Quantifying and Improving Robustness of Trust Systems (Doctoral Consortium)

Dongxia Wang
Supervisors: Yang Liu and Jie Zhang
School of Computer Engineering
Nanyang Technological University, Singapore
wangdx1204@gmail.com

ABSTRACT

Trust and reputation systems are widely used, facilitating interactions among agents based on trust evaluation. These systems may have robustness issues, which should be paid attention. Various attacks in existing trust and reputation systems have been identified. 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 systems based on the strength of the attacks it can resist. Our quantification is based on information theory. Our approach provides designers of trust systems a fair measurement of the robustness.

Categories and Subject Descriptors

H.4 [**Trust and reputation systems**]: Robustnessattacks measures, robustness measures

General Terms

Measurement, Security

Keywords

Robustness, Trust Systems, Information Theory

1. INTRODUCTION

Trust systems allow users to select trustworthy targets for interactions. Due to increasing stakes, these trust systems are increasingly under attack by malicious agents. It is important to make a trust system robust under these attacks. The effectiveness of attacks is limited in a robust trust system. Whether a trust system is robust or not influence the accuracy of trust evaluation.

We identify a collection of attacks in existing trust systems in [2]. Among these attacks, unfair rating attacks are a type of attack that affect the accuracy of trust evaluation, by having malicious advisors provide fake ratings. Unfair rating attacks are under close examination by the community, and robust defense mechanisms are the goal.

The designers of defense mechanisms or trust systems often verify the robustness of their method against specific attacks that are also modeled by themselves. This results in following problems: First, they do not consider all of 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. One may be superior w.r.t. one attack, and another superior w.r.t. another attack, but the reader cannot see whether an attack has been chosen just to put a certain system in a better light.

Verification of robustness of a trust system requires evaluations under all attacks, which may be infeasible. Hence, we propose to use theoretically strongest attacks for robustness verification. Given a type of attacks (e.g., unfair rating attacks), we argue that if a trust system functions well under the strongest attacks, then it should be considered robust. To compare the robustness of two trust systems, we need to be able to compare the strength of attacks that they are tested against. If a trust system resists stronger attacks, then should be considered more robust. Whether we use the strongest attack, or compare attacks' strength, we need to be able to measure the strength.

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

Based on this idea, we quantify and find the strongest attack for unfair rating attacks. We divide 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 ratings. We have modeled and analyzed this situation in [1]. In there, we model such attacks as in Figure 1.

Appears in: *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May 4–8, 2015, Istanbul, Turkey.*

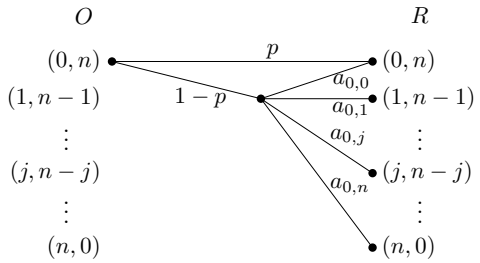

Figure 1: The extended rating model

In the model, 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 rating). Observations (O) or ratings (R) are of the form $R = (x, n-x)$; x is the number of successful interactions, $n-x$ the number of failed interactions, and n the total number of interactions. We introduce $a_{i,j}$ as the probability of an attacker reporting $R = (i, n-j)$ when the observation is $O = (i, n-i)$.

The information leakage of an advisors' true observations given its ratings is $H(O) - H(O|R)$. 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.

Some notable theoretical results that we prove: 1) even in the strongest attack, a user receives information even if more than half the advisors is malicious, 2) in attacks where the user receives 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. Unlike before, we cannot assign every attacker a strategy, but we have to assign a combined strategy to all attackers in a coalition. The combined strategy dictates the (probabilistic) actions of each attacker individually.

Again, we use information leakage to measure the strength of the attacks. This time, however, we measure the information leakage of all observations given all ratings. We use $\overline{x}$ to denote a vector of values:

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

Note that now the possible relationship among ratings should be considered, hence we use joint (conditional) entropy to represent the information carried with observations (given ratings).

Using this measurement, we first quantify the specific attacks found in literature. And compare their robustness based on information leakage of the attacks they use for verification. Then, we summarize 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 strategy is fairly complicated, and involves attackers reporting the truth surprisingly often.

The strongest attack in each type can be used to verify whether a trust system is robust to that type. 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 attacks should be taken into account.

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 interesting cases – the information leakage is non-zero. We propose to use the leaked information to help users make accurate trust opinions even under the strongest attacks.

For independent unfair rating attacks, we propose a general defense mechanism, named the induced trust computation, against the strongest attacks. It allows users to use ratings effectively. We compare the accuracy of the induced trust computation 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 to apply the strategies of the strongest attack to derive the accurate trust opinion.

5. FUTURE WORK

Beside unfair rating attacks, other attacks exist, such as whitewashing, camouflage, value imbalance exploitation, etc. We have concrete ideas to quantify whitewashing and camouflage attacks. In both of these attacks, the malicious advisors attempt to hide their bad reputation [2]. We propose to model the changes information in their behaviours over time, using additional random variables. We identify relationships between these random variables and the attackers' real opinions and ratings. For future work, we want to find a way to quantify all types of attacks, to derive measurements of the general robustness of trust and reputation 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

[1] D. Wang, T. Muller, A. Aravazhi Irissappane, J. Zhang, and L. Yang. Using information theory to improve the robustness of trust systems. In *Proceedings of International Autonomous Agents and Multi Agent Systems (AAMAS)*.

[2] D. Wang, T. Muller, Y. Liu, and J. 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. IEEE, 2014.