Quantifying and Improving the Robustness of Trust Systems

Dongxia Wang
School of Computer Engineering
Nanyang Technological University, Singapore
wang0915@e.ntu.edu.sg

Abstract

Trust systems are widely used to facilitate interactions among agents based on trust evaluation. These systems may have robustness issues, that is, they are affected by various attacks. Designers of trust systems propose methods to defend against these attacks. However, they typically verify the robustness of their defense mechanisms (or trust models) only under specific attacks. This raises problems: first, the robustness of their models is not guaranteed as they do not consider all attacks. Second, the comparison between two trust models depends on the choice of specific attacks, introducing bias. We propose to quantify the strength of attacks, and to quantify the robustness of trust systems based on the strength of the attacks it can resist. Our quantification is based on information theory, and provides designers of trust systems a fair measurement of the robustness.

1 Introduction

Trust systems allow users to select trustworthy targets for interactions, based on trust evaluation. The existence of various attacks in current trust systems affect the accuracy of trust evaluation, threatening the effectiveness of these systems. It is important to make a trust system function well under these attacks – being robust. A lot of defense mechanisms or trust systems have been proposed to defend against these attacks [Wang *et al.*, 2014].

The designers of defense mechanisms or trust systems often verify the robustness against specific attacks that are also modeled by themselves. This results in the following problems: first, they do not consider all attacks, hence they cannot ensure the robustness of their systems. Their models can only be declared robust against the attacks used in the verification. Second, the comparison among different trust systems or defense mechanisms under specific attacks may be biased. It is difficult to know whether an attack has been chosen for verification just to put a certain system in a better position.

Verification of robustness of a trust system requires evaluations under all attacks, which may be infeasible. For robustness verification against a type of attacks (e.g., unfair rating attacks), we propose to use the theoretically strongest attacks. We argue that if a trust system functions well under the strongest attacks, then it should be considered robust to the type. To compare the robustness of different trust systems under a type of attacks, we need to be able to compare the strength of attacks that they are tested against. If a trust system can resist stronger attacks, then it should be considered to be more robust. Either the scenarios above requires us to be able to measure the strength of attacks.

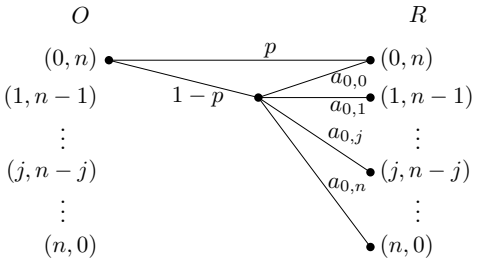

Figure 1: The extended rating model

We identify various types of attacks in existing trust systems in [Wang *et al.*, 2014], e.g., unfair rating attacks, on-off attacks, and re-entry attacks etc. Among these attacks, we currently deal with unfair rating attacks, where malicious advisors provide unfair ratings.

Consider measuring the strength of unfair rating attacks. A user aims to learn from recommendations provided by advisors about a target. We use information theory (specifically, information leakage) to measure how much the user can learn. Malicious advisors (attackers) reduce what the user can learn. We argue that unfair rating attacks are stronger if they have less information leakage, and the strongest attacks have minimal information leakage. The strength of attacks is quantified as such, since what matters is how effective a trust system is to a user; how much a user learns from ratings.

Based on this idea, we quantify and find the strongest attacks for unfair rating attacks. We study by dividing attacks into two types: independent unfair rating attacks and collusive unfair rating attacks.

2 Independent Unfair Rating Attacks

In independent unfair rating attacks, malicious advisors behave independently in providing recommendations to a user. We have modeled and analyzed this situation in [Wang *et al.*, 2015a], where the modeling is illustrated in Figure 1.

In the model, parameter p represents the probability that an advisor is honest (always reporting the truth), and $1-p$ represents the probability that the advisor is dishonest (strategically reports ratings). Observations (or real opinions about the target) (O) or ratings (R) of an advisor are of the form $R = (x, n-x)$; x and $n-x$ are the numbers of successful interactions, and unsuccessful interactions respectively. The total number of interactions is represented by n. We introduce $a_{i,j}$ as the probability of an advisor reporting $R = (j, n-j)$ when its observation is $O = (i, n-i)$.

The information leakage of an advisors' observations given its ratings is defined as $H(O) - H(O|R)$. When the behaviour pattern of an attack is parameterized using $p, n, a_{i,j}$, its information leakage of observations can be easily calculated. Also, we can measure the information leakage of the integrity of the target (denoted by T), which is $H(T) - H(T|R)$. We analyze the strongest attacks – minimal information leakage.

Regarding the strongest attacks, we prove some notable theoretical results: 1) under some attacks, a user gains information even if more than half the advisors are dishonest, 2) in attacks where the user gains no information, attackers sometimes report the truth, and 3) to minimize the information leakage of observations (O) and of the target's integrity (T), attackers need different rating strategies.

3 Collusive Unfair Rating Attacks

Attackers do not necessarily behave independently, as they may collude – collusive unfair rating attacks. Considering attackers in a coalition usually have a same purpose, we assign a combined strategy to them. The combined strategy dictates the (probabilistic) actions of each attacker individually.

Again, we use information leakage to measure the strength of attacks [Wang *et al.*, 2015b]. Unlike before, ratings in a coalition are not independent to each other, leaking extra information to the user. Hence, we cannot simply sum up the information leakage by individual attackers in a coalition, rather, we measure the information leakage of all observations given all ratings[1]:

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

In the formula above, the joint (conditional) entropy is used to represent the information carried with observations (given ratings).

Based on the information leakage measurement, we first quantify and compare the strength of specific attacks found in literature. Then, we study various types of collusive unfair rating attacks, as follows:

- I All attackers either promote (affiliated) targets, by ballot-stuffing, or degrade (unaffiliated) targets, by bad-mouthing.
- II All the colluding advisors lie regarding their true opinions.
- III The colluding advisors coordinate on their strategies in any arbitrary fashion.

For each type of attacks, we find a range of information leakage. And we found that the strongest attack happens in type III, with minimal information leakage $\frac{2^k}{\sum_{0 \le i \le k} \binom{m}{i}}$ bits. The strongest attack strategy is fairly complicated, and involves attackers reporting the truth surprisingly often.

We analyse several trust systems, and present that none of them are robust against the strongest attacks. We argue that for robust design of trust systems, the strongest collusive unfair rating attacks should be taken into account.

[1] We use $\overline{x}$ to denote a vector of variables.

4 Improve Robustness against Unfair Rating Attacks

We identify the strongest cases for both types of unfair rating attacks in sections above. For some of these strongest attacks, the information leakage is non-zero. And we are interested in making use of the leaked information to help users construct accurate trust opinions under the strongest attacks.

For independent unfair rating attacks, we propose a defense mechanism, named the induced trust computation (ITC). It allows users to derive accurate trust evaluation by exploiting the known strongest attack strategies. We compare the accuracy of trust opinions constructed by ITC with several other approaches, both under the strongest attacks and other types of attacks. We found that our defense achieves better accuracy in both cases. For collusive unfair rating attacks, we propose a similar method to base on the strategy of the strongest attack to derive the accurate trust opinions.

5 Future Work

Beside unfair rating attacks, other types of attacks exist in trust systems, such as whitewashing, camouflage, value imbalance exploitation, etc. In whitewashing and camouflage attacks, malicious advisors attempt to hide their bad reputation [Wang *et al.*, 2014]. We propose to model the information revealed in the changes of their behaviours over time, using additional random variables. We identify relationships between these random variables and the attackers' real opinions and recommendations. For future work, we want to find a way to quantify all types of attacks, to derive measurements of the general robustness of trust systems. More importantly, we want to design robust defense mechanisms to these attacks, using the strongest attack strategies. In so doing, we aim to improve the robustness of trust systems.

References

[Wang *et al.*, 2014] Dongxia Wang, Tim Muller, Yang Liu, and Jie Zhang. Towards robust and effective trust management for security: A survey. In *IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 511–518, 2014.

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

[Wang *et al.*, 2015b] Dongxia Wang, Tim Muller, Jie Zhang, and Yang Liu. Quantifying robustness of trust systems against collusive unfair rating attacks using information theory. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2015.