,j}{>}0$ may occur in ultimate attacks, which implies dishonest advisors may sometimes report the truth without leaking information.

It is common for trust and reputation systems, as well as some security-related systems (e.g. blockchains []), to have the precondition that at least half the participants are honest. Theorem 3 suggests that this requirement may be too strong. Theorem 3 indicates that R and O cannot be independent when $p{>}\frac{1}{n}$, which means that, for $n{>}2$ and $\frac{1}{2}{>}p{>}\frac{1}{n}$, in the frequentist interpretation, even if the advisee selects an advisor from a population that contains more attackers than honest advisors, the advisee would still learn from the rating. The larger n becomes, the larger the fraction of attackers in the population is allowed to be.

C. Minimizing Information Leakage

Ultimate attacks are the worst-case attacks since an advisee cannot learn anything about the truth. Although ultimate attacks cannot be achieved when $p{>}\frac{1}{n}$, some strategies should still be better at hiding the observations than others. To capture this, we quantitatively measure how much a rating is correlated or dependent to the observation, using *information leakage* (Definition 3 in Section III). The information leakage between R and O measures how much information R provides about O. There is 0 information leakage iff R and O are independent, i.e., ultimate attacks. Less information R leaks implies that O is hidden better. The *impact* of an attack can then be quantified by information leakage. An attack has a larger impact than another attack, when its information leakage is less.

Below, we aim to find the strategies with the largest impact, for $p{>}\frac{1}{n}$, namely the strategies that minimise the information leakage. We may refer to these strategies as the *worst-case* strategies. The attacker partially controls R given O, so $H(O|R)$ is variable, but $H(O)$ is not controlled by the attacker.

Definition 4. *Level strategy is the strategy where: $\forall j$, $\alpha_{j,j} = 0$ and $\forall i, i{\neq}j$, $\alpha_{i,j} = \frac{1}{n-1}$.*

The level strategy minimises information leakage:

Theorem 4. *For $p \geq \frac{1}{n}$, the level strategy minimises the information leakage of O given R.*

Proof. Let $h_j = p{+}(1-p)\sum_i \alpha_{i,j}$ for all i,j.

$$-H(O|R){=}\sum_j P(R{=}y_j)\sum_i P(O{=}x_i|R{=}y_j)\log P(O{=}x_i|R{=}y_j)$$

$$=^1 \frac{1}{n}\sum_j \Big(\sum_{i\neq j}(1-p)\cdot\alpha_{i,j}\log(\frac{(1-p)\cdot\alpha_{i,j}}{h_j})$$

$$+ (p + (1-p)\alpha_{j,j})\log(\frac{p+(1-p)\alpha_{j,j}}{h_j})\Big)$$

$$\geq^2 \frac{n-1}{n}\sum_i \frac{(1-p)(1-\alpha_{i,i})}{n-1}\log(\frac{(1-p)(1-\alpha_{i,i})}{n-1})$$

$$+ (p + \frac{\sum_j(1-p)\cdot\alpha_{j,j}}{n})\cdot\log(p+\frac{\sum_j(1-p)\cdot\alpha_{j,j}}{n})$$

$$\geq^3 p\cdot\log(p) + (1-p)\cdot\log(\frac{1-p}{n-1})$$

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

Finally, note that applying the level strategy from Definition 4 to term 1 yields term 3. Hence, the level strategy minimises information leakage. When $p{=}\frac{1}{n}$, the level strategy leads to zero information leakage, as proven in Theorem 2. □

Now, we have found the worst-case attack strategies for advisors with honesty degree $p \in [0, 1]$. Specifically, when $p < \frac{1}{n}$, the strategy requires a dishonest advisor to report the truth sometimes. When $p \geq \frac{1}{n}$, the strategy requires the advisor to uniformly choose a dishonest rating. Moreover, ultimate attack with zero information leakage can only be achieved when $p \leq \frac{1}{n}$. An advisee can still get some information about the truth when $p > \frac{1}{n}$.

To illustrate our results, we plot the information leakage of O in the worst-case attacks, as a variable of p (with n=3 and n=10) and n (with p=0.2 and p=0.8), in Figure 2. The figure shows that when $p \leq \frac{1}{n}$ or $n \leq \frac{1}{p} - 1$, the information leakage is zero. And when the difference between p and $\frac{1}{n}$ increases, the information leakage increases.

We have validated our information-theory based definition of the worst-case unfair rating attacks in [6] (see "Robustness Analysis" section). Three popular trust models BLADE [19], TRAVOS [33] and MET [24] were used. We presented that even when correct p values are provided, these models show poor accuracy in trust evaluation under the worst-case attacks. This is in line with our theoretical results: that when no information exists in ratings (ultimate attacks), the truth cannot be deduced, and when there exists minimal information (other worst-case attacks), the truth may be derived if the strategies are known (see ITC in [6]). In this paper, we do not present these simulations, as we want to focus on our new studies in Sections V and VI.

For attacks with non-zero amount of information, there does not exist straightforward relation between the amount of information leakage and the accuracy of trust evaluation, or other types of decision making. This means that more information leakage does not necessarily leads to more accurate trust evaluation. Different models or mechanisms may react differently to the same attack or attacks with the same amount of information. For example, bad-mouthing ratings are oppositely related to the truth (see their information leakage in [34]). They are filtered in some approaches, but are made use of in some others (see Section II). Models that are designed against some specific types of attacks may show lower accuracy under some other attacks, even if the latter have more information leakage. Therefore, we do not build explicit connection between the amount of information leakage and the accuracy of decision making, but observe the fact that information leakage puts an upper bound on the accuracy that is achievable. The worst-case attacks result in the tightest upper bound for accuracy. Hence, we argue that the worst-case attacks should be considered when designing a secure system.

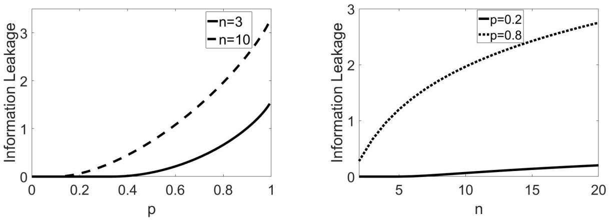


Fig. 2. The minimal information leakage of O varies with p and n ($n > 1$).

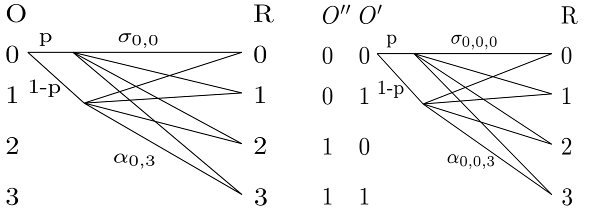
V. Quantifying Attacks under Subjective Rating

In reality, the deviation of ratings from the observed facts may not only come from dishonest intentions of advisors. Given the same observation, different honest advisors may also report different ratings. A very important reason here is that their opinions regarding the observation are subjective[7]. Subjectivity means "based on or influenced by personal feelings, tastes, or opinions" (Oxford Dictionary). Different advisors may have different subjectivity preferences. For example, they may put emphasis on different features, either when grading a target, or when suggesting an option. One honest user may rate a site unsafe due to excessive advertisements, whereas another honest user rates it safe, since it does not operate malware and delivers the promised functionality. Even if they emphasize the same features, they may have different expectations. In the example, when two honest users both take amount of advertisements as the criterion for safety, one user may find it excessive and rate it unsafe, whereas the other may find it acceptable and rate it safe.

Our interest in subjectivity is about to what extent honest ratings determine how the advisee would observe the truth. In the extreme case of the objective rating scenario from Section IV, an honest rating determines the truth completely. For example, if an honest advisor reports software as malware, then it is malware, and the advisee considers it as malware. In the subjective scenarios, there is no one-to-one link between the two. For example, a hotel room can be "good enough" or "not good enough" to an advisee, but advisors provide ratings in the form of 1 to 5 stars. In this example, a low star rating likely implies "not good enough" whereas a high star rating likely implies "good enough".

Subjectivity may reduce the usefulness of ratings. For example, a positive honest rating about a clean hotel in a bad neighbourhood may be found misleading by a user that cares about location rather than cleanliness. Both subjective ratings and dishonest ratings introduce inaccuracy. Some researchers treat them the same without differentiating the underlying motivation of advisors [13], [9], while some others orthogonally study them by distinguishing subjective advisors from the dishonest ones [28]. However, it was an open question whether (and if yes, how) subjectivity influences the effects of unfair rating attacks. We formally study this issue in this section.

[7] Discrimination by a target (where the target acts differently to different advisors) may also cause diverse opinions, but discrimination would result in different observations.



(a) Subjective rating model.

(b) Subject rating model distinguishing features (Sec. VI-A).

Fig. 3. Examples of subjective rating with 4 rating options.

A. Modelling Subjective Rating

Given an observation, various ratings may be chosen by an honest but subjective advisor, not only the rating that equals the observation. To include subjectivity, the proper objective rating model (e.g., depicted in Figure 1(b)) needs to be improved.

We still use notations $p, \alpha, O, R, H, \neg H$ with the same meaning as in the objective rating model. However, here we allow ratings to have different amount of options as the observations, with the outcomes of O/R being $\{0, \cdots, n_o{-}1\}$ / $\{0, \cdots, n_r{-}1\}$. We introduce an $n_o{\times}n_r$ matrix σ to characterize the rating behaviour of an honest advisor, with $\sigma_{d,r}{=}P(R{=}r|O{=}d, H)$. Subscript d, r also denote the row and column index of the entry $\sigma_{d,r}$, with both starting from 0. In the objective case, σ would be an identity matrix. The probability of receiving r when the truth is o is $P(r|o) = P(r|o, H) \cdot P(H|o) + P(r|o, \neg H) \cdot P(\neg H|o) = \sigma_{o,r} \cdot p + \alpha_{o,r} \cdot (1-p)$. The matrix μ is defined as a shorthand notation for $P(R|O)$, with $\mu_{o,r} = \sigma_{o,r} \cdot p + \alpha_{o,r} \cdot (1-p)$. The prior distribution of O is assumed to be uniform, with $H(O){=}\log(n_o)$.

We formally define subjective rating as follows:

Definition 5. σ-subjective rating *is a rating function with* $f_\sigma(o, \alpha)(r) = p \cdot \sigma_{o,r} + (1{-}p) \cdot \alpha_{o,r}$, $o{\in}\{0, \ldots, n_o{-}1\}, r{\in}\{0, \ldots, n_r{-}1\}$.

Function f_σ defines how an advisor's attributes p, σ, α decide its behaviour – the link between R and O. Note that the objective rating model in Figure 1(b) is a special case of f_σ: σ is an identity matrix I. We refer to it as f_I, which is an objective rating function.

A rating model for f_σ is presented in Figure 3(a). There are four outcomes of O. R also has four options. $\sigma_{0,0}$ denotes the conditional probability of an honest advisor reporting 0 when 0 is the observed fact.

In Figure 4, we depict 4 examples of σ matrices that define the rating behaviour of honest advisors: $\sigma_b, \sigma_c, \sigma_d$ and σ_e. Recall that rows are values for O, and columns are values for R. A value for O corresponds to the ground truth in the objective rating model, but in the subjective rating model, it corresponds to how the advisee would view the truth. Also differently from the objective rating model, there is no one-to-one correspondence between R and O e.g., $r = 0$ may correspond with both $o{=}1$ and $o{=}2$, and also they may potentially mean different things . In our example, the two are equinumerous.

$$\sigma_b = \begin{vmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{vmatrix} \qquad \sigma_c = \begin{vmatrix} 0.97 & 0.02 & 0.01 \\ 0.04 & 0.96 & 0.00 \\ 0.01 & 0.03 & 0.96 \end{vmatrix}$$

$$\sigma_d = \begin{vmatrix} 0.01 & 0.98 & 0.01 \\ 0.01 & 0.98 & 0.01 \\ 0.01 & 0.98 & 0.01 \end{vmatrix} \qquad \sigma_e = \begin{vmatrix} 0.70 & 0.20 & 0.10 \\ 0.50 & 0.40 & 0.10 \\ 0.25 & 0.05 & 0.70 \end{vmatrix}$$

Fig. 4. Several matrixes as examples

Matrix σ_b is the identity matrix, denoting objective rating. Ratings and opinions correspond in σ_b. Even if they do not, as long as an honest rating (R) completely determines what the opinion of the advisee would be (O) e.g., $r{=}2$ determines $o{=}2$, the rating is actually objective. Matrix σ_c is close to matrix σ_b, signifying low subjectivity. Values of R determine values of O to some degree, since for a given value of R, there is one highly probable value for O (and two improbable alternatives).

For matrix σ_d, the honest ratings do not determine the would-be opinion of the advisee, since R probably equals 1, in which case O is equally probably as 0, 1 or 2. Matrix σ_d is considered to be highly subjective. A possible alternative view would be to look at the extent to which the advisee's view on the truth (O) determines honest ratings (R), instead. Matrices σ_b and σ_c remain (nearly) objective in this view. Matrix σ_d could naively be considered less subjective than σ_c, since the value of R is determined to probably be a specific value given an o. However, the value of R cannot be determined by O or vice versa, as they are independent ($\forall o, r, p(r){=}p(r|o)$). To measure the extent to which the would-be opinion determines the honest rating, we would have to normalise the probability by dividing by the prior probability of the rating. Note that since $p(o|r) = \frac{p(r|o)}{n_o p(r)}$, meaning the conditional probabilities $p(o|r)$ for different o given a r are determined by the probabilities in a column $p(r|o)$, these two views actually coincide.

Matrices σ_b, σ_c and σ_d are unlikely to be the actual matrices of advisors, as they are extreme cases. Matrix σ_e is a more realistic example of a (highly) subjective rating matrix. If the value of R is i, then the most probable value for O is also i. Notice that the reverse is not true, since the advisor is most likely to rate $R = 0$ whenever the advisee would have opinion $O = 1$. Finally, the rating $R = 1$ weakly determines that O is probably 1, but it strongly determines that O probably is not 2. In the model we define in Section V-D, we take this into account when defining a partial order of subjectivity.

B. Information leakage

Use Definition 3 in Section III, we can compute information leakage of subjective rating f_σ:

Proposition 2. *Given strategy* α*, the information leakage of rating* f_σ *is:*

$$I(O; R) = \frac{1}{n_o} \sum_{o,r} \mu_{o,r} \log \frac{n_o \mu_{o,r}}{\sum_o \mu_{o,r}}$$

Given fixed attack strategy α, if we change σ (i.e. the model of the subjectivity of honest advisors), then the information leakage typically changes. The amount of information leakage in ratings reflects the impact (or its effect on the system) of the attack. Proposition 2 shows how subjectivity of honest advisors influences the impact of attacks.

Conversely, given fixed subjectivity σ, is it possible for attackers to find a strategy α such that no information is leaked – ultimate attacks or the worst case for an advisee? And if its possible, will there/and what would be conditions for the values of p, n_o, n_r, σ. We study these questions in Section V-C. In particular, we will investigate whether subjectivity changes the conditions for ultimate attacks compared to objective rating. In Section V-D we quantitatively study the relationship between the degree of subjectivity and the amount of information leakage – quantitative robustness comparison.

C. Ultimate Attacks

According to the definition in Section IV, ultimate attacks mean that there is zero information leakage of the observed fact O. No matter how sophisticated a system can be, the ratings are completely useless under ultimate attacks. Fortunately, the circumstances wherein an attacker can perform an ultimate attack are rare. Yet, for some settings it is rarer than others. Hence, we can use the difficulty to perform an ultimate attack as a proxy for the robustness of a system.

We can select α to get 0 information leakage for some values of p, σ, formally:

Theorem 5. *$\forall o, r$, let $p{<}1$ and $\sigma_{o^*,r} = \max_o \sigma_{o,r}$. There exists an attack strategy α, such that information leakage in Proposition 2 equals 0 iff $p \leq \frac{1}{\sum_r \sigma_{o^*,r}}$.*

Proof. Subscript o^*, r denotes the row and column index of the maximal element in column r of σ. Let $\alpha_{o^*,r}$ ($\mu_{o^*,r}$) denotes the entry in α (μ) which has the same index. The information leakage is zero iff O and R are independent, which holds iff given an arbitrary r, $P(r|o)$ equal for any o, including o^*. This means $p\sigma_{o,r}{+}(1{-}p)\alpha_{o,r}{=}p\sigma_{o^*,r}{+}(1{-}p)\alpha_{o^*,r}$, which can be rewritten as $\alpha_{o,r}{-}\alpha_{o^*,r}{=}\frac{p}{1-p}(\sigma_{o^*,r}{-}\sigma_{o,r})$. For "only if", remember that $\forall o, \sum_r \alpha_{o,r}{=}1$, hence $\frac{p}{1-p}\sum_r(\sigma_{o^*,r}{-}\sigma_{o,r}){+}\sum_r \alpha_{o^*,r}{=}1$. Note that $\sum_r \alpha_{o^*,r}{\geq}0$ and also $\sum_r(\sigma_{o^*,r}{-}\sigma_{o,r}){=}\sum_r \sigma_{o^*,r}{-}1$. Hence we get $\frac{p}{1-p}{\cdot}(\sum_r \sigma_{o^*,r}{-}1){\leq}1$ and $p{\leq}\frac{1}{\sum_r \sigma_{o^*,r}}$. For "if", we can simply set $\alpha_{o,r}{-}\alpha_{o^*,r}{=}\frac{p}{1-p}(\sigma_{o^*,r}{-}\sigma_{o,r})$. □

To better illustrate Theorem 5, take σ_e in Figure 4 as an example. The maximal entry in each column is $0.7, 0.4, 0.7$ respectively. Suppose $p{=}0.5{\leq}\frac{1}{0.7+0.4+0.7}$, then there exists an ultimate attack α. From the proof for the theorem, we have $\sum_r \alpha_{o^*,r}{=}1{-}\frac{p}{1-p}\sum_r(\sigma_{o^*,r}{-}\sigma_{o,r}){=}2{-}\sum_r \sigma_{o^*,r}{=}0.2$. W.l.o.g, Let $\alpha_{0,0}{=}0.2$ and as a result $\alpha_{1,1}{=}\alpha_{2,2}{=}0$. Based on $\alpha_{o,r}{-}\alpha_{o^*,r}{=}\frac{p}{1-p}(\sigma_{o^*,r}{-}\sigma_{o,r})$, we can get all entries of α, e.g., $\alpha_{0,2}{=}\alpha_{1,2}{=}0.6, a_{2,0}{=}0.65$.

Theorem 5 proves that it is possible for attackers to completely hide information if there are enough of them ($1 - p{\geq}1{-}\frac{1}{\sum_r \sigma_{o^*,r}}$), and the corresponding strategy depends on the values of p, σ. There already exist methods to distinguish subjectivity preferences of honest advisors, e.g., by clustering (as in [28] and [29]). Membership of a cluster determines how probable certain subjective behaviours are. The parameter σ characterises how probable certain actions are, given the context. In order not to underestimate the power of the attacker, we must assume that the attacker also has access to p and σ. Furthermore, it is likely that the attacker can arrive at the same result for p and σ, e.g. by performing the same computations – assuming ratings are public knowledge.

It is obvious that $\frac{1}{\sum_r \sigma_{o^*,r}} \geq \frac{1}{n_r}$, with equality only if all $\sigma_{o^*,r}$ equal 1. The identity matrix has all maximal elements equal to 1, and thus, it is easy to see that this generalises the results from Section IV. Even if there are not enough attackers to perform an ultimate attack on an objective-rating system ($\frac{1}{n_r}{<}p$), there may be sufficiently many attackers to do so on a system with subjectivity, namely when $\frac{1}{n_r} < p \leq \frac{1}{\sum_r \sigma_{o^*,r}}$. The introduction of subjectivity makes it easier for attackers to completely hide information, thus, leaving a system less robust.

Further, o^* in $\sigma_{o^*,r}$ denotes the observation under which reporting r is the most probable. The value of $\sigma_{o^*,r}$ reflects the subjectivity difference behind reporting r. The smaller $\sigma_{o^*,r}$ is, the more uniformly the values of $\sigma_{o,r}$ are distributed over o, and intuitively, the more probable that multiple o are reported as the observation behind r, which indicates more subjectivity difference. Theorem 5 implies that, with more subjectivity difference, the amount of attackers necessary in the population (from which we select our advisor) to make a rating completely useless decreases.

We can compare the examples from Figure 4, and compute what the probability needs to be that an advisor is an attacker, in order for the attacker to be able to perform the ultimate attack. For σ_b, it is $p \leq \frac{1}{1+1+1} \approx 0.333$; for σ_c, $p \leq \frac{1}{0.97+0.96+0.96} \approx 0.346$; for σ_d, $p \leq \frac{1}{0.01+0.98+0.01} = 1$; and for σ_e, $p \leq \frac{1}{0.7+0.4+0.7} \approx 0.556$. In the case for σ_d, as O and R are independent, attackers are even unnecessary to get 0 information.

D. Quantitative Robustness Comparison

Except ultimate attacks, we also want to investigate how the impact of attacks changes with the increase/decrease of the degree of subjectivity. First, we create an ordering of subjectivity, to be able to say that one advisor is more subjective than the other. Our ordering is not complete, so when an advisor is more subjective in one aspect than another advisor, but less so for another aspect, then the two advisors may be incomparable.

We define the ordering on matrices σ, which describe subjective rating behaviour, and take the natural extension to rating functions: $\sigma \preceq \sigma'$ iff $f_\sigma \preceq f_{\sigma'}$. There are some notions that any reasonable subjectivity ordering of matrices must have: 1) For $\sigma_b, \sigma_c, \sigma_d, \sigma_e$ from Figure 4: $\sigma_b \preceq \sigma_c \preceq \sigma_e \preceq \sigma_d$. 2) The relation must be reflexive and transitive (i.e. subjectivity is a preorder). No anti-symmetry, since two different matrices may be equally subjective. 3) An objective matrix I_o[8] is an maximal element (i.e. $\forall_\sigma I \preceq \sigma$): reporting with the identity matrix is objective reporting. 4) The uniform matrix U is a supremum (i.e. $\forall_\sigma \sigma \preceq U$): rating and observation are independent; honest ratings are unrelated to the truth.

[8] A matrix where every row and column has a single element equal to 1.

The definition of subjectivity assumes a ranking (π_i) of which ratings are more appropriate for which observations, in which case the more objective scenario should assign more probability to more appropriate ratings. It may be the case that some ratings are more likely to be provided a priori, which can be corrected for by normalising the terms by dividing by the prior probability. The prior probability is proportional to $\sum_i \sigma_{i,j}$, which we denote as σ_j. Formally:

Definition 6. *Let σ and σ' be $n \times n$ subjectivity matrices. A row σ_j is* less subjective *than σ'_j, denoted $\sigma_j \preccurlyeq \sigma'_j$ if there exist permutation π_j, s.t. $\pi(\sigma)_{i,j}$ and $\pi(\sigma')_{i,j}$ are non-increasing over i, and $\forall_{0 \leq k < n} \sum_{0 \leq j < k} \frac{\pi(\sigma)_{i,j}}{\sigma_j} \geq \frac{\pi(\sigma')_{i,j}}{\sigma'_j}$.*
Then, a subjectivity matrix σ is less subjective *than σ', denoted $\sigma \preceq \sigma'$ when for all i, $\sigma_i \preccurlyeq \sigma'_i$.*

The hard requirements for the Definition can be straightforwardly proven, bearing in mind that normalised majorisation is itself transitive, reflexive, has $(1, 0, 0, \ldots, 0)$ as supremum and $({}^1\!/\!{}_n, \ldots, {}^1\!/\!{}_n)$ as infimum.

Proposition 3. *The relationship $\preceq$ is reflexive and transitive, and for all σ, $I \preceq s \preceq U$.*

We rely on the observation that sometimes probability mass can move from one value to another, whilst decreasing information leakage:

Lemma 1. *Given $\mu_{i,j} \geq \frac{\mu_j}{n}$ and $\mu_{i,j'} \leq \frac{\mu'_j}{n}$, we can define μ^* equal to μ except at i, j and i, j' where $\mu_{i,j} \geq \mu^*_{i,j} \geq \frac{\mu^*_j}{n}$ and $\mu_{i,j'} \leq \mu^*_{i,j'} \leq \frac{\mu^*_{j'}}{n}$. Then $I(O; R_\mu) \geq I(O; R_{\mu^*})$.*

Proof. Jensens' Inequality (Thm 1) gives: $\sum_i \frac{\mu_{i,j}}{n} \log \frac{\mu_{i,j}}{\mu_j}$ $\geq \sum_i \frac{\mu^*_{i,j}}{n} \log \frac{\mu^*_{i,j}}{\mu^*_j}$ and $\sum_i \frac{\mu_{i,j'}}{n} \log \frac{\mu_{i,j'}}{\mu_j} \geq \sum_i \frac{\mu^*_{i,j'}}{n} \log \frac{\mu^*_{i,j'}}{\mu^*_{j'}}$. For other $j^\dagger$, $\mu_{i,j^\dagger} = \mu^*_{i,j^\dagger}$, so the remaining sums are: $\sum_i \frac{\mu_{i,j^\dagger}}{n} \log \frac{\mu_{i,j^\dagger}}{\mu^\dagger_j} = \sum_i \frac{\mu^*_{i,j^\dagger}}{n} \log \frac{\mu^*_{i,j^\dagger}}{\mu^*_{j^\dagger}}$. □

Using the order of subjectivity of reporting, we can formalise the notion that increasingly subjective reporting makes it easier for an attacker to decrease information leakage:

Theorem 6. *For any ratings $f \preceq f'$, for any $R_\mu = f(O, \alpha)$, there $\exists \alpha'$ such that $R_{\mu'} = f'(O, \alpha')$ with $I(O; R_\mu) \geq I(O; R_{\mu'})$.*

Proof. The ranking of the rows π_i from Def 5 may not rank the values of μ. Wlog, there exists α^*, s.t. $\mu^* = p\sigma + (1-p)\alpha^*$ where π_i ranks the values of $\frac{\mu^*_{i,j}}{\mu^*_j}$ and $I(O; R_\mu) \geq I(O; R_{\mu^*})$. The α^* can be obtained by applying Lemma 1 to move probability from overly high ranked values to overly low ranked values; since σ follows the ranking, the resulting μ^* has the property that $\mu^* - p\sigma \geq 0$ and thus $\alpha^* = \frac{\mu^* - p\sigma}{1-p} \geq 0$.

Remains to prove there exists μ' s.t. $I(O; R_{\mu^*}) \geq I(O; R_{\mu'})$. Let $\alpha^\dagger = \frac{\mu^* - p\sigma'}{1-p}$. Some values $\alpha^\dagger_{i,j}$ may be negative (meaning $\alpha^\dagger$ is invalid). For negative $\alpha^\dagger_{i,j}$, note that since $\alpha^* = \frac{\mu^*_{i,j} - p\sigma_{i,j}}{1-p}$ is non-negative, $\sigma'_{i,j} > \sigma_{i,j}$. Due the majorisation property, $\sum_{1 \leq k < \pi_i(j)} \frac{\sigma_{i,k}}{\sigma_k} > \sum_{1 \leq k < \pi_i(j)} \frac{\sigma'_{i,k}}{\sigma_k}$, and since $\frac{\mu^*_{i,j}}{\mu^*_j}$ respects the ranking, $\sum_{1 \leq k < \pi_i(j)} \frac{\mu_{i,k}}{\mu_k} > \sum_{1 \leq k < \pi_i(j)} \frac{p\sigma'_{i,k}}{p\sigma'_k}$. Therefore, there exists j' s.t. $\pi_i(j') < \pi_i(j)$, and $\alpha^\dagger_{i,j'} > 0$, so we can apply Lemma 1 to increase $\mu^\dagger_{i,j}$ (and decrease $\mu^\dagger_{i,j'}$), whenever $\alpha^\dagger_{i,j} < 0$. Hence, negative values can be eliminated by applying Lemma 1, preserving $I(O; R_\mu) \geq I(O; R_{\mu'})$. □

$$\begin{vmatrix} .485 & .010 & .505 \\ .020 & .480 & .500 \\ .015 & .005 & .98 \end{vmatrix} \quad \begin{vmatrix} .485 & .260 & .255 \\ .270 & .480 & .250 \\ .015 & .005 & .980 \end{vmatrix} \quad \begin{vmatrix} .485 & .260 & .255 \\ .270 & .480 & .250 \\ .125 & .025 & .850 \end{vmatrix}$$

μ (IL: 0.478) $\quad$ μ^* (IL: 0.441) $\quad$ μ' (IL: 0.293)

Fig. 5. Three relevant matrices for applying Theorem 6 to σ_c and σ_e.

To illustrate the Theorem, say we want to prove that attacks on σ_e can have more impact than on σ_c. So, given an arbitrary attack on σ_c, there is a worse (or equal) attack on σ_e. Take, for example, the attack where the attacker always rates option 2, so $\alpha_{i,2} = 1$. Let $p = {}^1\!/\!{}_2$. Taking $\mu = p\sigma_c + (1-p)\alpha$ (depicted in Figure 5), we see that entries in μ are ranked differently than in σ_c. By moving mass away from column 3 to columns 1 and 2 where appropriate, we obtain μ^* (with lower information leakage, due to Lemma 1). Note that $\mu^*_{i,j} > p \cdot (\sigma_c)_{i,j}$, so α^* is well-defined, but $\alpha^\dagger$ is not, since $\mu^*_{2,0} < p \cdot (\sigma_e)_{2,0}$ and $\mu^*_{2,1} < p \cdot (\sigma_e)_{2,1}$. Observe that μ' is the result of moving probability mass from $2, 2$ to $2, 0$ and $2, 1$, and that $\mu'_{2,0} = p \cdot (\sigma_e)_{2,0}$, and $\alpha' \geq 0$. Again, information leakage decreased.

VI. Robustness of Existing Approaches to Deal with Subjectivity

We consider two types of ways proposed by system designers to deal with subjectivity: feature-based rating, which is popularly applied in reality to help resolve conflicting emphasis on features in overall rating, and clustering advisors, which is proposed in literature to distinguish advisors with different subjectivity. These approaches aim to mitigate the influence of subjectivity, so it is interesting to study whether they improve the robustness against unfair rating attacks.

A. Feature-based rating

Feature-based rating refers to settings where advisors need to rate each feature of an observation (or a target) instead of providing an overall rating. For example, in Booking.com or Expedia.com, consumers can score over multiple features of a hotel after their accommodation, such as cleanliness, comfort, location, facilities, staff and value for money. Compared with overall rating, distinguishing features helps avoid subjective ratings induced by different emphasis on features. Also, if all ratings are honest, potential consumers are presented a more comprehensive view of the hotel.

The rating function f_σ modelled in Figure 3(a) does not distinguish features, and we name it as overall subjective rating function. The modelling of a feature-based rating can be directly derived from f_σ. To better illustrate this, we first reformulate the modelling of f_σ. There, O represents an observation regarding a target, which contains all related features of the observation that advisors care. We can split all features into two groups regarding whether the advisee cares, with O' and O'' representing the group of features that the advisee cares and does not care respectively.We assume that cared features and uncared features are independent, i.e., $O' \perp\!\!\!\perp O''$. Outcomes of O' (or O'') are the true values of the corresponding features, e.g, the score of a hotel's location. The total number of outcomes of O' (O'') is denoted as $n_{O'}$ ($n_{O''}$). As there are three variables in rating, 2D behaviour matrices σ, α in overall rating now become 3D matrices with dimensions $n_{O'} \times n_{O''} \times n_r$. Figure 3(b) presents the reformulated subjective rating model. Specifically, $\sigma_{0,1,3}$ means the probability that an honest advisor reports 3 when he observes 0 for O' and 1 for O''.

The subjective rating behaviour regarding features O' can be characterized by conditional probabilities $p(r|o')$. Let $\sigma_{o',r} = \frac{1}{n_{O''}} \sum_{O''} \sigma_{o',o'',r}$ and $\alpha_{o',r} = \frac{1}{n''_O} \sum_{O''} \alpha_{o',o'',r}$. Then, $p(r|o') = \sum_{o''} p(r, o''|o') = p \cdot \sigma_{o',r} + (1-p) \cdot \alpha_{o',r}$. The information leakage of both O' and O'' is $I(O', O''; R)$. An advisee is interested in the information of O': $I(O'; R)$, which is determined by $\sigma_{o',r}$ and $\alpha_{o',r}$.

In a feature-based rating scenario, there are no uncared features. Hence, only variables O' and R remain in the reformulated rating model (see Figure 3(b)). In this way, the only difference between a feature-based rating and f_σ in Figure 3(a) is that variable O becomes O'. To be distinguished from $\sigma_{o',r}, \alpha_{o',r}$ in the reformulated model, we use $\varsigma_{o',r}$ and $\beta_{o',r}$ to characterize subjective and strategic behavior in a feature-based rating respectively. $\varsigma_{o',r}$ and $\beta_{o',r}$ determine the information leakage of O'. Theorem 5 still holds: the condition to achieve ultimate attacks in feature-based rating is $p < \frac{1}{\sum_r \varsigma_{o^*,r}}, \varsigma_{o^*,r} = \max_o \varsigma_{o,r}$.

According to the definition of information leakage, as long as $p(r|o')$ remains the same for given any o', which means $p \cdot \varsigma_{o',r} + (1-p) \cdot \beta_{o',r} = p \cdot \sigma_{o',r} + (1-p) \cdot \alpha_{o',r}$, the feature-based rating would have the same amount of information leakage compared with the overall rating. This implies that, when a given overall rating framework is changed to a corresponding feature-based rating framework, the information an advisee can learn may remain the same. For instance, if subjectivity differences over feature O' remain unchanged, which means $\forall o', r, \varsigma_{o',r} = \sigma_{o',r}$, then attackers can choose the same strategy $\beta_{o',r} = \alpha_{o',r}$ to leak the same amount of information. By choosing proper β, attackers may even cause less information leakage in feature-based rating. For ultimate attacks, if $\sum_r \varsigma_{o^*,r} < \sum_r \sigma_{o^*,r}$, the feature-based rating relaxes the condition on achieving ultimate attacks.

In summary, although help reduce subjectivity induced by different emphasis, feature-based rating does not necessarily improves robustness compared with overall rating. Intuitively, although subjectivity by different emphasis is reduced in feature-based rating, subjectivity induced by various expectation gets re-distributed as every advisor is forced to consider each feature separately. Hence, it is hard to judge whether subjectivity difference becomes less in feature-based rating.

B. Clustering Advisors

Thus far, we have taken an approach where we use a single matrix to model the subjectivity of all advisors. This is trivially sufficient when reasoning about a single advisor, as we did in Section V. There are two other cases where a single matrix is sufficient to model subjectivity for all advisors: The first case is the rather unrealistic case where all advisors have the same subjective preferences. The second case is where we cannot distinguish subjective preferences of different advisors.

While these three cases are interesting, they are insufficient. Generally, we have multiple advisors with different behaviour that we have some historical data about. Hence, in this section, we introduce a model that help reason advisors with different subjectivity matrices.

1) Modelling: To deal with the general case, we base our analysis on the popular clustering approach [27], [28]. Therein, advisors are assigned to clusters based on their (past) behaviour, and each cluster has their own behaviour model. Not only can the subjectivity matrices differ from cluster to cluster, we also allow p-value to differ, to model clustering based on degree of honesty. Finally, we assume that the attacker knows which cluster he is in. This implies that his strategy matrix is chosen to minimise total information leakage.

Assume there are k advisors in total, with R_i represents the rating of i^{th} advisor, and $\overline{R} = R_0, \ldots, R_{i-1}$. Let there be clusters $c', c^{\dagger}, \ldots,$ with symbols $', ^{\dagger}, ^{\S}$ denote association to a cluster. $\overline{R}'$ refer to ratings generated by all k' advisors in c'. Associated with cluster c' is: probability of its advisors' honesty p', subjectivity matrix σ' and strategy matrix α'. The random variable C_i dictates to which cluster the i^{th} advisor belongs, and $\overline{C} = C_0, \ldots, C_k$. In other words, $p(r_i|d, C_i = c') = p'\sigma' + (1 - p')\alpha' = \mu'$. We use c'_j to mean j^{th} advisor in cluster c'. Clustering is typically based on previous behaviour of the advisors, and thus not related to what the observed facts are; so $\overline{C}$ is independent of O.

2) Robustness of clustering: The crucial question is whether clustering increases the robustness of the system, which is the equivalent as whether clustering increases the expected information leakage. Clustering indeed increases expected information leakage, except in the case where C has no impact on the relationship between O and R:

Theorem 7. *$I(O; \overline{R}|\overline{C}) \geq I(O; \overline{R})$, with equality iff C and O are conditionally independent under R.*

Proof. First, note: $I(O; \overline{R}|\overline{C}) = H(O|\overline{C}) - H(O|\overline{R}, \overline{C})$, and $I(O; \overline{R}) = H(O) - H(O|\overline{R})$. Since O and $\overline{C}$ are independent, it suffices to prove $H(O|\overline{R}, \overline{C}) \leq H(O|\overline{R})$. This is a known property of conditional entropy, and equality holds only if C and O are conditionally independent under R. □

Therefore, clustering increases the expected information leakage. No matter what the attacker's behaviour is, more information is expected after clustering. However, we strongly conjecture that the benefits of clustering are even greater: Clustering always outperforms not clustering.

A *naive* interpretation of the conjecture is that for all $\overline{C}$, $I(O; \overline{R}|\overline{C} = \overline{c}) \geq I(O; \overline{R})$. Not only does Theorem 7 not suffice to prove this version, it turns out to be false. As a counter example, assume for one cluster $c^{\S}$ its $\sigma^{\S}$-matrix has no information leakage – its a cluster where useless advisors are put in. There is a non-zero probability that all advisors are put in this matrix, namely when the user is unfortunate enough only to select useless advisors. In this case, $I(O;\overline{R}|\overline{C}{=}\overline{c}){=}0 < I(O;\overline{R})$. In the remainder of this section, we will nuance our claim that clustering always outperforms not clustering.

Intuitively, the reason the counterexample fails, is that moving an advisor from one cluster to another changes the expected behaviour of a randomly chosen advisor. It only makes sense to compare a given clustering to a situation where the expectation of the behaviour is the same. We introduce the *universal cluster* for that purpose. The universal cluster, $c^{\S}$, has the desired property that $p(o|r,c^{\S}){=}\sum_i \frac{p(o|r_i,c_i)}{k}$. Specifically, let $p^{\S}{=}\frac{p'k'}{k}+\frac{p^{\dagger}k^{\dagger}}{k}+\dots$ – so the universal p-value is simply the averaged p-value. And, $\forall o,r$ let $\sigma^{\S}_{o,r}{=}\frac{k'p'\sigma'_{o,r}}{k'p'}+\frac{k^{\dagger}p^{\dagger}\sigma^{\dagger}_{o,r}}{k^{\dagger}p^{\dagger}}+\dots$ – so each cell in the universal behaviour matrix is a weighted average of the corresponding cells in the matrices of the clusters. Similarly, $\forall o,r$ let $\alpha^{\S}_{o,r}{=}\frac{k'(1-p')\alpha'_{o,r}}{k'(1-p')}+\frac{k^{\dagger}(1-p^{\dagger})\alpha^{\dagger}_{f,d,r}}{k^{\dagger}(1-p^{\dagger})}+\dots$. Now, $\alpha^{\S}$ may not minimise the information leakage for the given p and σ. Let $\alpha^{\$}$ be the attacker behaviour that minimises information leakage. Then let $c^{\$}$ be the minimal universal cluster, where $p^{\$}{=}p^{\S}$, $\sigma^{\$}{=}\sigma^{\S}$, but $\alpha^{\$}$ is the matrix that minimises $I(O;\overline{R}|\overline{c}^{\$})$

Using the new terminology, we can express our conjecture as $I(O;\overline{R}|\overline{c}) \geq I(O;\overline{R}|\overline{c}^{\$})$. Note that it suffices to prove a simpler inequality to prove the conjecture:

Proposition 4. *Let $\overline{c}$ be $c'_0,\dots c'_{k'-1},c^{\dagger}_{k'},\dots,c^{\dagger}_{k'+k^{\dagger}}$, and $\overline{c^{\$}}$ its minimised universal cluster. If $H(O|\overline{R},\overline{c}) \leq H(O|\overline{R},\overline{c}^{\S})$ then for any collection of clusters $\overline{b}$, $I(O;\overline{R}|\overline{b}){\geq}I(O;\overline{R}|\overline{b}^{\$})$.*

Proof. First, note that it suffices to prove for $c^{\S}$, since $I(O;\overline{R}|\overline{c}^{\$}) \leq I(O;\overline{R}|\overline{c}^{\S})$ by definition. Second, without loss of generality, we reorder the list of advisors such that each cluster's advisors occur consecutively. Then, if we can prove the proposition for two clusters, we can replace any pair of clusters by its universal cluster, and inductively apply the proposition. Remains to prove the proposition for two clusters: If $H(O|\overline{R},\overline{c}) \leq H(O|\overline{R},\overline{c}^{\S})$, then $H(O) - H(O|\overline{R},\overline{c}) \geq H(O) - H(O|\overline{R},\overline{c}^{\S})$. Due to independence of O and C, $H(O|\overline{c}){-}H(O|\overline{R},\overline{c}) \geq H(O|\overline{c}^{\S}){-}H(O|\overline{R},\overline{c}^{\S})$, which suffices to prove the proposition. □

3) Whether to exclude clusters: We discussed the robustness of clustering advisors. There exist various ways to deal with clusters. Some researchers choose to exclude clusters where the advisors are considered dishonest [29], while some others propose to learn from clusters where advisors are even strategic [28]. We study the impact of these different ways on robustness.

Theorem 8 states that if a cluster provides no information about O, then it does not impact the correlation between O and other clusters.

Theorem 8. *If $\overline{R}'$ is independent of O, then $I(\overline{R}^{\dagger};O|\overline{R}') = I(\overline{R}^{\dagger};O)$, also $I(\overline{R}^{\dagger};O|\overline{R}',\overline{R}^{\S}) = I(\overline{R}^{\dagger};O|\overline{R}^{\S})$.*

Proof. $I(\overline{R}^{\dagger};O|\overline{R}') = \sum_{\overline{r}'} p(\overline{r}')\cdot I(\overline{R}^{\dagger};O|\overline{r}')$

$$= \sum_{\overline{r}'}\sum_{\overline{r}^{\dagger}}\sum_{o} p(o)\cdot p(\overline{r}^{\dagger},\overline{r}'|o)\cdot \log\left(p(\overline{r}^{\dagger}|o) - p(\overline{r}^{\dagger}|\overline{r}')\right)$$

$$=^1 \sum_{\overline{r}'}\sum_{\overline{r}^{\dagger}}\sum_{o} p(o)\cdot p(\overline{r}^{\dagger}|d)p(\overline{r}')\cdot \log\left(p(\overline{r}^{\dagger}|d) - p(\overline{r}^{\dagger})\right)$$

$$= I(\overline{r}^{\dagger};O)$$

Equality 1 follows due to the independence between $\overline{R}'$ and O, and also conditional independence between $\overline{R}'$ and $\overline{R}^{\dagger}$ given O. The second equality in the theorem can be proved in the similar way. □

The independence between $\overline{R}'$ and O means $I(\overline{R}'|O){=}0$. Following immediately from Theorem 8, we get Corollary 1.

Corollary 1. *If $I(\overline{R}';O){=}0$, then $I(\overline{R}',\overline{R}^{\dagger},\overline{R}^{\S};O) = I(\overline{R}^{\dagger},\overline{R}^{\S};O)$.*

Proof. Based on the chain rule of mutual information, $I(\overline{R}',\overline{R}^{\dagger},\overline{R}^{\S};O){=}I(\overline{R}';O){+}I(\overline{R}^{\dagger};O|\overline{R}'){+}(\overline{R}^{\S};O|\overline{R}',\overline{R}^{\dagger})$. Considering $I(\overline{R}';O){=}0$ and Theorem 8, the corollary can be easily proved. □

Corollary 1 implies if a cluster has no information about O, then it can be completely excluded without making a user lose any information. Remember the conditions for a cluster to have 0 information leakage is that the probability of advisors being honest, p, needs to be below a threshold (Theorem 3 or 5). The value p is a parameter of advisors. However, the value is not obvious to an advisee, and clustering mechanisms may not estimate p accurately. If a cluster is considered dishonest and gets excluded, when its true p is above the threshold, then a user loses useful information. Similarly, if a cluster is considered very reliable, when its p is actually below the threshold, then an advisee gets no information.

VII. Conclusion

In this paper, we proposed a quantitative measurement of unfair rating attacks based on information theory. How much information ratings provide about the truth determines the impact of attacks. We studied the scenario that an arbitrary advisor is rating a given subject. And we found the worst-case strategies against a user – meaning causing the minimal information leakage – that an individual attacker can undertake in his rating.

We first considered the scenario where rating is assumed to be objective. A probabilistic rating model was built to reason about possible rating behaviour of an arbitrary advisor who can have any degrees of honesty. We found that if we select an advisor randomly from a population, that advisor can hide the truth completely (perform the ultimate attack), if the population contains at least $n-1$ times more attackers than honest users – where n is the number of rating options. And some of them need to report the truth. Otherwise, the truth can still be learned from the ratings even if more than half of the advisors are strategic.

Considering that subjective rating is typically unavoidable in reality, we then improved the proposed rating model in several aspects: 1) given an observation, we allow honest advisors to choose different ratings; 2) we allow the options of ratings to be different from the options of observations; 3) we distinguish features of an observation that an advisee cares and does not care about.

We found that the introduction of subjectivity makes it easier for attackers to completely hide the truth. we also introduced an ordering of subjectivity, and found that more subjective rating makes a system less robust against unfair rating attacks. Since subjectivity decreases robustness, we studied whether existing methods of mitigating subjectivity difference would improve robustness. Splitting ratings up, such that individual features are rated does may mitigate or exacerbate the problem. Clustering advisors with similar subjective preference, however, improves robustness. Clusters with sufficiently many attackers may be ignored or blocked without consequence.

References

[1] C. Dellarocas, "The digitization of word-of-mouth: Promise and challenges of online reputation systems," *Management Science*, vol. 49, no. 10, pp. 1407–1424, 2003.

[2] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 17, no. 6, pp. 734–749, 2005.

[3] M. Li, X. Sun, H. Wang, Y. Zhang, and J. Zhang, "Privacy-aware access control with trust management in web service," *World Wide Web*, vol. 14, no. 4, pp. 407–430, 2011.

[4] S. Wang, Z. Zheng, Z. Wu, M. R. Lyu, and F. Yang, "Reputation measurement and malicious feedback rating prevention in web service recommendation systems," *IEEE Transactions on Services Computing*, vol. 8, no. 5, pp. 755–767, 2015.

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

[6] D. Wang, T. Muller, A. A. Irissappane, J. Zhang, and Y. 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)*, 2015, pp. 791–799.

[7] D. Wang, T. Muller, Y. Liu, and J. Zhang, "Towards robust and effective trust management for security: A survey," in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2014 IEEE 13th International Conference on*. IEEE, 2014, pp. 511–518.

[8] J. Zhang and R. Cohen, "Evaluating the trustworthiness of advice about seller agents in e-marketplaces: A personalized approach," *Electronic Commerce Research and Applications*, vol. 7, no. 3, pp. 330–340, 2008.

[9] J. Weng, Z. Shen, C. Miao, A. Goh, and C. Leung, "Credibility: How agents can handle unfair third-party testimonies in computational trust models," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, vol. 22, no. 9, pp. 1286–1298, 2010.

[10] S. Liu, A. C. Kot, C. Miao, and Y.-L. Theng, "A dempster-shafer theory based witness trustworthiness model to cope with unfair ratings in e-marketplace," in *Proceedings of the 14th Annual International Conference on Electronic Commerce*. ACM, 2012, pp. 99–106.

[11] S. Liu, J. Zhang, C. Miao, Y.-L. Theng, and A. C. Kot, "An integrated clustering-based approach to filtering unfair multi-moninal testimonies," *Computational Intelligence*, vol. 30, no. 2, pp. 316–341, 2014.

[12] Y. Yang, Y. L. Sun, S. Kay, and Q. Yang, "Defending online reputation systems against collaborative unfair raters through signal modeling and trust," in *Proceedings of the 2009 ACM symposium on Applied Computing*. ACM, 2009, pp. 1308–1315.

[13] 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*, 2013.

[14] H. Yu, Z. Shen, C. Miao, B. An, and C. Leung, "Filtering trust opinions through reinforcement learning," *Decision Support Systems*, vol. 66, pp. 102–113, 2014.

[15] Y. Liu and Y. Sun, "Anomaly detection in feedback-based reputation systems through temporal and correlation analysis," in *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. IEEE, 2010, pp. 65–72.

[16] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100–115, 1954.

[17] X. Wang, L. Liu, and J. Su, "Rlm: A general model for trust representation and aggregation," *IEEE Transactions on Services Computing*, vol. 5, no. 1, pp. 131–143, 2012.

[18] Y. Sun and Y. Liu, "Security of online reputation systems: The evolution of attacks and defenses," *IEEE Signal Processing Magazine*, vol. 29, no. 2, pp. 87–97, 2012.

[19] K. Regan, P. Poupart, and R. Cohen, "Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change," in *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 2006, pp. 1206–1212.

[20] 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*, vol. 193, pp. 149–185, 2012.

[21] J. WENG, C. MIAO, and A. GOH, "An entropy-based approach to protecting rating systems from unfair testimonies," *IEICE TRANSACTIONS on Information and Systems*, vol. 89, no. 9, pp. 2502–2511, 2006.

[22] T. Muller, Y. Liu, and J. Zhang, "The fallacy of endogenous discounting of trust recommendations," in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2015, pp. 563–572.

[23] 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)*, vol. 22, no. 12, pp. 1766–1780, 2010.

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

[25] J. Zhang, R. Cohen, and K. Larson, "Combining trust modeling and mechanism design for promoting honesty in e-marketplaces," *Computational Intelligence*, vol. 28, no. 4, pp. 549–578, 2012.

[26] Y. Liu and J. Zhang, "An incentive mechanism designed for e-marketplaces with limited inventory," *Electronic Commerce Research and Applications*, 2013.

[27] S. Liu, J. Zhang, C. Miao, Y.-L. Theng, and A. C. Kot, "iclub: an integrated clustering-based approach to improve the robustness of reputation systems," in *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 3*, 2011, pp. 1151–1152.

[28] H. Fang, J. Zhang, and N. Magnenat Thalmann, "Subjectivity grouping: learning from users' rating behavior," in *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2014, pp. 1241–1248.

[29] Z. Noorian, S. Marsh, and M. Fleming, "Multi-layer cognitive filtering by behavioral modeling," in *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2011, pp. 871–878.

[30] ——, "Prob-cog: An adaptive filtering model for trust evaluation," in *Proceedings of the IFIP International Conference on Trust Management (IFIPTM)*, 2011, pp. 206–222.

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

[32] J. L. W. V. Jensen, "Sur les fonctions convexes et les inégalités entre les valeurs moyennes," *Acta Mathematica*, vol. 30, no. 1, pp. 175–193, 1906.

[33] 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*, vol. 12, no. 2, pp. 183–198, 2006.

[34] D. Wang, T. Muller, J. Zhang, and Y. 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*, 2015, pp. 111–117.