Quantifying Robustness of Trust Systems against Collusive Unfair Rating Attacks Using Information Theory

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

Abstract

Unfair rating attacks happen in existing trust and reputation systems, lowering the quality of the systems. There exists a formal model that measures the maximum impact of independent attackers [Wang *et al.*, 2015] – based on information theory. We improve on these results in multiple ways: (1) we alter the methodology to be able to reason about colluding attackers as well, and (2) we extend the method to be able to measure the strength of any attacks (rather than just the strongest attack). Using (1), we identify the strongest collusion attacks, helping construct robust trust system. Using (2), we identify the strength of (classes of) attacks that we found in the literature. Based on this, we help to overcome a shortcoming of current research into collusion-resistance – specific (types of) attacks are used in simulations, disallowing direct comparisons between analyses of systems.

1 Introduction

Trust and reputation systems are designed to help users select trustworthy agents for interactions. Due to their rising popularity (e.g., e-marketplaces), various attacks have been discovered, and are threatening the security of these systems. Considering the threats, trust and reputation systems should be designed to mitigate these attacks (i.e. be *robust*). In this paper, we analyze collusive unfair rating (CUR) attacks.

CUR attacks are among various unfair ratings attacks, which aim to influence the trust evaluations made by users. CUR attacks differ from other unfair rating attacks, as malicious users (*attackers*) collude to achieve a same goal. Robustness issues of trust and reputation systems against collusion attacks are extensively studied. Some work focuses on detection of collusive unfair rating behaviours, while others aim to design mechanisms to defend against such attacks.

There can be various ways of colluding. Designers of trust and reputation systems sometimes verify the robustness only under specific CUR attacks [Qureshi *et al.*, 2010; Swamynathan *et al.*, 2010; Li *et al.*, 2013; Weng *et al.*, 2010]. This results in the following problems: first, these systems can only be known to be robust against the assumed attacks. Hence, one cannot know whether they are also robust to all other kinds of CUR attacks. Second, comparing the robustness of two trust models under specific attacks is not fair. The designer may design a system to be robust against a given attack, and use that specific attack to compare his system with another. Such a comparison is biased in favour of the proposed system.

Considering these two problems, we argue that if the general robustness of a trust system against CUR attacks are measured, then it should be tested against the strongest CUR attacks. If a trust system functions well under the strongest attack, then it is generally robust. Otherwise, if trust systems are merely robust to some given CUR attacks, which they are tested against, then we need to be able to compare the strength of these attack(s). In both the cases, we need to measure the strength of CUR attacks.

In [Wang *et al.*, 2015], we introduced information theory to identify and measure the worst-case (strongest) attacks by independent attackers. There, we focused on general robustness only, which depends on how effective the system is, given the strongest attacks. Here, we focus on how strong arbitrary attacks are (also using information theory), which allows us to reason about the quality of the validation of a trust system. Moreover, we shift focus from independent attackers to colluding attackers (including Sybil attackers). The formalism from [Wang *et al.*, 2015] must be fundamentally altered to allow for measurements of coalitions of attackers.

Contributions: We extend the methodology from [Wang *et al.*, 2015] to 1) quantify and compare CUR attacks found in the literature (Section 4), 2) quantify types of CUR attacks (Section 5), and 3) identify the strongest possible CUR attacks. Doing this, we found 1) attacks from the literature are not suitable to stress-test trust systems, 2) strongest CUR attacks are not considered in the literature, 3) always assuming the strongest attacks barely reduces the effectiveness, but greatly increases the robustness of trust systems. We consider the results from [Wang *et al.*, 2015] explicitly as a special case of CUR attacks.

2 Preliminaries

Our approach uses concepts from information theory:

Definition 1. *(Shannon entropy [McEliece, 2001]) The Shannon entropy of a discrete random variable X is:*

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

The Shannon entropy measures the uncertainty of a random variable, which reflects the expected amount of information carried in it. W.l.o.g. let 2 be the base of the logarithms. Since $x\log(x)$ is a common term, we introduce the shortcut $\mathbf{f}(x) = x\log(x)$. For practical reasons, we let $0\log(0) = 0$.

Definition 2. *(Conditional entropy [McEliece, 2001]) The conditional entropy of discrete random variable X given Y is:*

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

The conditional entropy measures the uncertainty of a random variable, given that another random variable is known. In other words, it measures the reduction of the expected information of a random variable when another random variable is revealed. Note that $H(X|Y) = H(X)$ if and only if X and Y are independent random variables, and $0 \leq H(X|Y) \leq H(X)$.

Definition 3. *(Joint entropy [Plunkett and Elman, 1997]) The joint entropy of discrete random variables X, Y (given Z) is:*

$$H(X,Y) = -\sum_{x_i \in X}\sum_{y_j \in Y} \mathbf{f}(P(x_i, y_j))$$

$$H(X,Y|Z) = -\sum_{z_h \in Z} P(z_h) \cdot \sum_{x_i \in X}\sum_{y_j \in Y} \mathbf{f}(P(x_i, y_j|z_h))$$

The joint entropy of X and Y is at most equal to the sum of the entropy of X and Y, with equality holds only if X and Y are independent.

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

Information leakage measures the gain of information about one random variable by learning another random variable. This definition coincides with mutual information [Papoulis and Pillai, 2002]. Information leakage is zero, if and only if the two variables are independent.

Definition 5. *(Hamming distance [Hamming, 1950]) The Hamming distance between $\overline{a} = a_0, \ldots, a_n$ and $\overline{b} = b_0, \ldots, b_n$, denoted $\delta(\overline{a}, \overline{b})$ is the number of $0 \leq i \leq n$ where $a_i \neq b_i$.*

Theorem 1. *(Jensen's inequality [Jensen, 1906]) For a convex function f:*

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

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

For brevity, we may use $\overline{X}$ to represent a collection of random variables, e.g., $\{X_1, X_2, \ldots, X_n\}$. Moreover, we may write $p(a)$ to mean $p(A = a)$, or even $p(\overline{a})$ to mean $p(A_1{=}a_1, \ldots, A_n{=}a_n)$.

3 Modeling CUR Attacks

In trust systems, users make trust decisions based on the available information. Part of such information comes from ratings. A user aims to learn from ratings to make deductions about an advisor's real opinion. For the user, an attacker hinders such learning, hiding its real opinion. In [Wang *et al.*, 2015], we proposed to quantify how much a user can learn as the information leakage of the ratings, which can then be used to measure the strength of the corresponding attacks.

$O \backslash R$	00	01	10	11
00	0	0	0	1
01	0	1	0	0
10	0	0	1	0
11	1	0	0	0

Table 1: Strategy matrix of the colluders from Example 1.

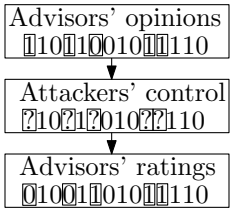

Figure 1: Advisors mo-deled as channel.

From an information theoretic perspective, advisors can be seen as a (noisy) channel. See Figure 1. The opinions of the advisors are being output as ratings, where the ratings need not match the true opinions (i.e., noise). Like in digital channels, not just the amount of noise matters, but also the shape of the noise. Thus, not just the amount of attackers matters, but also their behaviour. The difference in noise-per-advisor is often ignored in the literature, potentially skewing analysis of the attacker-resistance.

However, in [Wang *et al.*, 2015], only the strongest independent attackers are considered. In this paper, we extend the quantification method there to cover 1) collusive attackers and 2) non-strongest attacks. To measure the strength of a collusion attack, we need to measure the information leakage of the coalition as a unit. The ratings provided by colluders are interdependent, (potentially) revealing extra information to the user. Hence, when measuring the information leakage of a coalition, we cannot simply sum up the information leakage of individuals. (as can be seen below)

Example 1. *Consider a trust system with 4 advisors, 2 of which are colluding attackers.*[1] *Take the perspective of a user; he gets a rating from each of the advisors about their opinions of the target, and he does not know which two advisors are colluding. We assume that the opinions are positive and negative with 50% probability each. Non-colluding advisors always report the truth. Colluding advisors have one shared strategy. Here, the strategy dictates that if the attackers agree, then they both lie, and if they disagree, then they report the truth. The user received four ratings, three positive and one negative.*

[1]Note that there exist many methods which can estimate the number of colluders in a system [Liu *et al.*, 2008; Allahbakhsh *et al.*, 2013]. Hence we make the reasonable assumption in this work that the number or percentage of colluders is known.

We model the attackers' strategy in this example with the matrix in Table 1. The left column represents the real opinions (observations) of the two colluding advisors, represented as the combinations of positive: 1 and negative: 0. The top row represents the ratings which are in the same form. The cells provide the probability that the attackers report the column's rating, given the row's observation.

The user wants to learn about the observations of all advisors from the received ratings. We use random variables $O_i, R_i, i \in \{1, \ldots, 4\}$ to represent the observation and the rating of advisor i respectively. And we use random variable C_2 to represent two colluding advisors.

Before receiving the ratings, the information that the user has about the observations can be represented using joint entropy (Definition 3): $H(O_1, O_2, O_3, O_4)$. The joint entropy expresses the uncertainty associated with these four observations. Given the ratings, the information the user has of the observations becomes the conditional entropy: $H(O_1, O_2, O_3, O_4|R_1, R_2, R_3, R_4)$. Thus, the information that the ratings leak about the observations (information leakage) can be represented as follows:

$$H(O_1, O_2, O_3, O_4) - H(O_1, O_2, O_3, O_4|R_1, R_2, R_3, R_4)$$

The conditional entropy of observations given the ratings is:

$$\begin{aligned}
&H(O_1, O_2, O_3, O_4|R_1{=}1, R_2{=}1, R_3{=}1, R_4{=}0)\\
&= -\sum_{o_1,o_2,o_3,o_4} \mathbf{f}(p(o_1, o_2, o_3, o_4|R_1{=}1, R_2{=}1, R_3{=}1, R_4{=}0))\\
&= -\sum_{o_1,o_2,o_3,o_4} \mathbf{f}(\sum_{i,j} p(C_2{=}(i,j)|R_1{=}1, R_2{=}1, R_3{=}1, R_4{=}0)\\
&\cdot p(o_1, o_2, o_3, o_4|R_1{=}1, R_2{=}1, R_3{=}1, R_4{=}0, C_2{=}(i,j)))\\
&= -\frac{1}{2}\log(\frac{1}{12}) \approx 1.79
\end{aligned}$$

All combinations of O_1,O_2,O_3,O_4 are captured in o_1,o_2, o_3, o_4. The second equality follows from the law of total probability.

The entropy $H(O_1, O_2, O_3, O_4) = \log(2^4)$, since each O_i is positive or negative with exactly 50% probability. By Definition 4, therefore, we get $\log(2^4) - (-1/2 \log(1/12)) \approx 2.21$ bits of information leakage.

We now consider more general collusion attacks from the perspective of users. There are m advisors, k attackers ($0 \leq k \leq m$), and we assume non-attacking advisors always report the truth. The random variable O_i represents the opinion of the i^{th} advisor. We assume maximum entropy for the random variables $\overline{O}$, meaning $H(\overline{O}) = m$. Similarly, R_i represents the rating of the i^{th} advisor. For non-attacking advisors, $O_i = R_i$. For attacking advisors, we use $\sigma_{\overline{o},\overline{r}}$ to represent the probability that attackers that have observed $\overline{o}$ provide the ratings $\overline{r}$. The random variable C represents the coalition; its outcomes are, therefore, sets of advisors. The probability that a set c of k advisors are colluding is $p(c) = 1/\binom{m}{k}$. The strategies of attackers are expressed using the matrix in Table 2. The sum of each row equals 1, since, given an observation, the sum of the probabilities of all ratings is one.

The information leakage of all advisors' observations given their ratings is

$O \backslash R$	$0\cdots0$	$\cdots$	$1\cdots1$
$0\cdots0$	$\sigma_{0\cdots0,0\cdots0}$	$\cdots$	$\sigma_{0\cdots0,1\cdots1}$
$0\cdots1$	$\sigma_{0\cdots1,0\cdots0}$	$\cdots$	$\sigma_{0\cdots1,1\cdots1}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
$1\cdots1$	$\sigma_{1\cdots1,0\cdots0}$	$\cdots$	$\sigma_{1\cdots1,1\cdots1}$

Table 2: Strategy matrix of general collusion attacks

$$H(\overline{O}) - H(\overline{O}|\overline{R}). \tag{1}$$

The conditional entropy of observations given ratings is as follows:

$$\begin{aligned}
H(\overline{O}|\overline{R}) &= -\sum_{\overline{r}} p(\overline{r}) \cdot H(\overline{O}|\overline{r})\\
&= -\sum_{\overline{r}} p(\overline{r}) \sum_{\overline{o}} \mathbf{f}(p(\overline{o}|\overline{r}))\\
&= -\sum_{\overline{r}} p(\overline{r}) \sum_{\overline{o}} \mathbf{f}(\sum_{c} p(c|\overline{r}) \cdot p(\overline{o}|\overline{r}, c))
\end{aligned}$$

The general modeling and measurement of collusion attacks in this section can give us a way to quantify different attacks in the literature.

4 Quantifying CUR Attacks

Collusive unfair rating attacks have been studied by many researchers [Jiang *et al.*, 2013; Swamynathan *et al.*, 2010; Li *et al.*, 2013; Weng *et al.*, 2010; Jurca and Faltings, 2007]. They propose various methods to counter such attacks, e.g., detection based approaches [Qureshi *et al.*, 2010], incentive based approaches [Jurca and Faltings, 2007], and defense-mechanism design based approaches [Swamynathan *et al.*, 2010; Jiang *et al.*, 2013]. Based on simulations or experiments, they verify how effective their methods are in minimizing the influence of collusion attacks, namely improving the robustness of trust models. Sometimes such verification only covers specific attacks that are assumed or designed by the researchers themselves [Qureshi *et al.*, 2010; Swamynathan *et al.*, 2010; Li *et al.*, 2013; Weng *et al.*, 2010]. To compare the robustness of two trust models under specific attacks is unfair. One that fails under these attacks may behave better for some other attacks.

We argue that to equitably compare the robustness of two trust models, we need to compare the strength of attacks that they are tested against. A trust model should be considered more robust if it is verified under stronger attacks. From the section above, we know that information leakage of ratings can be used to measure the strength of attacks. We apply the method to some attacks found in the literature.

The authors in [Weng *et al.*, 2010] propose to mitigate the influence of unfair ratings by helping users to evaluate the credibility of the advisors, based on which to further filter and aggregate ratings. For collusive unfair ratings, the authors only consider the case in which malicious advisors provide unfairly high ratings for the colluding target, to boost its trustworthiness – *ballot-stuffing*. When evaluating the method, such attacks are configured as various percentages of attackers, namely 20%, 40%, 60%, 80%.

We use parameter m=100 to represent the number of all advisors in the system, then the number of attackers can be $0.2m, 0.4m, 0.6m, 0.8m$. The expected information leakage is 61.13, 39.69, 23.72, 10.84 bits, respectively[2].

The "FIRE+" trust model is proposed in [Qureshi *et al.*, 2010]. FIRE+ aims to detect and prevent collusion attacks. It considers two kinds of collusion attacks: in the first type, advisors collude with the target under evaluation, providing false positive ratings to promote the target as trustable. In the second type, advisors may collude to degrade the target, by providing false negative ratings. When evaluating the performance of "FIRE+", the authors consider three types of advisors: 10 honest advisors who always report the truth, 20 attackers that report all others as trustworthy, and 20 attackers that report the opposite of the truth. The information leakage for the second type of advisors is 6.79 bits[1], and for the third type of advisors is 22.52 bits[3].

The authors in [Li *et al.*, 2013] design a SocialTrust mechanism to counter the suspicious collusion attacks, the patterns of which are learned from an online e-commerce website. Instead of filtering collusive unfair ratings or preventing collusion behaviours, SocialTrust adjusts the weights of detected collusive ratings based on social closeness and interest similarity between a pair of nodes.

To evaluate the mechanism, it considers three attack scenarios: pairwise collusion in which two agents promote each other, multiple agents collusion in which agents all promote a boosted agent, and the collusion in which multiple agents promote each other. All the three scenarios are essentially about colluders ballot-stuffing to boost trustworthiness of other attackers that are under evaluation. The attacks for testing are configured as 9 pretrusted nodes and 30 attackers. Based on Theorem 3, the information leakage is 5.96 bits[1].

The authors in [Swamynathan *et al.*, 2010] aim to design a reliable reputation systems against two leading threats, one of which is user collusion. For performance evaluation of their system, the three same collusion scenarios as in the SocialTrust are considered. For the configuration of attacks, the percentage of attackers are varied from 10 to 50 percent. Based on Theorem 3, the information leakage is between 31.33 bits (for 50% attackers) to 75.97 bits (for 10% attackers)[1].

In summary, the CUR attacks used above are essentially ballot-stuffing and lying. From the quantification, we get following results: 1) referring to [Swamynathan *et al.*, 2010; Weng *et al.*, 2010], the information leakage of ballot-stuffing CUR attack decreases with the increase of the percentage of attackers. 2) Given the same number of attackers, the attack strategy of ballot-stuffing leads to much less information leakage than the strategy of lying (6.97<22.52 bits).

By relating these results with the strength of attacks, we get following conclusions. The ballot-stuffing attack gets stronger as the number of attackers increase. On the other hand, with the same amount of attackers, ballot-stuffing attack is stronger than the lying based attack.

[2]The computation of these values is omitted, as their generalization is provided in Theorem 3.

[3]The computation of these values is omitted, as their generalization is provided in Theorem 4.

5 Quantifying Types of CUR Attacks

The papers we discussed in the previous section only consider specific attacks in the verification of trust systems. As a result, we cannot know whether these systems are also robust against other attacks. We argue that to verify the robustness of a trust system, it should be tested against all kinds of attacks. However, this is not generally feasible. Therefore, we identify the strength of each attack type (and if it is a range, we identify the strongest attack within the type). To verify the robustness of a trust system to a type of attacks, we propose to test it against the strongest attack in that type.

We summarize various types of collusive unfair rating attacks from the literature. Information leakage measures the strength of the attacks. The information leakage is totally ordered, with an infimum (zero information leakage), therefore, there exists a least element. We refer to these least elements as the strongest attacks.

The types of collusion attacks are summarized as follows:

- I There is no colluding among malicious advisors, and they are behaving independently.
- II All attackers either boost (affiliated) targets, by unfairly providing good ratings (ballot-stuffing), or degrade (unaffiliated) targets, by unfairly providing bad ratings (badmouthing).
- III All the colluding advisors lie regarding their true opinions. As ratings are binary, they always report the opposite.
- IV The colluding advisors coordinate on their strategies in any arbitrary fashion.

The first attacking type is a special case of the collusion attacks, namely those that coincide with the independent attacks. The second type of attacks is commonly found in the literature, where all attackers are either ballot-stuffing or badmouthing (see, e.g., [Swamynathan *et al.*, 2010; Weng *et al.*, 2010; Li *et al.*, 2013]). There are also papers considering the other two types of attacks [Jurca and Faltings, 2007]. The last type of attacks is interesting, since it covers the three preceding types, as well as all other conceivable types.

We use the general CUR attack model from Section 3 to quantify these types of attacks. We use m, k to represent the number of advisors and the number of attackers, and use 0 to represent negative ratings, and 1 to represent positive ratings.

In the first type of attacks, all attackers operate independently. Each attacker chooses to report the truth with probability q, and lie with probability $1-q$.

Theorem 2. *The information leakage of any attacks of type I, is* $m - \sum_{d=0}^{k} \binom{m}{d} \cdot \mathbf{f}(\binom{m-d}{k-d} \cdot \frac{(1-q)^d \cdot q^{k-d}}{\binom{m}{k}})$.

Proof sketch. Since, when $\delta(\overline{o}, \overline{r}) > k$, $p(\overline{o}|\overline{r}) = 0$, wlog,

$$\sum_{\overline{o}} p(\overline{o}|\overline{r}) = \sum_{d=0}^{m} \sum_{\overline{o}:\delta(\overline{o},\overline{r})=d} p(\overline{o}|\overline{r}) = \sum_{d=0}^{k} \sum_{\overline{o}:\delta(\overline{o},\overline{r})=d} p(\overline{o}|r).$$

Moreover, since if $\exists_{i \notin c} r_i \neq o_i$ then $p(\overline{o}|\overline{r}, c) = 0$, wlog, $\sum_c p(\overline{o}|\overline{r}, c)p(c|\overline{r}) = \sum_{c:\forall_{i \notin c} r_i = o_i} p(\overline{o}|\overline{r}, c)p(c|\overline{r})$.

Substituting $p(\overline{r})$ by $1/2^m$, $p(c|\overline{r})$ by $1/\binom{m}{k}$ and $p(\overline{o}|\overline{r}, c)$ by $(1-q)^d q^d$, we obtain the information leakage. □

In the second type of attacks, we use $x, (1-x)$ to represent the probability that all attackers are ballot-stuffing and bad-mouthing, respectively. For $x = 1$, attackers are always ballot-stuffing and for $x = 0$, attackers are always bad-mouthing. To express this using the general attack model in Table 2, we assign the probability of "all ratings are 0" (meaning bad-mouthing) to be $1-x$ and "all ratings are 1" (meaning ballot-stuffing) to be x.

Before showing the next theorem, we introduce a shorthand notation. Let $\alpha_{k,h,y,0} = 0$, let $\alpha_{k,h,y,\text{-}} = \frac{\binom{h}{k}\cdot y}{\binom{m}{k}\cdot 2^{m-k}}$ and let $\beta_{k,i,j,z} = 1/2^k - \sum_{\ell=1}^{k}\binom{i}{\ell}\cdot \mathbf{f}(\frac{z\cdot\binom{i-\ell}{k-\ell}}{z\cdot\binom{i}{k}\cdot(1-z)\cdot\binom{j}{k}\cdot 2^k}) + \binom{j}{\ell}\cdot \mathbf{f}(\frac{(1-z)\cdot\binom{j-\ell}{k-\ell}}{z\cdot\binom{i}{k}\cdot(1-z)\cdot\binom{j}{k}\cdot 2^k})$. And let i be the number of "1" ratings, $j = m - i$ the "0" ratings, and $\mathcal{R}_{i,j}$ be the set of all ratings with i "1" ratings and j "0" ratings.

Theorem 3. *If $i<k$, let $z=0$; if $j<k$, let $z=1$; otherwise, let $z=x$. The information leakage of any attack of type II, is $m - \sum_{\overline{r}\in\mathcal{R}_{i,j}}(\alpha_{k,i,x,z} + \alpha_{k,j,1-x,1-z})\cdot\beta_{k,i,j,z}$ bits.*

Proof sketch. Note that $i < k$ and $j < k$ cannot simultaneously be the case, since at least k attackers rated "1" or k attackers rated "0". If $i < k$, then the attackers must have degraded (and if $j < k$, then boosted). The analysis of these two cases contains the same elements as the general case, which we prove below.

If $i \geq k$ and $j \geq k$, then the conditional entropy follows (via Definition 2), as $-\sum_{\overline{r}} p(\overline{r})\cdot\sum_{\overline{o}}\mathbf{f}(p(\overline{o}|\overline{r}))$. Remains to prove that $p(\overline{r}) = \alpha_{k,i,x,z} + \alpha_{k,j,1-x,1-z}$ and that $\sum_{\overline{o}}\mathbf{f}(p(\overline{o}|\overline{r})) = \beta_{k,i,j,z}$.

The equality $p(\overline{r}) = \alpha_{k,i,x,z} + \alpha_{k,j,1-x,1-z}$ follows from simple combinatorics, given that $p(\overline{r}) = \sum_c p(\overline{r}|c)\cdot p(c)$.

The equality $\sum_{\overline{o}}\mathbf{f}(p(\overline{o}|\overline{r})) = \beta_{k,i,j,z}$ also follows from total probability over c via $\sum_{\overline{o}}\mathbf{f}(\sum_c p(\overline{o}|\overline{r},c)\cdot p(c|\overline{r}))$. Straightforwardly, $p(\overline{o}|\overline{r},c) = 1/2^k$, provided for all $\ell \notin c$, $o_\ell = r_\ell$, and zero otherwise. Furthermore, $p((|\overline{r}) = \frac{x}{(1-x)\cdot\binom{j}{k}+x\cdot\binom{i}{k}}$ if for all $i \in c$, $r_i = 1$, and symmetrically when for all $i \in c$, $r_i = 1$. If neither is the case $p(c|\overline{r}) = 0$. The equality $\sum_{\overline{o}}\mathbf{f}(p(\overline{o}|\overline{r})) = \beta_{k,i,j,z}$ follows by applying these substitutions. □

In the third type of attacks, with probability q, all attackers report their true opinion, and with probability $1-q$, they all report the opposite. In the strategy model, we assign probabilities of reporting the opposite ratings as $1-q$, and probabilities of other cases as q. Then we compute the information leakage as follows:

Theorem 4. *The information leakage of the attack of type III, is $m + ((1-q)\cdot\log(\frac{1-q}{\binom{m}{k}}) + q\cdot\log q)$ bits.*

Proof sketch. Either $\overline{o} = \overline{r}$, or $\delta(\overline{o},\overline{r}) = k$, since either all attackers tell the truth, or all lie. Wlog $\sum_{\overline{o}}\mathbf{f}(p(\overline{o}|\overline{r})) = \sum_{\overline{o}:\delta(\overline{o},\overline{r})=k}\mathbf{f}(p(\overline{o}|\overline{r})) + \mathbf{f}(p(\overline{O}=\overline{r}|\overline{r}))$.

Straightforwardly, $p(\overline{o}|\overline{r}) = \frac{1-q}{\binom{m}{k}}$, when $\delta(\overline{o},\overline{r}) = k$, and $p(\overline{o}|\overline{r}) = q$, when $\overline{o} = \overline{r}$; substituting these terms yields the theorem. □

$O\backslash R$	00	01	10	11
00	1/11	2/11	2/11	6/11
01	2/11	1/11	6/11	2/11
10	2/11	6/11	1/11	2/11
11	6/11	2/11	2/11	1/11

Table 3: Ex. strongest collusion attack strategy matrix

For $q = 0$ and $k/m < 1/2$, $\log(\binom{m}{k}) \approx mH_2(k/m)$, where $H_2(p)$ is the entropy of a bernoulli distribution with parameter p. Thus, for small k and $x = 1$, information leakage roughly equals $m(1-H_2(k/m))$, which models the entropy m transmissions of bits that arrive intact with probability k/m.

In the fourth type of attacks, attackers are allowed to take any strategies, including the cases when they coordinate on different strategies. We aim to find a range of the strength of all of these attacks:

Theorem 5. *The information leakage of any attack of type IV, is between m and $\frac{2^k}{\sum_{0\leq i\leq k}\binom{m}{i}}$ bits.*

Proof sketch. The upper bound happens when $H(\overline{O}|\overline{R}) = 0$, which is satisfied $\overline{R}$ completely decides the value of $\overline{O}$. The crux is the lower bound; the minimal information leakage of type IV.

No matter what the attackers' strategy matrix is, k attackers can change at most k values. Therefore, $\sum_{\overline{o}} p(\overline{o}|\overline{r}) = \sum_{\overline{o}:\delta(\overline{o},\overline{r})\leq k} p(\overline{o}|\overline{r})$. There are $\zeta = \sum_{i=0}^{k}\binom{m}{k}$ possibilities for $\overline{o}$ given $\overline{r}$.

By Jensen's inequality (Theorem 1):

$$\sum_{\overline{o}:\delta(\overline{o},\overline{r})}\mathbf{f}(p(\overline{o}|\overline{r})) \geq \zeta\cdot\mathbf{f}(\frac{\sum_{\overline{o}} p(\overline{o}|\overline{r})}{\zeta}) \tag{2}$$

The equality holds iff for all $\overline{o}$ with $\delta(\overline{o},\overline{r}) \leq k$, $p(\overline{o}|\overline{r})$ is equal; meaning $p(\overline{o}|\overline{r}) = 1/\zeta$ iff $\delta(\overline{o},\overline{r}) \leq k$. Note that the number of $\overline{o}$ with $\delta(\overline{o},\overline{r}) \leq k$ is the same for any $\overline{r}$, hence the minimum of each $H(\overline{O}|\overline{r})$ is the same, allowing us to ignore $p(\overline{r})$. Filling in $p(\overline{o}|\overline{r}) = 1/\zeta$, the minimal information leakage can be computed: $\frac{2^k}{\sum_{0\leq i\leq k}\binom{m}{i}}$. □

The corresponding strategy matrix that leads to the minimal information leakage can be easily derived:

$$\sigma_{o_k,r_k} = \frac{1}{\zeta\cdot\binom{m-i}{k-i}\cdot p(c_k)}, \tag{3}$$

where $0 \leq i \leq k$ represents the Hamming distance between observations $\overline{o}$ and ratings $\overline{r}$. To give a concrete example of such a strategy, let there be 4 advisors, 2 of which are colluding, the strongest attack strategy is given in Table 3. Naively, one may expect the attackers to always lie, to ensure the probability that a given rating is truthful is half. However, each attacker in Table 3 reports the truth 3/11 times on average, disproving the naive view.

Finally, we show that attacks of types I, II and III are not the strongest attacks. In other words, the strongest attacks only occur in type IV (except in edge cases like $k=1$ or $k=m$):

Theorem 6. *For $1 < k < m$, there are attacks of type IV, such that every attack in type I, II or III has strictly more information leakage.*

Proof sketch. The proof of Theorem 5 applies Jensen's inequality, to prove that setting all $p(\overline{o}|\overline{r})$ equal (when $\delta(\overline{o},\overline{r}) < k$) provides the optimal solution. Jensen's inequality is strict when not all those $p(\overline{o}|\overline{r})$ are equal. Thus, it suffices to prove that for types I, II and III, with $1 < k < m$, $p(\overline{o}|\overline{r}) \neq 1/\zeta$.

For types I, II and III, there is only one degree of freedom (q, x and q, respectively). For no value for q or x, for all $\overline{o}$ with Hamming distance below k, $p(\overline{o}|\overline{r}) = 1/\zeta$. □

6 Discussion

We model a group of m advisors, containing k attackers, as a channel transmitting m bits, of which k bits are subject to noise (Figure 1). Like there are different types of noise, there are different types of attackers. We studied attack models found in the literature, and the types themselves. In this section, we analyze our findings, and put them into context.

The information leakage of the attacks from the literature (Section 4) is high compared to the minimum (e,g, FIRE+ provides 5.79 or 22.52 bits, whereas 0.03 bits is optimal). Higher information leakage means that it is easier for a user to learn from ratings. Models of attacks with high information leakage may not be suitable to stress-test a trust system, since it would be too easy to learn from ratings.

When interpreting these results, we must keep in mind that existing papers do not aim to minimize information leakage, but to faithfully model existing attacks. However, underapproximating attacks is undesirable. This is why we focus on the strength of attacks, even if minimizing the information leakage is not the original intention of the attackers. In fact, a robust system may never underapproximate the strength of attacks, linking robustness to the strength of attacks.

We now propose a method of designing robust trust systems based on the above. Given the attackers' behavior, trust evaluation becomes relatively simple. In [Muller and Schweitzer, 2013], the authors provide a general formula that allows mathematically correcting trust evaluations, given the attackers' behaviour. We propose to use computations in the formula, based on the assumption that the attackers' behaviour is the strongest possible attacks – minimal information leakage. Since such a model, by definition, can resist the strongest attacks, the system is robust. Whenever the system makes a trust evaluation, the actual information content of the evaluation can only exceed the systems' estimate.

The possible downside of assuming the strongest attacks, is that information available when attacks are weaker, is not being used effectively. However, the amount of information leakage when $k \ll m$ is high, even in the strongest attacks. When, on the other hand, $k \approx m$, the information leakage is significantly lower in the strongest attacks. We argue, however, that if $k \approx m$, it is unsafe to try to use the ratings as a source of information anyway. When the group of attackers is too large, no robust solution should use the ratings, as using the ratings would open a user to be easily manipulated. For small groups of attackers, the robust solution loses little performance, and for large groups of attackers, non-robust solutions are not safe. Therefore, we propose robust solution (and the strongest attack) to be the standard.

We distinguish four types of attacks. Attacks without collusion (I), attacks where the coalition boosts or degrades (II), attacks where the coalition lies (III), and the class of all attacks with collusion (IV). The former three are all instances of the latter. Attacks I, II and III deserve extra attention, since most unfair rating attacks in the literature are instances of them. However, while attacks I, II and III are interesting, they are trivially special cases of IV. Moreover, per Theorem 6, there are attacks in IV, not present in either I, II, and III – particularly, the strongest attacks.

The differences between the strongest attacks of type I and IV are remarkably small. For attacks of type I, we can only set one parameter, p, whereas for IV, we have $k(k-1)$ parameters that we can set. However, if, for example, we take m=30 and k=10, then attacks of type IV have at most a conditional entropy of 25.6597 (at least 4.3403 bits information leakage), whereas attacks of type I have at most a conditional entropy of 25.6210 (at least 4.3790 bits information leakage). The difference in conditional entropy is less than one part in a thousand. We conjecture that the minute difference is an artifact of the fact that the size of the coalition is given, and that if we remove that, the difference disappears entirely. Effectively, we suppose that the coalition does not effectively help minimize information leakage about observations, but rather help minimize information leakage about the shape and size of the coalition.

In Section 3, we have made several assumptions about ratings, advisors and targets. Non-binary ratings are also common in the literature. Our approach can generalize to other rating types by extending the alphabet of the ratings, at the expense of elegance. Our assumption that observations are 1 or 0 with 50% probability is just a simplifying assumption. In reality, these probabilities depend on the target. Since we are not interested in the target, but rather the advisors, we assumed maximum entropy from the target. The entropy is, therefore, lower for real targets – meaning the user has more information in practice than in theory.

7 Conclusion

In this paper we quantify and and analyze the strength of collusive unfair rating attacks. Sybil attacks where Sybil accounts provide unfair ratings are important examples of such attacks. Compared with independent attackers, the additional attacker strength gained by collusion is surprisingly small.

We apply our quantification to collusive unfair rating attacks found in the literature, where ballot-stuffing/badmouthing form the most well-studied types. The attacks in the literature are not maximally strong. We also quantify different types of attacks. And we identify the strongest possible collusive unfair attacks. Based on these strongest attacks, we propose trust systems robust against unfair rating attacks.

By this paper, the approach of applying information theory to quantify the strength of attacks becomes general and adaptable. Different types of unfair rating attacks (whitewashing, camouflage, etc.) can be quantified if the assumptions are changed accordingly.

Acknowledgement

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

References

[Allahbakhsh *et al.*, 2013] Mohammad Allahbakhsh, Aleksandar Ignjatovic, Boualem Benatallah, Elisa Bertino, and Norman Foo. Collusion detection in online rating systems. In *Web Technologies and Applications, Lecture Notes in Computer Science*, volume 7808, pages 196–207. Springer, 2013.

[Hamming, 1950] Richard W Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[Jensen, 1906] Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193, 1906.

[Jiang *et al.*, 2013] Siwei Jiang, Jie Zhang, and Yew-Soon Ong. An evolutionary model for constructing robust trust networks. In *Proceedings of the 12th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2013.

[Jurca and Faltings, 2007] 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.

[Li *et al.*, 2013] Ze Li, Haiying Shen, and Karan Sapra. Leveraging social networks to combat collusion in reputation systems for peer-to-peer networks. *IEEE Transactions on Computers*, 62(9):1745–1759, 2013.

[Liu *et al.*, 2008] Yuhong Liu, Yafei Yang, and Yan Lindsay Sun. Detection of collusion behaviors in online reputation systems. In *Proceedings of the 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1368–1372. IEEE, 2008.

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

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

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

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

[Qureshi *et al.*, 2010] Basit Qureshi, Geyong Min, and Demetres Kouvatsos. Collusion detection and prevention with fire+ trust and reputation model. In *Proceedings of the 10th IEEE International Conference on Computer and Information Technology*, pages 2548–2555. IEEE Computer Society, 2010.

[Swamynathan *et al.*, 2010] Gayatri Swamynathan, Kevin C Almeroth, and Ben Y Zhao. The design of a reliable reputation system. *Electronic Commerce Research*, 10(3-4):239–270, 2010.

[Wang *et al.*, 2015] 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*, 2015.

[Weng *et al.*, 2010] 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*, 22(9):1286–1298, 2010.