NANYANG
TECHNOLOGICAL
UNIVERSITY

QUANTIFYING AND IMPROVING THE ROBUSTNESS OF TRUST SYSTEMS

DONGXIA WANG

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING

2017

QUANTIFYING AND IMPROVING THE ROBUSTNESS OF TRUST SYSTEMS

DONGXIA WANG

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University

in partial fulfillment of the requirement for the degree of

Doctor of Philosophy

2017

Abstract

Trust systems help users evaluate trustworthiness of partners, which support users to make decisions in various scenarios. Evaluating trust requires evidences, which can be from ratings of other users (*advisors*). Rating are especially helpful when direct experiences of a user (*advisee*) are not enough. However, not all ratings are useful, and sometimes they may even be misleading. Either dishonesty (*unfair rating attacks*) or subjectivity of advisors can cause ratings deviating from the truth. Misleading ratings make trust evaluation inaccurate, and reduce the quality of trust-based decision making.

There exist various approaches to defend against unfair rating attack, aiming to make a trust system robust. Most of them are passive regarding attackers – they prepare for known attack strategies. When facing unknown attack strategies in the future, there is no guarantee whether they will be robust. Moreover, the robustness of these approaches is typically verified and compared under specifically constructed attacks, the results of which are not convincing. First, we do not know whether there exist worse attacks under which their performance may not remain. Second, it is not clear whether the specifically constructed attacks used for robustness comparison are more advantageous to some approaches. Last, different approaches may have different models for same types of attacks. There lacks a unified modeling of unfair rating attacks.

Thus, instead of passively defending after attacks are uncovered, we study unfair rating attacks in an active way, starting from considering all possible attack strategies. We propose a probabilistic modeling of attacks in any settings where ratings have discrete options. Our modeling is flexible in both space and time dimensions: allowing any number of advisors and rating levels, and also allowing attackers to change strategies over time. Given the uncertainty in predicting future attacks, we propose to emphasize

the strongest or worst-case attacks. From a security viewpoint, how well a system would perform under the worst case should be a key consideration in its design. We propose to use information theory, specifically information leakage to quantify the strength of an attack: less information leakage means the attack is stronger. We then analyze and compare the robustness of several trust systems based on the strength of attacks (especially the strongest attacks) they can handle. Different from existing approaches, our quantification is independent of specific systems, allowing a fair comparison of their robustness. We study attacks from two dimensions, 1) whether attackers are independent or collusive, and 2) whether their behavior patterns are static or dynamic. For each type of attacks, we identify the strongest attack strategies. Compared to them, the commonly studied attacks are far from being really threatening, which are thus not suitable to stress-test robustness.

There are approaches which not only consider dishonest ratings but also subjective ratings. They deal with dishonesty and subjectivity orthogonally – they distinguish the effects of dishonest ratings and honest but subjective ratings. However, subjectivity may twist with unfair rating attacks, in which way, influence the robustness of trust systems. We study their interplay, specifically: whether and how subjectivity, and different treatments of subjectivity may affect robustness. We also formally analyze two types of methods used to mitigate the effects of subjectivity: feature-based rating and clustering (advisors or ratings). We found that feature-based rating may deteriorate robustness, whereas clustering improves robustness. We also found that finer clustering enhances robustness, with tracking individual advisors as the extreme case.

In summary, our work provides a new perspective on studying unfair rating attacks and robustness of trust systems. Probabilistic modeling allows it to be flexible and active towards gaining robustness, compared with most existing approaches. Information theory-based measurement allows it to be general, and to enable a fair comparison of both attacks and robustness across various systems. Exposed worst-case attacks have drawn the attention of both relevant researchers and system designers to design more robust systems against more threatening attacks. Finally, some non-intuitive theoretical

results provide new insights for researchers, and also suggestions for system designers in practice.

Acknowledgements

There are plenty of people that I want to offer my sincere gratitude to. Without them, I may not reach what I have achieved during my PhD.

My supervisors, Prof. Yang Liu and Prof. Jie Zhang, have been playing a crucial role in leading me to a qualified researcher. I remember the very beginning of my PhD study, when I was quite new to doing research. Prof. Liu did not force me to study some specific problems immediately. Instead, he provided me freedom to get a wide horizon of the area first. And he allowed me to choose the research problems I like to solve. Prof. Zhang always provide very valuable opinions in evaluating my ideas and paper drafts. He offers me significant guidance on how to make my research more valuable and recognizable. Besides, he is very generous in spending his time on us: from discussing ideas and plans, to helping fix some writing problems. I am very grateful for both Profs supervision in the last four years. Whenever I get myself too relaxed, they remind me that I can do better, without which, I cannot be productive in research.

Tim, who was a post-doc when cooperating with me (and now is a lecturer in Oxford), has been very supportive. When I got stuck in small questions or technical problems, he helped to drag me out and taught me how to think in a higher level. As a mentor, he has been showing considerable patience. He would never feel troubled to discuss with me, or teach me how to fix even tedious problems. He has influenced me not only on the way of thinking, but also on exploring for meaningful research. Tim is also a very nice friend, he often invited us to cook, BBQ or have some drinks, and we really enjoyed the time. In a word, it has been my great pleasure to be a friend with him.

I also want to thank my TAC members, Prof. Alwen Tiu, Prof. Bo An, and Prof. Rongxin Lu. They supervised my research progress regularly, and brought up questions to help me reflect on my solutions or methods.

During the last four years, I have been surrounded by a group of nice fellows, who are also my friends. I appreciate their company, which enriches my PhD life and brings me beautiful memories. They are Zhimin Wu, Hao Xiao, Guozhu Meng, Xiaoning Du, Sa Gao, Wen Song, Zehong Hu, Yunwei Zhao, Yanhai Xiong, Qingyu Guo, MengChen Zhao, Naipeng Dong, Zhu Sun and so on. We encourage each other not only to overcome difficulties (e.g., not being too frustrated facing the rejection of papers), but also to live a healthy and relaxed life (e.g., by regularly playing sports and traveling together). These friends are by no doubt a treasure I got beyond my research.

Either my study or life is strongly backed by my family, especially my parents and sisters. Although they cannot directly help with my research, they encourage and support me spiritually, making me more confident in myself. They also always remind me to maintain a healthy diet and lifestyle, which is of great importance for both myself and my work.

Last but not the least, I own thanks to Jingyi Wang, my boyfriend. He takes care of me in various aspects, from my work to my daily life. His company and encourage made me feel not lonely to pursue a PhD in a foreign country.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Trust Systems

Trust systems help make decisions based on evaluating trustworthiness or reputation. Trust systems are applied extensively, e.g., in e-commerce (Amazon, Taobao), multi-agent systems, wireless sensor networks, vehicular ad-hoc networks, cloud computing, robotics and autonomous systems and so on. Specifically, trust systems can: compute reputation of sellers for buyers as a reference [1–3] (trust systems for e-commerce); select reliable routes for wireless sensor networks [4–6] (trust-based secure routing); and decide whether to grant access to an entity [7–9] (trust-based access control), and so on.

To evaluate trustworthiness of a target needs evidences, which can be from direct experiences or observations of a user. When there is insufficient direct evidence, trust systems tend to rely on ratings provided by other users (e.g., see BLADE [1], TRAVOS [10], MET [2], HABIT [11]). Throughout this thesis, we name those who provide ratings as advisors, those who receive ratings as advisees, and objects that are under rating as targets.

To provide reviews or ratings is a typical form of information sharing, which is popularly applied in practice, like in e-commerce (e.g., Amazon, Taobao, eBay) and

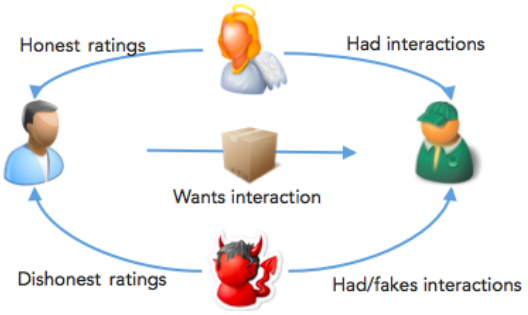


Figure 1.1: An example of unfair rating

some online P2P social networks (e.g., Yelp, Douban, and NetEase reviews). In e-commerce, when a customer is not sure about how to evaluate or select products, (s)he may refer to ratings given by some other customers, who have bought the products before. Douban for instance provides a platform where anyone can share their reviews and scores for a movie or a book. Rating and other forms of information sharing are also playing more and more important roles in fields such as Internet of Things (IoTs) and cyber security [12–14].

1.2 Misleading Ratings in Trust Systems

Shared information like ratings is not necessarily useful to its receivers. It may be irrelevant, redundant, or even misleading. Misleading ratings may result from the dishonesty of advisors. Dishonest (usually malicious) advisors deliberately report ratings that deviate from their observations, or even fabricate ratings for which they actually have no observations. For example, in Taobao, some sellers bribe a group of buyers to fabricate good ratings via fake transaction records, for the purpose of promoting their reputation. Such a type of rating behavior is also known as unfair rating attacks, which is a prevailing threat to trust systems. Figure 1.1 presents a scenario where before an interaction takes place, both honest ratings and dishonest ratings are received by an advisee (the blue icon).

Unfair ratings attacks reduce the quality of trust-based decision making, by introducing inaccuracy in trust evaluation. Take a trust-based access control system as an example. Unfair rating attacks may cause wrong decisions like granting access to a

malicious entity. The term *robustness* is usually used to describe a trust system's effectiveness against various attacks. Throughout this thesis, robustness specifically refers to how well a trust system deals with unfair rating attacks[1].

Besides dishonesty, misleading ratings may also result from subjectivity difference between an advisee and honest advisors. Given the same observation, different honest advisors may have different opinions, due to their different subjective dispositions. For example, if an honest advisor has expected much more about a target than another, (s)he may tend to give a lower rating. Subjectivity here means "based on or influenced by personal feelings, tastes, or opinions" (Oxford Dictionary). Although subjective ratings may be misleading, in some scenarios, they are very useful. For instance, it is assumed in recommender systems that honest advisors with similar subjective preferences tend to provide similar ratings, based on which recommendations are generated [16]. For the trust system in [17], trust is measured by the subjectivity similarity between users.

In this thesis, we consider these two factors impacting the quality of ratings – dishonesty and subjectivity. We specifically study how they influence the robustness of trust systems.

1.3 Problem Statement

In research about trust systems, there are various approaches aiming to improve the quality of ratings, by dealing with dishonesty (unfair rating attacks), subjectivity or both. By mitigating the negative effect of unfair rating attacks, these approaches aim to improve the robustness of trust systems. Several types of methods exist: to model advisors' degrees of honesty [3, 10, 18–22], to reward honest rating behavior [23, 24], and to re-interpret ratings regardless of whether they are honest [1, 11]. Generally, they have following problems:

[1] Various other types of attacks also exist in trust systems such as Sybil attacks, proliferation, on-off attacks and so on [15].

First, these approaches only study attacks that already exist in practice or have been attracting much attention. For example, ballot-stuffing attacks (where malicious advisors always report fake positive ratings to promote a target), which are the commonly known attacks in e-commerce, are also the popularly studied attacks. Although solving existing attacks is useful, to assume they consist of all threats overlooks attackers' flexibility, which cannot ensure the robustness of a system. Attackers may adapt their strategies, either over time or across systems. For example, after some attacks are uncovered, the designers of system A prepare defending mechanisms, which, do not work well as time goes by, because attackers already changed their strategies. If designers of system B assume that the attacks happened to A are the only threats they may face, then attackers may succeed with high probability, when they launch a new attack.

In the example above, the reason why attackers may succeed in the two systems, is that designers cannot "catch up with" attackers – their solutions are prepared for discovered (known) attacks but not for unknown ones. Such a kind of approaches is passive regarding improving the robustness. We do not know whether these approaches still work when faced with unknown attacks. And in order to "catch up with" attackers, they need to predict what strategies that attackers will undertake in future.

Second, even for known types of attacks, designers usually pay less attention in modeling or analyzing attackers' behavior, compared with in evaluating the effectiveness of their systems under attacks. In some related research, attacks appear merely to be testing tools for the robustness of systems or algorithms [1, 2, 25].

Third, when verifying a system's robustness, designers tend to use specifically-simulated attacks. For example, Weng et al. [20] consider ballot-stuffing attacks, and in evaluation, they consider several percentages of attackers: 20%, 40%, 60%, 80%. In Qureshi et al. [26], two attacks are used in evaluation: 20 attackers report all others as trustworthy, and 20 attackers report the opposite of the truth.

Positive verification results do prove that a system is robust against those specifically configured attacks, but do not prove that the system is robust under other attacks. Moreover, comparing the robustness of two systems under specific attacks is not fair. The designer may design a system to be robust against a given (type of) attack, and use

that specific attack to compare the system with another. Such a comparison is biased in favor of the proposed system.

And sometimes, attacks of same types are configured differently in different systems. For example, while both choose ballot-stuffing attacks for testing, Li et al. set 9 honest nodes and 30 attackers [27], while Swamynathan et al. vary attackers percentage from 10% to 50% [28]. There is a lack of work to formally model attacks independent of specific systems. A model which characterizes attackers' behavior serves to better analyze attacks, and more importantly, to make the design of defending approaches more targeted.

Last, there is no consensus regarding measuring the strength of attacks, or how harmful an attack is. Some researchers define based on their heuristics, e.g., Sybil attacks and camouflage attacks together would be stronger than either the single type [2]. Some others measure attacks based on how much ratings deviate from an advisee's opinion. Also, attacks may be measured based on their effect on a system – worse effect means more harmful attacks.

In practice, an honest advisor's ratings are typically subjective, which may differ from an advisee's opinion or from other advisors' ratings. Subjectivity is another important factor of misleading ratings, but it is intrinsically different from dishonesty. If defenses mechanisms for dishonesty change, then attacks may change, but subjectivity remains unchanged. Existing approaches generally cope with subjectivity and dishonesty orthogonally. Interestingly, subjectivity may change unfair rating attacks, e.g., introducing an attack where dishonest advisors camouflage as honest-but-subjective [29], thus changing robustness of trust systems. Therefore, the interplay of dishonesty and subjectivity needs to be studied.

1.4 Objectives and Contributions

We describe our objectives in dealing with unfair rating attacks, regarding all the problems mentioned in the section above. And also we present our main contributions on the way of achieving the objectives.

Being able to handle only existing attacks is not enough for robustness, as attackers are flexible in adapting their strategies. On the other hand, to accurately predict attackers' strategies beforehand is difficult. Thus, we suggest preparing a trust system for the *worst-case* attacks. Different trust systems may have different purposes, e.g., selecting reliable sellers (Amazon, Taobao), deciding whether to grant access or selecting secure routes (MANETs, WSNs, VANETs)[2]. We want the definition of worst-case attacks to be general and not confined to a specific trust system. For this reason, the definition cannot be attacks that minimize a system's effectiveness in achieving its purpose.

We focus on the nature of rating that belongs to all trust systems, and ignore specific purposes of the systems. Advisors provide ratings to share their observations (or experiences), to advisees who lack enough observations to make decisions. The prerequisite of decision making is to obtain useful information from ratings, namely the information about the (true) observations, which we treat as the core utility of advisees.

We use information theory, specifically information leakage to quantify the information that a rating provides about an observation. A stronger attack causes less utility for an advisee, meaning it causes less information leakage. The worst-case attacks lead to minimized information leakage. The measurement of attacks can be applied to measure robustness. If a system cannot function well under the worst-case attacks, then it is not robust. If a system can function well under a specific attack, then it is robust against the attack. If a system allows attacks with less information leakage, then it is less robust.

We categorize unfair rating attacks based on two considerations: whether attackers are independent or collusive (independent vs. collusive attacks), and whether an attacker's behavior pattern changes over time (static vs. dynamic attacks). We propose a probabilistic model for rating where attackers are assumed to be independent and static – a basic model. A group of parameterized probabilities, characterize an advisor's rating behavior, namely how various observations are re-interpreted into various options of ratings. We consider both the possibilities that an arbitrary advisor is honest or dishonest. By allowing probabilities' values range from 0 to 1, our model includes all possible

[2]MANETs represent mobile ad-hoc networks. WSNs represent wireless sensor networks. VANETs represent vehicle as-hoc networks.

behavior or strategies of an independent attacker. Besides, the model is flexible regarding the percentage of attackers and the number of rating levels. We then generalize the basic model to cover collusive attacks and dynamic attacks. We assume colluders share same goals and strategies (e.g., promoting a seller together), and they can report different ratings given same observations. For dynamic attacks, we use a random process to model the behavior of an advisor over time, and the basic model applies in each time step.

Based on the proposed rating models, we identify the worst-case strategies for each type of attacks. And we find some popularly studied attacks are far from the strongest to stress-test a system. We propose mechanisms which can effectively exploit ratings under the worst-case attacks, or weaker attacks, when being integrated into several existing systems. The rating models also allow us to analyze whether and how some system properties influence the harm of attacks, specifically through influencing information leakage. Based on our study, we propose suggestions for system designers on how to improve robustness.

Currently, the effect of subjective ratings and unfair rating attacks are distinguished. We are the first to formally analyze the interference of subjectivity and robustness. We first study the impact of subjectivity, and different ways of dealing with subjectivity, on robustness. We find that the introduction of subjectivity decreases robustness, and also higher degree of subjectivity means less robust.

1.5 Thesis Organization

In Chapter 2, we survey related work. We first present how trust systems are used to support making decisions, especially in security systems. Trust systems have been applied to solve several challenges faced with traditional security systems. Robustness is a key consideration to ensure the effectiveness of trust systems. Multiple types of attacks have been found in existing trust systems. We discuss recent existing work on dealing with these attacks, especially unfair rating attacks. Honest but subjective ratings

may also influence the accuracy of trust evaluation. Approaches which consider both dishonest ratings and subjective ratings are analyzed.

In Chapter 3, we introduce a list of concepts and theorems, most of which are from information theory. They will be used throughout this thesis to support our theoretical work.

In Chapter 4, we study the basic type of unfair rating attacks, where attackers are independent and have static behavior pattern over time. We base on information theory to quantify the strength of attacks. We propose a rating model, with which we compute the worst-case attacks. We formally prove that if there are not sufficiently many attackers, then ratings may still be useful. Our evaluations on several popular trust models show that they cannot provide accurate trust evaluation under the worst-case as well as many other types of unfair rating attacks. Our way of explicitly modeling dishonest advisors induces a method of computing trust accurately, which can serve to improve the robustness of the existing trust models.

In practice, attackers may not always be independent when rating, rather, they collude using a shared strategy (e.g., in e-marketplace, some buyers are gathered specifically for promoting the sellers who bribed them). In Chapter 5, we analyze collusive attacks. We alter the methodology proposed to be able to reason about collusive unfair rating attacks (CUR attacks) as well. We also extend the method to be able to measure the strength of any attacks (rather than just the strongest attack). We identify the strongest CUR attacks, helping construct robust trust systems. We also identify the strength of (classes of) attacks that have been used in the literature. Based on these, we help to overcome a shortcoming of current research in collusion-resistance – specific (types of) attacks are used in simulations, disallowing direct comparisons between robustness of systems.

Attackers may adapt their behavior or strategies over time, to better react to a system's defense mechanisms. In Chapter 6, we allow attackers to change their behavior in arbitrary ways, forming dynamic attacks. In the literature, camouflage attacks are the most studied dynamic attacks. But an open question is whether more harmful dynamic attacks exist. We generalize rating models for static attacks to cover dynamic attacks,

using stochastic processes. Information theory is still applied to measure the harm of attacks. The harm of an attack is influenced by an advisee's ability to learn from the past. We consider three types of advisees: blind advisees, aware advisees, and general advisees. We find for all the three types, camouflage attacks are far from being the most harmful. We identify the most harmful attacks, under which we find the ratings may still be useful to advisees.

While unfair rating attacks reduce the quality of ratings, subjectivity of honest advisors may also make ratings confusing. Existing approaches orthogonally deal with subjectivity (to make ratings more useful) and dishonesty (to improve robustness against unfair rating attacks). How subjectivity interplays with robustness remains an open question. In Chapter 7, we focus on whether (and how much) subjectivity – and different ways of dealing with subjectivity – change susceptibility towards robustness. We use information theory to measure the impact of subjectivity on robustness, and discover that increased subjectivity decreases robustness. We formally analyze two methods used to mitigate the effects of subjectivity: feature-based rating and clustering, and find that feature-based rating may deteriorate robustness, whereas clustering improves robustness. We also find that finer clustering enhances robustness, with tracking individual advisors as the extreme case.

In Chapter 8, we conclude this thesis, and also suggest several research topics for future.

Chapter 2

Related Work

In this chapter, we review existing research that is relevant to this thesis. The thesis deals with the robustness of trust systems, specifically robustness against unfair rating attacks. To provide a comprehensive view of the field, we consider various degrees of relevance to the thesis: ranging from most relevant works which also study unfair rating attacks, to less relevant works which study applications of trust systems. The robustness-related works are presented in the second section, where we emphasize unfair rating attacks. The applications of trust systems are presented in the first section. Trust systems are used to support decision making, in areas such as e-commerce, finance, public services[3], wireless networks (wireless communications, MANETs, WSNs, VANETs etc.) [5, 30–35]. When being used for decision making, trust systems act as a type of soft security mechanisms, and we focus on trust-based security applications.

2.1 Trust Systems for Security

Traditional security mechanisms are faced with challenges, which can be elegantly addressed with trust systems. Traditional mechanisms do not address security threats caused by internal malicious agents, while trust systems can detect them by evaluating their trustworthiness or reputation. Further, traditional security mechanisms are not

[3]For example, Ant Financial creates sesame credit score for each user, which serves as authorized reference in multiple services, such as loan, public facilities borrowing, public transportation, communication etc.

designed to protect against threats from malicious service providers [36], while trust systems can be used to protect both service consumers and providers.

Although the purpose of different trust-based security approaches is similar – namely, to guide security decisions based on trust evaluation results – the requirements to make them effective vary with regard to different security problems and circumstances. For example, trust-based access control requires privacy preserving, service-orientation, and revocation of invalid on-going access. Therefore, it is necessary to find and summarize the critical issues in designing trust systems for each type of security problems. We discuss four categories of trust-based security solutions – authentication, access control, secure service provision and secure routing. For each, we identify the important requirements, and see whether solutions in the literature adhere to the requirements. Note that our requirements overlap with those in the traditional security perspective.

2.1.1 Trust-based Authentication

In authentication, one verifies that the identity of a person or object is what it claims to be [37]. Trust systems are introduced to facilitate authentication in various applications [38–41]. Here, we give a brief introduction to some of such approaches.

In Park et al. [39], an authentication protocol is built to allow entities from one cluster to communicate with entities in another cluster. An agent that wants to communicate with the target agent in a new cluster needs to present certificates of its trust value, which are used for authentication. These certificates are signed by introducer in its original cluster.

In VANETs, where the number of authenticating executors is small compared to the number of on-board units (OBUs), OBUs need to wait for the nearest authenticating executor to authenticate before it can access services. TEAM in Chuang and Lee [40] is designed to reduce the waiting time. An OBU is regarded as trusted after being authenticated successfully, and will be authorized to authenticate not-yet-trusted OBUs. This mechanism builds chains of transitive trust relationships rooted in authenticating executors, which speeds up authentication.

In a federated identity management system, an agent's authentication assertions can be created and propagated across different authorities. This requires service providers to evaluate the trustworthiness of the agent's identity. In Gomi [38], authentication trust of a agent is used to evaluate whether the identity is legitimate.

Agents may not want to authenticate by providing personal information to untrustworthy entities. Therefore, in Husseini et al. [41], trust evaluation happens before the authentication process. Only when trust evaluation result is higher than a certain threshold, the authentication phase will be started.

Regarding these models, there are some key issues that are worthwhile to highlight and discuss below.

2.1.1.1 Mutual Authentication

In centralized systems, agents may have to unilaterally authenticate to a server. However, mutual authentication is vital in environments where two entities know little about each other, or where authentication protocols cannot always operate normally.

In Park et al. [39], the target agent also needs to provide a certificate of its trust value to the requesting agent. In TEAM [40], when the authenticating executor authenticates an OBU, the OBU needs to ensure that the authenticating executor is genuine.

2.1.1.2 Global Newcomers

A global newcomer is an agent which is new to the whole network. It has no past interactions, which means there is no evidence to evaluate its trustworthiness. Global newcomers must be considered by the authentication mechanism.

In Park et al. [39], a newcomer will first be monitored by all the other agents in a cluster for a certain time, based on which trust value will be computed. In Husseini et al. [41], risk assessment (based on second-hand information) is used for dealing with agents which are not evaluated before. However, these agents are not global newcomers as we define. How global newcomers are treated is not specified.

	Park et al. (2009) [39]	Gomi (2010) [38]	Chuang et.al (2011) [40]	ElHusseini et al. (2013)[41]
Mutual Authentication	Yes	Yes	N/A	N/A
Global Newcomer	Monitor	N/A	N/A	N/A
Privacy Preservation	N/A	Anonymity location privacy	N/A	Non-sensitive data used

TABLE 2.1: Properties of trust-based authentication models

2.1.1.3 Privacy Preservation

Protecting the privacy of agents being authenticated is important. Agents are reluctant to provide too much personal information for authentication. Authentication should avoid this.

Non-sensitive information is used in the evaluation of trust in Husseini et al. [41]. TEAM satisfies anonymity and location privacy.

In conclusion, an effective trust-based authentication protocol should achieve mutual authentication, privacy preservation, and be able to deal with global newcomers. Table 2.1 summarizes the properties of several recent trust-based authentication models. It can be seen that none of the models are sufficiently effective in achieving the desired properties above.

2.1.2 Trust-based Access Control

In distributed networks where resources for each agent are limited (e.g., limited processing power, memory space, battery life and bandwidth), resource discovery is vital. Unconstrained resource discovery, however, may lead to security threats. Access control is needed to restrict unauthorized access to resources based on security policies of the system.

In highly dynamic networks like MANETs, traditional access control approaches which rely on identity (e.g., mandatory access control, role based access control) are not

feasible. In trust based approaches, access rights are decided based on trust evaluation of the requesters and the security policies. Here, we discuss some key issues regarding trust based access control approaches we surveyed.

2.1.2.1 Service-Oriented Access Control

Not all services of a provider or device require the same level of security. For example, write access to a file may require a different security level than read access. Different security levels require different degrees of trust. Hence, a uniform trust threshold for all the services is infeasible. Thus, trust based access control should be service oriented, rather than device oriented.

In Li et al. [8], accesses to services with different security levels are assigned different trust thresholds, allowing dubious data requesters with low trust values to access some low-risk services but not high-risk services.

2.1.2.2 Privacy Protection

Privacy protection is crucial when agents' personal information is being collected and used, especially in e-business environments.

In Li et al. [8], an authorization of a data item depends on the requested time interval. Personal information is only kept for the period required to serve its purpose. Bhatia and Singh [42] builds a model which allows data owners to control the degree of data disclosure according to its privacy level. Personal data items are classified into different privacy levels based on the privacy preference of the data owner. Data item with higher privacy levels will be kept for a shorter time period.

2.1.2.3 Continuity of Access Rights

Continuity means the presence of on-going access rights [43]. After access is granted, new requester events (e.g., malicious behaviors) and system attributes may be received

by the access control manager. If these events indicate that the requester cannot be trusted anymore, on-going access should be revoked.

Re-calculation and re-evaluation systems are introduced in Li et al. [43] and [8]. The re-calculation system is responsible for re-calculating the trust value of the requester based on new evidence received during on-going access. The re-evaluation system is used to check if the access control rules are violated. On-going access rights would be revoked if either of these two systems receives negative results.

2.1.2.4 Ratings of Competitors

Providers of the same service are often competitors, trying to maximize their own revenue. As a result, they may be reluctant to warn each other about malicious requesters, or even provide dishonest ratings. An effective trust evaluation method takes this into account, rather than blindly incorporating ratings. In fact, this issue relates to the robustness property of trust systems that is discussed in more details in Section 2.2.

In Gupta et al. [44], providers are assumed to only be able to use their own data. But situations of inefficient first-hand experience are ignored in this case.

In summary, for effective trust based access control, following properties are desired, namely service oriented access control, privacy preserving, revocation of invalid on-going access and filtering malicious ratings from competitors. The properties of the recent access control models are listed in Table 2.2. None of the models meet all of the requirements we discuss.

2.1.3 Trust-based Secure Service Provision

Authentication and access control protect service providers from malicious requesters. The reverse, protecting requesters from malicious service providers, however, can also be of importance [36]. Secure service provision exists in service provision networks where trust already plays a role (e.g., eBay or Amazon) and in networks where providers have more control over the data of the requesters (e.g. mobile agent systems and cloud computing environments).

	Li et al. (2009) [43]	Li et al. (2011) [8]	Gupta ea. (2011) [44]	Bhatia (2013)[42]
Service Oriented	No	Different trust thresholds	No	Service specific
Privacy Protection	Yes	Yes	N/A	Yes
Continuity of Access Rights	Re-calculation Re-evaluation	Re-calculation Re-evaluation	N/A	Access rights revocation
Competitors' Ratings	N/A	N/A	Rely on own data	N/A

TABLE 2.2: Properties of trust-based access control models

Service provision networks like Amazon offer lots of open trading opportunities for consumers and providers, allowing providers to be malicious. Louta and Michalas [45] evaluate trust relations in a normal way. Providers are evaluated based on whether they honor the agreements built with consumers in their past performance. The approach combines first-hand experiences with second-hand evidence from other consumers.

In mobile agent systems, agents need to be protected from malicious execution hosts (execution service providers), which can cause agents' code to be disclosed, agents' data to be changed, and agents sent to wrong destinations. In MobileTrust [46], execution trust – which is the measure of trustworthiness of a host – is used to detect and eliminate malicious hosts.

In hybrid cloud computing environments, where both private and public clouds exist, the customers' control over their data is diminishing once their data is processed by third-party clouds [47]. In this situation, consumers need to ensure the trustworthiness of the cloud providers. Abawajy [48] builds a trust and reputation system to enable cloud customers to evaluate the trustworthiness of cloud providers, and to select best cloud services.

In conclusion, these models all attempt to select a service provider by evaluating its trustworthiness or reputation. Unlike other security problems, privacy is the main concern of trust based secure service provision. Besides, service-contract consistency

should also be considered in trust evaluation, to select both secure and high-quality services.

2.1.4 Trust-based Secure Routing

Secure routing is a routing technique in which the sender of a packet determines the complete sequence of agents through which to forward the packet [49]. Routing is vital for systems where agents cannot communicate directly to the destination agents (e.g., in MANETs and WSNs, agents can only communicate to neighbors within radio range). Message forwarding has to depend on collaborations among agents. Due to selfish or malicious intent, however, some agents may not collaborate as expected. Moreover, defective agents may also introduce faults. Both of these misbehaving agents threaten routing security.

The goal of trust based routing is to select trustworthy neighbors as packet forwarding candidates. Typically, each neighbor is assigned a routing score, and the agent with the highest score will be selected [50–55]. Here, we do not detail each trust based routing model, rather, we discuss some key issues in the models. On the basis of this, we want to highlight the requirements for an effective trust based routing scheme.

2.1.4.1 Trust Metrics

Different trust metrics capture different aspects of security of routing. An optimal metric should consider all of the potential security threats in routing.

In the routing model THWMP [50], trust is simply based on packet loss. Each agent calculates the packet loss by its upstream neighbor, with which it updates trust value of the neighbor.

In most of the models, however, there is a collection of trust metrics. In ATSR [52], eight trust metrics are combined, while each one stands for an aspect of security concern, such as forwarding (to detect agents denying to forward packets), packet precision (to ensure that no unexpected modification has occurred). These metrics are based on detectable events and can be used to measure them inversely. For instance, packet

modification can be measured by the packet precision metric. By these metrics, trust evaluation can well capture various types of misbehavior, which can then help improve the ability to resist them. In both Chen et al. [53] and Han et al. [55], direct trust evaluation depends on the percentage of successful interactions, which is defined by Han et al. as forwarding the message to the correct peer. Although it is claimed by Chen et al. [53] that an agent's trust is based on quality of service characteristics, such as packet forward and data rate, there is no explanation how they are implied in defining successful interactions.

2.1.4.2 Trust Evidence Propagation

In networks where agents are highly mobile, such as MANETs, VANETs, and CNRs, the neighbors of an agent change frequently, which causes it to have a smaller number of interactions with a larger number of partners [33]. As a result, there are not enough experiences for an agent to evaluate arbitrary partners. Therefore, effective trust evaluation should be based on both direct experiences and indirect trust evidence.

All of the surveyed models incorporate indirect trust evidence into calculation.

2.1.4.3 Routing Score

For security, trust value should be a factor of routing score which is used to select the next-hop. At the same time, it would be better if routing distance is incorporated, which impacts routing efficiency.

THWMP [50] decides whether to add agents to a path solely based on trust evaluation results. The remaining models all consider both trust value and the distance to destination, for the purpose of selecting trustworthy agents with less physical latency to the destination.

ATSR [52], takes remaining energy in the agent into consideration. Regardless of computation complexity, models considering distance would be more efficient in packet forwarding. In each model, a weighted sum function is proposed to aggregate these metrics.

DTEGR [51] optimizes the static weighting scheme in ATSR. It selects agents with trust values above a threshold to form a forwarding list, from which the agent with the closest distance to the destination will be chosen as the next hop.

In conclusion, there are three requirements for effective trust based routing. First, trust evaluation should capture as much potential misbehavior as possible. Second, indirect trust evidence should be incorporated (correctly) when the direct experiences are not enough for trust evaluation. Third, functional requirements on the routes should be considered, and a balance must be achieved between secure routes and efficient routes.

In summary, all models have some trust metrics, in various degrees of detail. Trust evidence propagation is present in all models, albeit implemented differently. Except THWMP, all models use both trust values and routing distance in the score.

2.1.5 Discussion

Different security problems have some common requirements on trust systems (e.g., privacy protection). We will discuss these in detail below. Further, we compare trust-based security mechanisms to traditional approaches.

2.1.5.1 Common Requirements

We identify three common requirements:

- *Privacy protection* is a consideration in all of security problems above. Authentication should avoid requiring private information. In access control, data owners need to specify security policies for data of different privacy levels, which should be combined with trust decisions. Protection of consumers' data is also an important component of service provision trust. In secure routing, data (including personal information) is often encrypted, to prevent internal agents from snooping.

- *Trust evidence propagation* is desired in environments where first-hand experience is insufficient to make effective trust decisions. In the aforementioned security problems, this is the case; most prominently in secure routing (Section 2.1.4). The requirement for trust evidence propagation, however, depends on the characteristics of the environment. Ratings of competitors may be dishonest and should be inspected (Section 2.1.2). Generally, this is a misleading feedback attack (Section 2.2.2) which can come from either service providers or requesters. Regardless of the type of security problems, it should be considered in trust systems where second-hand evidence is used for trust evaluation.

- *Global newcomers* should explicitly be taken into consideration, in trust systems where they occur.

2.1.5.2 Difference with the Traditional Security

Being social control mechanisms, trust-based security schemes are unlike traditional security mechanisms. The former can be regarded as *soft security* approaches, while the latter can be regarded as *hard security* approaches [56]. Hard security strives to guarantee that secure components work as intended. However, it is not feasible to guarantee security of all components in all systems. Without trust management, the system would be left unprotected. Soft security acknowledges the existence of malicious entities and behaviors, and it attempts to detect them and accordingly decreases the impact caused by them. Additionally, in traditional security, there are typically no security levels, just secure or not secure – hence the term *hard* security. In trust systems, trust evaluation provides a quantitative value for the object, which can represent various levels of security.

2.1.5.3 Combine with the Traditional Mechanisms

Trust systems can be combined with traditional mechanisms to support security. Trust systems evaluate entities based on their behaviors, while traditional security relies on rigorous mechanisms (e.g., certificates, credentials). In Lin and Varadharajan [46],

the evidence results from these two are combined to make security decisions. There are models in which trust systems are combined with role-based access control(see CATRAC [57] and the model in Ray et al. [58]). Specifically, in Ghali et al. [57], access is granted if both a client's trust level exceeds a threshold and the global role and permissions are correct.

2.2 Robustness of Trust Systems

Trust systems help to identify trustworthy entities as secure. However, to maximize profit, malicious entities may strategically attack trust systems. For example, malicious entities may provide dishonest ratings trying to defame an honest agent. A weak trust system may not function as desired under these attacks. Hence, robustness is crucial in the design of trust systems for security.

2.2.1 The Role of Robustness in Trust Systems

The accuracy of trust evaluation is closely related to the robustness of trust systems, which can further impact trust-based security decision making. Jøsang and Golbeck state that the correctness of the computed trust score is influenced by two factors: robustness of trust systems, and attack incentives [59]. The lack of incentives can reduce the number of attacks, and more robust systems can mitigate these attacks, both leading to increased accuracy. It has been shown that current trust systems can be easily attacked [60]. A robust trust system is a system where there are less attacks, or where the attacks' effectiveness is limited. For accurate trust evaluation to support security, we cannot ignore the robustness.

2.2.2 Attacks and Solutions

In order to study robustness, we must study potential attacks. We study the typical attacks. Attacks result from agents with malicious intent. We cannot detect or prevent malicious intents. We can, however, mitigate the damage caused by malicious behaviors

– behavior resulting from malicious intents – or disincentivise attacks. For example, we can detect unfair ratings and filter them out or detect the dishonest raters first and abandon their ratings [19]. When the attack relies on the vulnerabilities of the system, then the system must be fixed. For example, if the system assumes one account per agent, then it is crucial for the authentication system to identify multiple identities registered by one agent. Otherwise the so-called Sybil attack or the newcomer attack (see below) can happen. Although a comprehensive solution against all of existing attacks does not (yet) exist, researchers have proposed methods to mitigate some of them.

2.2.2.1 Unfair Rating Attacks

The unfair rating attacks, also known as misleading feedback attacks, have been recognized as an important threat especially in trust systems [31, 60, 61]. Below we first review existing approaches that only consider mitigating (the effect of) unfair rating attacks. We focus on trust systems, where such attacks attract the most attention. There are usually three types of approaches: filtering or discounting ratings (sometimes based on advisors trustworthiness) [3, 10, 18–22, 27], incentivizing honestly rating [23, 24], measuring and exploiting information [1, 11]. The work in this thesis belongs to the last type.

The first type attempts to filter or discount unfair ratings. The reputation of a target is derived by aggregating the weighted filtered/discounted ratings. The weights are mostly decided by evaluating the trustworthiness of advisors. There are several representative examples. Based on the beta probability density function, TRAVOS [10] examines the reliability of an advisor's previous ratings, based on their difference from an advisee's observations. Then such reliability values act as weights when combing advisors' evidence. Besides, Sun et al. [18] propose a method which is characterized as follows: Ratings from advisors with higher degrees of trustworthiness are more capable of propagating. Ratings from advisors with lower degrees of trustworthiness have smaller impacts on decision-making. Liu et al. [22] compute advisors' trust values based on both local (ratings about the concerned target) and global rating information

(ratings about other targets). In SocialTrust mechanism [27], ratings are weighted based on social closeness and interest similarity between a pair of advisors.

Data mining approaches such as clustering and classification are also applied to identify dishonest advisors [3, 21]. In both Liu et al. [21] and Irissappane et al. [3], advisors are clustered, based on the difference between their ratings and a buyer's. Advisors that are in the same cluster as the buyer are regarded as honest, since they have smaller rating difference, and thus similar rating behavior with the buyer. Irissappane et al. [3] use the difference of Spearman's rank correlation to decide trustworthiness of advisors, when an advisee has little direct experience. Besides, multi-criteria ratings are considered, where different clusters are formed for different sets of criteria. Ratings with multiple levels are considered by Liu et al. [21]. The numbers of transactions rated with different levels form a rating vector.

To investigate advisors' trustworthiness is necessary to evaluate their ratings, however, we also need to understand how the advisors behave when they are perceived to be dishonest. The trust models above have the common assumption that dishonest advisors only adopt some simple strategies. Moreover, these models only consider attacks with known characteristics (existing attacks), which makes them passive faced with new attacks. When new attacks occur, we do not know whether their approaches will still be robust.

Instead of discounting unfair ratings, BLADE [1] and HABIT [11] aim to deduce the true observations behind them. Both the models learn the statistical correlation between an advisor's ratings and an observation, based on which the advisor's future ratings are re-interpreted. For example, if an advisor keeps reporting positive ratings when the truth is negative, then his ratings will be reversed before being used. In BLADE, the learned correlation forms an evaluation function, which describes how the advisor rates (i.e., how its ratings distribute given its observations). HABIT allows any types of rating representations, while BLADE only allows discrete representations of ratings [11].

Incentive and punishment-based schemes aim to reduce the occurrence of fake ratings, by rewarding truthful feedback. In Zhang et al. [62], reputable buyers would be provided the increased quality of products with decreased prices. Each buyer keeps

a group of advisors, consisting of the most trusted fellow buyers. Sellers identify reputable buyers based on the number of advisor groups they belong to. Honest buyers will benefit from its ratings by gaining more profitable transactions. Liu and Zhang [63] provide each seller a limited inventory, where buyers compete with each other to get the purchase. In a naive system, buyers would provide negative feedback about high-quality sellers, since they are scarce. The authors propose an incentive mechanism where buyers providing truthful ratings are assigned higher scores, which makes them have more chance to transact with reputable sellers.

There are trust models trying to address some specific unfair rating attacks. For example, Feng et al. [64] study three attacks, namely RepBad, RepSelf and RepTrap, and propose defenses against them. Jiang et al. [2] propose a trust model based on evolutionary computation (called MET) to effectively cope with four typical attacks and their combinations. Liu et al. [65] propose a fuzzy logic based trust model, to effectively resist the attacks that exist in a Cyber competition where human participants compete to break down a trust system. However, it is still difficult to say that these trust models will be robust to all possible unfair rating attacks. Moreover, to compare robustness under specific attacks is unfair. We do not know whether these attacks are chosen to be beneficial to certain trust models. Also, if given some other attacks, we do not know whether the comparison results would be reversed.

2.2.2.2 Discrimination Attack

In a discrimination attack, the service provider provides high quality services to some groups, but low quality services to others. This induces contradictory ratings among these groups, which may impact their trust value as advisors. If a group identifies dishonest advisors based on rating difference to its own (like using the detection-based schemes in Section 2.2.2.1), then the group which provides contradictory ratings will be regarded as dishonest. To defend this attack, self-experiences should not be set as the only benchmark to identify dishonest advisors. We found no effective solution for this attack.

2.2.2.3 On-off Attack

On-off attack means malicious entities behave inconsistently over time, exploiting the trust computation algorithm, while remaining undetected [18]. For example, an agent firstly accumulates a high trustworthiness through good behavior. Then, additional ratings play a smaller role in changing its reputation, and it starts behaving badly while maintaining an acceptable reputation. This suggests that older behavior records may indicate less about an agent's current behavior.

To address this problem, the most commonly used approach is to introduce a forgetting factor [18]. However, a fixed forgetting factor can also be used by malicious entities to facilitate the on-off attack. With a long forgetting factor, the computed trust value does not reflect the current state of the agent, whereas with a short forgetting factor, the behaviors are forgotten quickly, and the agent regains its trust too easily. Sun et al. propose an adaptive forgetting scheme [18]. When the trust value is below the threshold, a longer forgetting factor is used, otherwise, a shorter forgetting factor will be used. Therefore, the trust value can keep up with the change in the agent's behaviors, and moreover, recovery from a low trust value requires enough good behaviors.

In P2P systems, the on-off attack is called dynamic personality of peers. Xiong and Liu [66] propose an adaptive time window-based algorithm to react to such personalities. The idea is to adaptively choose a smaller time window to collect the most recent behavior records of a peer, when its performance drops. The trust value computed from those most recent records will be compared with the one computed from all records in a larger time window. If it is lower than a certain threshold, which indicates the peer is performing badly recently, then it will be set as the peer's trust value.

2.2.2.4 Sybil Attack

The Sybil attack comes from malicious entities who freely create several identities. The attacker can use different identities each time to behave maliciously, and then the blame will be shared by all of these identities, instead of being afforded by itself. Also, relying on its multiple identities, the attacker can give multiple ratings over the same service

object, unfairly increasing its influence on the service's reputation. Countermeasures against Sybil attacks are usually confined to a particular network (e.g., VANETs [67], P2P [68], WSNs [69]). In [70], admission control is used to block unnecessary raters when there is enough information to predict the rating value of a service item. Based on this intuition, only ratings from the reliable raters will be used for prediction of the rating value.

2.2.2.5 Newcomer Attack

An agent may cause a newcomer attack if it can easily register a new identity, also known as whitewasher attacks. By re-registering, the attacker can easily get rid of its previous bad behavior history, and bad reputation. The newcomer attack is also called the re-entry attack [59]. Similar countermeasures as against the Sybil attack may work here. In addition, a penalty for new agents works effectively against newcomer attacks (however, punishing new agents may be unacceptable in many settings).

2.2.2.6 Value Imbalance Exploitation

Typically, ratings do not indicate the value of the services. A malicious agent can gain high profits and also reputation by providing more high quality services with low value, while providing low quality services with high value. To defend this, one simple way is to assign weights to ratings as a function of the value of services [59].

2.2.3 Attack Model

In this section we summarize some attacks and existing solutions in trust systems. These attacks happen in different stages of a trust system: the Sybil attack and newcomer attack happen in the login phase; the discrimination attack, the on-off attack and value imbalance exploitation happen in the transaction phase; and the misleading feedback attack happens in the trust evaluation phase. Jøsang proposes a concrete model for attack functional phases and attack vectors in trust and reputation systems [71]. Although

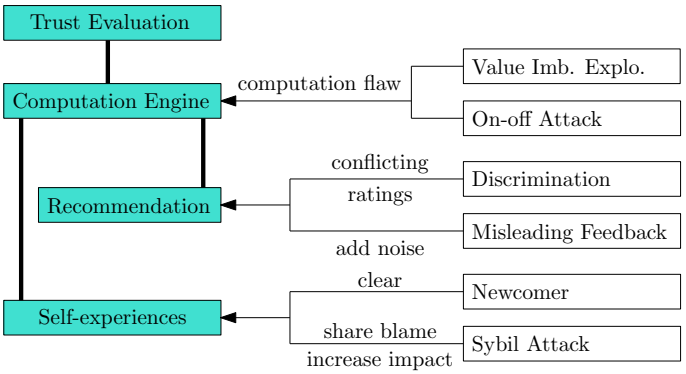


Figure 2.1: Attack model

occurring in different stages, we conclude that these attacks ultimately aim at trust evaluation. Their intentions and targets are presented in Figure 2.1.

First, value imbalance exploitation and on-off attacks actually attempt to exploit flaws on computation algorithms, to behave maliciously without proportionate negative impact on their reputation. In the former, when computing trust values, if the computation algorithm does not weight ratings for service of different values, then attackers can gain unfairly high profits while maintaining a disproportionately good reputation. In the latter, if the computation algorithm does not correctly weight old and recent behavior, then attackers can adaptively oscillate between good and bad behaviors, and remain undetected.

Second, attacks such as unfair rating attacks and discrimination aim at second-hand evidence, thus impacting trust evaluation results. A misleading feedback attack consists of an agent providing dishonest ratings. Discrimination itself seems to have no direct harms (Section 2.2.2), however, discriminating groups will cause contradiction among ratings of these groups, which can influence their recommendation trust in each other.

Third, newcomer and Sybil attacks can impact both first-hand and second-hand evidence. Newcomer attackers attempt to get rid of its bad behavior history, which may be first-hand or second-hand evidence to others. Sybil attackers can provide multiple ratings over the same service object and/or they diminish blame for their bad behavior by spreading it over fake identities.

In this thesis, we focus on studying unfair rating attacks. On one hand, unfair ratings attacks are the easiest to launch with minimum cost. Dishonesty is enough to generate

unfair ratings, without the need of creating new accounts or adapting to trust computation mechanisms. For this reason, unfair rating attacks are very prevalent in practice. On the other hand, as an important form of second-hand information, unfair ratings directly (or indirectly) impact the quality of decision making. The more a system relies on ratings, the more the quality of ratings matters. Moreover, unfair ratings are intrinsically a type of false information. The study on unfair ratings can provide us hindsight on how to deal with general false information, not only in trust systems, but also in many other areas (see Future Work section in Chapter 8).

2.3 Subjective Ratings

Both subjectivity and unfair rating attacks are obstacles that trust systems must cope with. Subjective ratings from honest advisors may also be misleading, but they should not be mistakenly treated as unfair rating attacks. Mehta [72] introduces the notion of shilling attacks into collaborative filtering. Fang et al. [25] propose to explicitly distinguish dishonesty and subjectivity difference in the modeling of advisors' trustworthiness. In collaborative filtering advisor systems, users' subjective tastes are matched based on their ratings, which serve as references to provide recommendations [73, 74].

There are various ways to eliminate the negative effect of subjectivity. One way is to explicitly rate for individual aspects or features of a target [17]; a widely applied approach in e-commerce. Clustering advisors based on their rating similarity is another typical way, as ratings of honest advisors reflect their subjective dispositions. Some clustering-based approaches do not differentiate advisors' honesty [17], while some others filter dishonest advisors before clustering [29, 75].

How to deal with clusters also varies among these approaches. In Noorian [29], malicious advisors identified would be excluded. In Fang [75], ratings of some dishonest advisors may also be used. There are ways which do not distinguish honesty and subjectivity, but learn any correlation between ratings and the truth for individual advisors [1, 11]. We treat such individual correlation-learning ways as an extreme case of clustering, where each individual advisor forms a cluster.

As we see, the existing work distinguish the negative effects of subjective ratings and unfair ratings. We are interested in whether subjectivity would influence the effects of unfair rating attacks. We argue in favor of a holistic approach, where advisors are both subjective to some degree and potentially dishonest (See Chapter 7).

Chapter 3

Preliminaries

Our approach is mostly supported by concepts and theorems in probability theory and information theory, as presented below. The modelling of various types of unfair rating attacks is based on probability theory. For example, we use conditional probabilities to characterize how ratings are generated from true observations. Dynamic rating behaviour is modelled as stochastic processes. Information theory provides us the mathematical foundation of measuring the strength of attacks. For example, Shannon entropy is used to quantify the expected information of an observation or a rating variable. Information leakage is used to quantify the information that a rating reveals about the underlying true observation. Most definitions or theorems will be applied almost in each of the four studies in this thesis.

Definition 3.1. (Surprisal [76]) The surprisal (or self-information) of an outcome x of discrete random variable X is given as: $I(X{=}x) = -\log(P(X{=}x))$. Surprisal can be generalized for continuous random variable Y as: $I(Y{=}y) = -\log(p_Y(y))$.

Definition 3.2. (Shannon entropy [77]) The Shannon entropy of a discrete random variable X is given:

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

The Shannon entropy gets maximum when all possible outcomes are equiprobable. Further, it can be generalized to differential entropy for continuous random variables Y as:

$$h(Y) = \mathbf{E}(I(Y)) = -\int_Y p(y) \cdot \log(p(y)) \, \mathrm{d}y$$

The Shannon entropy measures the expected amount of information carried in a random variable, which is decided by the uncertainty of the random variable. The base of the logarithm throughout this thesis is set as 2, following the corresponding definitions in information theory. 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 3.3. (Conditional entropy [77]) The conditional entropy of discrete random variables X under Y is given as:

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

It can be generalized to continuous X and Y as:

$$H(X|Y) = -\int_Y p(y) \cdot \int_X \mathbf{f}(p(x|y)) \, \mathrm{d}x \, \mathrm{d}y$$

The conditional entropy measures the expected amount of information in one random variable when another random variable is known. $H(X|Y)$=$H(X)$ iff X and Y are independent. For brevity, we leave out the cases where only one of X and Y is continuous. Note that $0 \leq H(X|Y) \leq H(X)$.

Proposition 3.4. *For any random variables X, Y: $H(X) = H(X|Y)$ iff $P(X)$=$P(X|Y)$, or X and Y are independent.*

Definition 3.5. (Joint entropy [78]) The joint entropy of discrete random variables X, Y (given Z) is:

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

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

Since $H(X,Y)=H(X)+H(Y|X)$, and $H(Y|X)\leq H(Y)$, with equality holds iff X and Y are independent. The joint entropy of X and Y is at most equal to the sum of the entropy of X and Y.

Definition 3.6. (Cross entropy [78]) The cross entropy for two distributions P and Q is given as:

$$H(P,Q) = E_P[-log(Q)] = H(P) + D_{KL}(P||Q)$$

The cross entropy measures the distance between the probability distribution the data actually follows and the distribution that is assumed. $D_{KL}(P||Q)$ named Kullback-Leibler divergence is a non-symmetric measure of the difference between distributions P and Q [79]. It measures the information gained if the real distribution P is used instead of its approximation Q. When $P{=}Q$, $H(P,Q){=}H(P)$, $D_{KL}(P||Q){=}0$, which are their minimal values.

Definition 3.7. (Information leakage) The information leakage of X under Y is given as:

$$H(X) - H(X|Y)$$

Information leakage is the gain of information (or the reduction of uncertainty) about one random variable, when knowing another random variable. This definition is the same with mutual information [80]. Information leakage is zero, iff the two variables are independent.

Definition 3.8. (Hamming distance [81]) The Hamming distance between $\vec{a} = a_0, \ldots, a_n$ and $\vec{b} = b_0, \ldots, b_n$, denoted $\delta(\vec{a}, \vec{b})$ is the number of $0{\leq}i{\leq}n$ where $a_i{\neq}b_i$.

Theorem 3.9. *(Jensen's inequality) For a convex function f:*

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

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

In this thesis, Jensen's inequality is mainly used to solve optimization problems like calculating the strongest attack strategies.

Throughout this thesis, we introduce some shortcuts as follows: 1) for random variable X, we use x for its outcomes, and $p(x)$ to mean $p(X{=}x)$; 2) We write $\forall x$ ($\sum_x$) to denote (sum of) all outcomes of X; 3) we may use $\overline{X}$ to represent a collection of random variables, e.g., $\{X_1, X_2, \ldots, X_n\}$, and $p(\overline{a})$ to mean $p(A_1{=}a_1, \ldots, A_n{=}a_n)$; 4) we may write $\overrightarrow{X_i}$ to mean $X_i, \ldots, X_1$, or an empty list, when $i{=}0$.

Chapter 4

Independent and Static Unfair Rating Attacks

By injecting fake information, unfair rating attackers aim to impact decision making. From a security view, to design a robust system[4] needs preparation for the worst-case attacks, which yield the least useful information. In this chapter, we study the basic type of unfair rating attacks: independent and static attacks, where attackers are assumed to be independent and their behavior patterns remain unchanged over time [82]. The basic type of attacks is the most studied type. The existing work typically focus on evaluating advisors' trustworthiness [3, 10, 11, 19]. However, we argue that knowing only the trustworthiness of advisors is not sufficient. For a complete picture, we also need to understand how the advisors behave when they are dishonest.

We formally characterize behavior pattern (or strategy) of an arbitrary advisor, whose properties are characterized by statistical parameters. We then calculate what behavior causes the worst-case attack, through solving an optimization problem, namely minimizing information leakage of a rating. We propose methods to exploit ratings under the worst-case attacks, aiming at improving the robustness of the existing trust models. In this chapter, honest advisors are assumed to be objective in rating, which means they report same ratings given a same observation.

[4] In this thesis, regarding unfair rating attacks, we treat robustness as a security issue of trust systems.

There are some notable theoretical contributions. First, we prove that in the worst case, even if the fraction of dishonest advisors is larger than the commonly asserted threshold 0.5, an advisor can still obtain information from ratings. Second, we prove that, even in the case where an advisor obtains zero information, dishonest advisors may still sometimes report the truth (observation). Third, we also prove that, for dishonest advisors, to minimize the information leakage of their observations and that of the integrity (or reliability) of targets, they need to perform different attack strategies.

Based on the explicit modeling of the worst-case attacks and also the formal theoretical analysis, we propose an induced trust computation method (ITC), which can ensure the accuracy of trust evaluation under the worst case. The simulation results demonstrate that under the worst case, ITC predicts either the integrity of targets or the observations of dishonest advisors, with much higher accuracy compared to three representative trust models: TRAVOS [10], BLADE [1] and MET [2]. To defend against unfair rating attacks, always assuming the worst case is a safe but may not always be the most accurate choice. Hence, we also compare the performance of ITC with TRAVOS, BLADE and MET under various weaker attacks. And it shows that ITC still has higher accuracy. All these results confirm that our method can effectively improve the robustness of the trust models.

4.1 The Worst Case: Minimizing Information Leakage

An advisee aims to learn (or obtain information) from ratings, based on which it makes decisions. Note that this does not simply mean the advisee would believe the ratings. The advisee can calibrate the interpretation of ratings based on the trustworthiness or the strategies of advisors, to make accurate decisions (e.g., accurate trust evaluation for trust systems). For example, BLADE proposes to re-interpret ratings based on the correlation between an advisor's ratings and the truth [1]. Therefore, whenever there is information in ratings, there can be a way to make use of it. The worst case for the advisee is: *there is little information in ratings, or dishonest advisors attempt to minimize that information*.

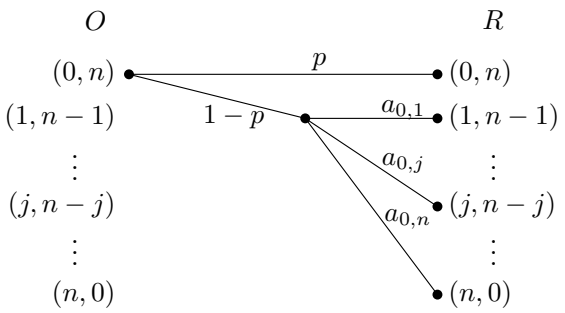


Figure 4.1: The naive rating model

Below, we quantify the worst case and analyze what kinds of rating strategies constitute that, based on the rating model introduced below.

4.1.1 Rating Model

There are advisors, advisees and targets under rating in a rating process. A typical example is rating-based reputation systems in e-marketplaces, which we use as backgrounds for our modeling. A rating process in an e-marketplace includes three kinds of entities: buyers, advisors, and sellers. Note that the role of buyer, advisor or even seller may be fulfilled by a same entity, e.g., considering C2C platform Taobao. We assume the success (or failure) of a transaction with a seller completely depends on the seller's integrity (or reliability, trustworthiness), which is represented by random variable T. We assume that T follows Bernoulli distribution. With probability $T{=}t$, a transaction is successful, and with probability $T{=}1{-}t$, a transaction is failed. Below, we analyze the worst-case ratings provided by advisors to a single buyer regarding a single seller.

We first consider a single advisor. The advisor reports its interaction history with the seller to the buyer. We assume that for the buyer, the number of interactions between the advisor and the seller is a known quantity, $n \in \mathbb{N}$; the only thing unknown is what fractions are successes and failures. The random variables O and R represent the true and the claimed interaction history of the advisor about the seller, respectively. O is decided only by T[5], while R depends not only on T, but also on the honesty and strategy of the advisor. We assume that from the buyer's perspective, before receiving R, the

[5]Observation is purely decided by the integrity of the seller.

truth O and T have the highest uncertainty, thus they are uniformly distributed based on the maximum entropy principle.

The advisor may not always report the truth to the buyer. We set a variable P to describe its honesty, namely with probability p the advisor is honest: $p{=}p(P{=}1)$, and with $1{-}p$ it's dishonest: $1{-}p = p(P{=}0)$. Honesty can refer to "free of deceit" as well as "truthful". We in this thesis interpret it as the former. Hence, dishonesty means that the advisor strategically provides ratings, and we treat dishonest advisors as attackers. Given an observation $O{=}(i, n-i)$, with i as the number of successful interactions, the probability that the advisor reports $R{=}(j, n-j)$ is $a_{i,j}$. For example, $a_{0,1}$ represents the probability that the advisor reports $R{=}(1, n-1)$ when $O = (0, n)$ is observed. As $R{=}(j, n-j), (j{=}0, 1, \ldots, n, j{\neq}i)$ constitutes all possible ratings when the advisor is dishonest, we have $\sum_{j\neq i} a_{i,j}{=}1$. Matrix $a_{i,j}$ decides the rating strategy of an advisor. We assume advisors are independent and the population is large enough, to treat probabilities $p, a_{i,j}$ of an arbitrary advisor as equal to the corresponding frequencies, e.g., p equals the ratio of honest advisors[6]. For simplicity, below we use $O{=}i$, $R{=}j$ to represent $O{=}(i, n-i)$ and $R{=}(j, n-j)$ respectively.

The set-up with a single advisor can be generalized to multiple advisors, with all of them assigned the same probabilistic parameters $n, p, a_{i,j}$. Here, p, $(1{-}p)$ can also be approximately treated as the rate of honest (dishonest) advisors. Also, $a_{i,j}$ can be treated as the rate of advisors reporting $R{=}j$ when $O{=}i$ is observed. In this way, our analysis for a single advisor is also explainable for multiple advisors.

We consider two types of worst-case unfair rating attacks performed by advisors: misbehaving advisors aiming at minimizing (hiding) the information of their true observations, and misbehaving advisors aiming at minimizing the information of the integrity of the seller.

[6]The honesty and strategy of each advisor are identically distributed random variables, following Bernoulli distributions with parameters p and $a_{i,j}$. According to Chernoff bounds [83], the difference between $p, a_{i,j}$ and their corresponding frequencies is inversely proportional to the number of advisors.

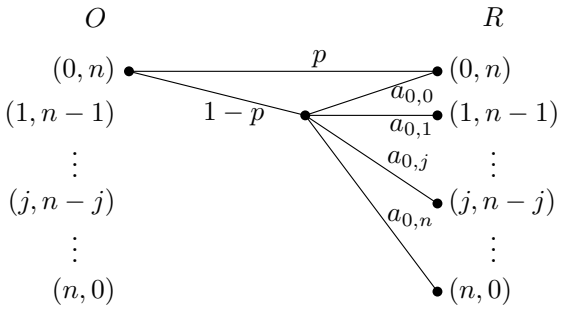


Figure 4.2: The extended rating model

4.1.2 Attackers Hiding their True Observations

This kind of attackers aims to hide their observations from the buyer. The rationale is that, by hiding the observations, attackers make it difficult for the buyer to construct an accurate trust opinion about the seller. In the worst case, an attacker can completely hide its observation, i.e., the rating is independent of the observation.

Theorem 4.1. *In the naive rating model shown in Figure 1, a rating R is independent of the observation O iff $p = \frac{1}{n+1}$ and $a_{i,j} = \frac{1}{n}(i \neq j)$.*

Proof. If R is independent of O, then $P(R{=}j|O{=}i) = P(R{=}j|O{=}i')$, for all j, i, i'.

$$P(R{=}j|O{=}i) = \begin{cases} p & \text{if } i = j \\ (1-p)a_{i,j} & \text{if } i \neq j \end{cases} \tag{4.1}$$

Therefore $p{=}(1-p)a_{i,j}$ and $a_{i,j}{=}\frac{p}{1-p}$, where $i{\neq}j$. Since $\sum_j a_{i,j} = 1$ where $i \neq j$, $n \cdot \frac{p}{1-p} = 1$ and $p{=}\frac{1}{n+1}$. As the result, $a_{i,j}{=}\frac{1}{n}$, where $i{\neq}j$.

On the other hand, if $p = \frac{1}{n+1}$ and $a_{i,j} = \frac{1}{n}$, $i \neq j$, then

$$P(R{=}j|O{=}i) = \frac{1}{n+1} \tag{4.2}$$

$$P(R{=}j) = \sum\nolimits_i P(O = i) \cdot P(R{=}j|O{=}i) = \frac{1}{n+1} \tag{4.3}$$

As $P(R{=}j|O{=}i) = P(R{=}j)$ holds for any i and j, R and O are independent. □

Intuitively, we expect that the lower values of p (more dishonest advisors) make it easier to hide O. However, Theorem 4.1 implies that when $p{<}\frac{1}{n+1}$ the observations

cannot be perfectly hidden, whereas for $p{=}\frac{1}{n+1}$, it can. Therefore, we need to alter the naive model to accommodate for the case $p{<}\frac{1}{n+1}$. When $p{<}\frac{1}{n+1}$, the independence of O and R implies $\sum_{j\neq i} a_{i,j}{<}1$, which is impossible in the naive model. This is caused by the fact that the advisor is forced to lie (with n fixed) if the advisor is strategical in the naive model. 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 tell the truth, as part of deceit. As a real-world scenario: consider 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 if he states Queen, and Alice has the King, Alice would know that Bob has the Ace. Thus, as a strategical player, Bob should sometimes report the truth to deceive Alice. Hence, here we introduce an alternative option $a_{j,j}$ (e.g., $a_{0,0}$ when $j{=}0$), as depicted in the extended rating model in Figure 4.2.

Theorem 4.2. *In the extended rating model shown in Figure 2, the rating R is independent of the observation O iff $0 \leq p \leq \frac{1}{n+1}$ and $a_{ij} = \frac{p}{1-p} + a_{jj}$.*

Proof. If rating R is independent of O, then $P(R{=}j|O{=}i) = P(R{=}j|O{=}i')$, for all j, i, i'.

$$P(R{=}j|O{=}i) = \begin{cases} p + (1-p)a_{i,j} & \text{if } i = j \\ (1-p)a_{i,j} & \text{if } i \neq j \end{cases} \tag{4.4}$$

Therefore $p + (1-p)a_{j,j} = (1-p)a_{i,j}$, $a_{i,j} = \frac{p}{1-p} + a_{j,j}$. Since $\sum_{j\neq i} a_{i,j} = 1 - a_{i,i}$, $\frac{np}{1-p} + \sum_{j\neq i} a_{j,j} = 1 - a_{i,i}$, we get $\sum_j a_{j,j} = \frac{1-(n+1)p}{1-p}$. Since $\sum_j a_{j,j} \geq 0$ and $0 \leq p \leq 1$, we get $0 \leq p \leq \frac{1}{n+1}$.

On the other hand, if $0 \leq p \leq \frac{1}{n+1}$ and $a_{ij} = \frac{p}{1-p} + a_{jj}$

$$P(R{=}j|O{=}i) = P(R{=}j) = p + (1-p)a_{j,j} \tag{4.5}$$

holds for any i, j. Hence, R and O are independent. □

When $\sum_j a_{j,j} = 0$ and $a_{i,j} = \frac{p}{1-p}$, Theorem 4.2 becomes Theorem 4.1. Note that $\sum_j a_{j,j}>0$ is allowed when R is independent of O, which implies even when the buyer learns nothing, still some dishonest advisors (attackers) may tell the truth.

Intuitively, ratings are only useful when less than half of the advisors are attackers. Remarkably, Theorem 4.2 proves otherwise. It implies that R and O cannot be independent when $p>\frac{1}{n+1}$. This means that, for $n>1$, in the case that over half of the advisors are dishonest (i.e., $(1-p)>\frac{1}{2}$), the buyer can still learn information from the ratings.

Although no strategy can achieve the independence when $p > \frac{1}{n+1}$, some strategies are still better at hiding the observations than others. To capture this, we generalize the measure of dependency between ratings and observations to information leakage (Definition 3.7 in chapter 3). The independence of R and O holds iff R leaks zero information about O. Low information leakage about O means that O is hidden well. Below, we aim to find the strategy that minimizes the information leakage for $p > \frac{1}{n+1}$. As $H(O)$ is unchangeable to the buyer, to minimize information leakage, it suffices to minimize $-H(O|R)$.

Definition 4.3. (Level strategy) is the strategy where: for all $0 \leq j \leq n$, $a_{j,j} = 0$, and for all $0 \leq i \neq j \leq n$, $a_{i,j} = \frac{1}{n}$.

Theorem 4.4. *The level strategy minimizes information leakage of O given R for $p \geq \frac{1}{n+1}$.*

Proof. Given $h_j = p + (1-p)\sum_i a_{i,j}$, $0 \leqslant i, j \leqslant n$,

$$
\begin{aligned}
&-H(O|R) = \sum\nolimits_j P(R{=}y_j) \sum\nolimits_i P(O{=}x_i|R{=}y_j) \log P(O{=}x_i|R{=}y_j) \\
&=^1 \frac{1}{n+1} \sum\nolimits_j \Big(\sum_{i \neq j} (1-p) \cdot a_{i,j} \log(\frac{(1-p) \cdot a_{i,j}}{h_j}) \\
&\quad + (p + (1-p)a_{j,j}) \log(\frac{p + (1-p)a_{j,j}}{h_j}) \Big) \qquad (4.6) \\
&= \frac{1}{n+1} \sum_j \sum_{i \neq j} h_j \cdot f(\frac{(1-p) * a_{i,j}}{h_j}) \qquad (4.7) \\
&\quad + \frac{1}{n+1} \sum_j h_j \cdot f(\frac{p + (1-p) * a_{j,j}}{h_j}) \qquad (4.8) \\
&\geq^2 \frac{n}{n+1} \sum\nolimits_i \frac{(1-p)(1-a_{i,i})}{n} \log(\frac{(1-p)(1-a_{i,i})}{n}) \\
&\quad + (p + \frac{\sum_j (1-p) \cdot a_{j,j}}{n+1}) \cdot \log(p + \frac{\sum_j (1-p) \cdot a_{j,j}}{n+1}) \\
&\geq^3 p \cdot \log(p) + (1-p) \cdot \log(\frac{1-p}{n})
\end{aligned}
$$

Inequality 2 is derived based on the Jensen's inequality (Theorem 3.9 in Chapter 3). Inequality 3 is derived based on the property that $xlog(x)$ is superlinear and $p \geq \frac{1}{n+1}$.

Finally, note that applying the strategy from Definition 4.3 to term 1 yields term 3. Thus, term 3 represents the information leakage under the level strategy. Since term 3 is the minimum, the level strategy minimizes information leakage. For $p = \frac{1}{n+1}$, the level strategy leads to zero information leakage, as we proved in Theorem 4.1. □

In summary, we have found the strategies that minimize the information leakage about O for all $p \in (0,1)$. Specifically, for $p < \frac{1}{n+1}$, the strategy requires a fraction of dishonest advisors to report the truth. For $p \geq \frac{1}{n+1}$, the strategy requires dishonest advisors to uniformly choose a lie. Further, zero information leakage (independence) is only achieved when $p \leq \frac{1}{n+1}$. We may name attacks that cause zero information leakage as ultimate attacks throughout this thesis. The buyer can still get some information for $p > \frac{1}{n+1}$.

To illustrate our results, we plot the information leakage of O in the worst-case attacks (minimized information leakage), as a variable of p (with $n = 5$) and n (with $p = 0.25$), in Figure 4.3. From the figure, we learn that when $p \leq \frac{1}{n+1}$ or $n \leq \frac{1}{p} - 1$,

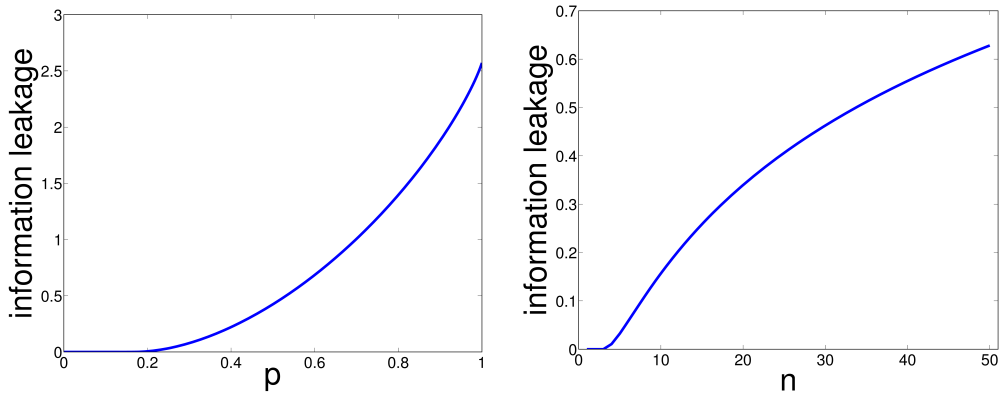


Figure 4.3: The minimal information leakage of O varies with p and n

the information leakage is zero. And when the difference between p and $\frac{1}{n+1}$ increases, the information leakage increases. This will further be demonstrated in the simulation in Section 5.

4.1.3 Attackers Hiding the Integrity of the Seller

This kind of attackers aims to hide the integrity, T, of the seller from the buyer. The rationale is that the buyer's trust opinion is about the integrity of the seller. Therefore, to make it difficult for the buyer to construct an accurate trust opinion about the seller, the advisor aims to hide information about the integrity of the seller.

Intuitively, hiding the observations (which correspond to interaction histories in this chapter) may seem equivalent to hiding the integrity of the seller. As we will prove in Theorem 4.6, they are not the same. However, they do coincide whenever they can avoid any information leakage.

Theorem 4.5. *There is zero information leakage of T, iff there is zero information leakage of O.*

Proof. From Proposition 3.4 in Chapter 3, zero information leakage of T (O) given R is equivalent to T (O) being independent of R. If O is independent of R, we have

$$\begin{aligned} P(T{=}t|R{=}j) &=^1 \sum\nolimits_i P(T{=}t|R{=}j, O{=}i) \cdot P(O{=}i|R{=}j) \\ &=^2 \sum\nolimits_i P(T=t|O=i) \cdot P(O=i|R=j) \\ &=^3 \sum\nolimits_i P(T=t|O=i) \cdot P(O=i) \\ &=^4 P(T=t) \end{aligned} \tag{4.9}$$

which holds for any t, j, implying that T is independent of R. Term 2 follows because T and R are conditionally independent given O.

On the other hand, if T is independent of R, we have

$$\begin{aligned} P(O{=}i|R{=}j) &=^1 \sum\nolimits_i P(O{=}i|T{=}t, R{=}j) \cdot P(T{=}t|R{=}j) \\ &=^2 \sum\nolimits_i P(O=i|T=t) \cdot P(T=t|R=j) \\ &=^3 \sum\nolimits_i P(O=i|T=t) \cdot P(T=t) \\ &=^4 P(O=i) \end{aligned} \tag{4.10}$$

which holds for any i, j, implying that O is independent of R. Term 2 follows because O and R are conditionally independent given T. Thus we prove Theorem 4.5. □

Note that since zero information leakage of O requires $p \leq \frac{1}{n+1}$, zero information leakage of T also requires $p \leq \frac{1}{n+1}$.

Theorem 4.6. *The level strategy does not minimize information leakage of T, for all n, p that satisfy $p > \frac{1}{n+1}$.*

Proof. It suffices to provide a counterexample. For $n = 2$, $p = \frac{2}{3}$, using the level strategy, we obtain $-H(T|R) = 0.2192$. When we set

$$a = \begin{pmatrix} 0 & 0.2938 & 0.7063 \\ 0.4922 & 0.0156 & 0.4922 \\ 0.7063 & 0.2938 & 0 \end{pmatrix},$$

$-H(T|R) = 0.1934$. Since $0.1934 < 0.2192$, the level strategy does not minimize information leakage of T. □

Below, we aim to find the strategy that minimizes information leakage of T given R when $p > \frac{1}{n+1}$. As $H(T)$ is unchangeable, it suffices to minimize $-H(T|R)$.

$$\begin{aligned} -H(T|R) &= -\sum\nolimits_j P(R=j)H(T|R=j) \\ &= \sum\nolimits_j P(R=j)\int_0^1 f_T(t|R=j)\cdot\log f_T(t|R=j)\,\mathrm{d}t, \end{aligned}$$

where $P(R=j)$ as before, and

$$\begin{aligned} f_T(t|R=j) &= \sum\nolimits_i f_T(t|O=i,R=j)\cdot P(O=i|R=j) \\ &= \sum\nolimits_i f_\beta(t;i{+}1,n{-}i{+}1)\cdot \begin{cases} \frac{p+(1-p)a_{j,j}}{h_j} & \text{if } i=j \\ \frac{(1-p)a_{i,j}}{h_j} & \text{if } i\neq j \end{cases} \end{aligned} \tag{4.11}$$

Note that $P(O=i|R=j)$ is the posteriori probability about O known R, which can be computed from $P(R=j|O=i)$ based on Bayes' theorem. And $P(R=j|O=i)$ is decided by the rating strategy.

For our analysis, we use a local search heuristic to find good strategies for the advisors. Our heuristic is initialized with the level strategy. We iterate over all $a_{i,j}$, where, for each $a_{i,j}$, we increase $a_{i,j}$ with a fixed value (at the expense of the other $a_{i,j'}$) until $-H(T|R)$ stops decreasing. We perform the iteration multiple times, with decreasing step sizes. In the limit, the heuristic is a gradient search.

To illustrate the analysis above for T, we plot the information leakage of T in the worst case, as a variable of p (with $n=5$) and n (with $p=0.25$), in Figure 4.4. From the figure, we learn that when $p \leq \frac{1}{n+1}$ or $n \leq \frac{1}{p}-1$, the information leakage is zero. And when the difference between p and $\frac{1}{n+1}$ increases, the information leakage increases. This will also be further demonstrated in the simulations in Section 5.

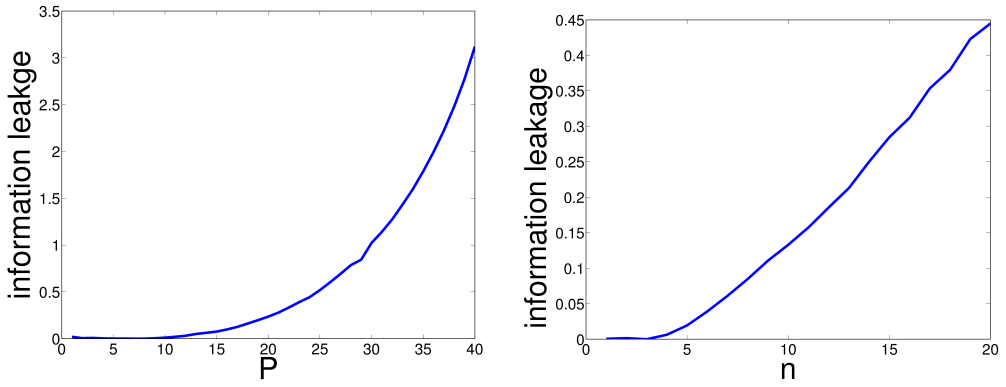


Figure 4.4: The minimal information leakage of T varies with p and n

4.1.4 Induced Trust Computation (ITC)

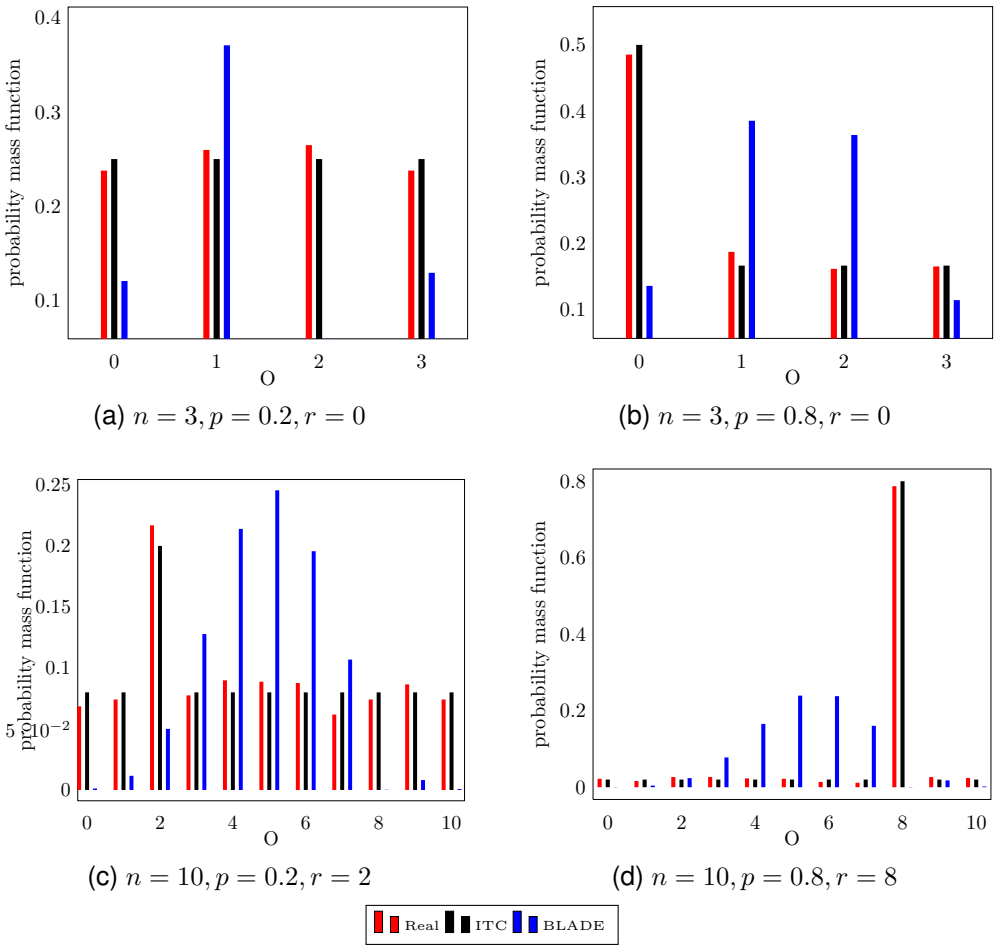
Given the strategies of the advisors, the buyer can construct accurate trust opinions under the worst-case attacks. Here, we propose an induced trust computation method, which assumes the worst-case attack strategies. A trust opinion is a distribution $f_T(t|\phi)$, where ϕ consists of the knowledge of the buyer (direct experiences and ratings) [84].

Muller and Schweitzer [85] prove the following theorem under the assumptions that if ratings and observations are conditionally independent given the strategies of the sellers and advisors, and that their strategies are independent:

Theorem 4.7. *For any collection of ratings and direct observations* φ *and* ψ, $f_T(t|\varphi,\psi) \propto f_T(t|\varphi) \cdot f_T(t|\psi)$.

With Theorem 4.7, the knowledge of the buyer can be broken down into cases for which we have explicit computations. The case where the knowledge of the buyer is direct experience, has already been solved [84]. If the knowledge of the buyer is a single rating, then $\phi = R$ and the trust opinion is $f_T(t|R)$. In the worst-case attack, $f_T(t|R)$ can be computed known the strategy (matrix $a_{i,j}$) of the advisors based on Equation (4.11). In this chapter, we follow the assumption of Theorem 4.7. And the integrity of a seller can be computed based on $f_T(t|R)$.

Note that the accuracy of computing $f_T(t|S)$ is influenced by the accuracy of p. As a description of the trustworthiness of an advisor, p is usually estimated by the trust models (as done by TRAVOS [10] and many other classic models [11, 19]). *In this work, we are not trying to build a new robust trust model. We are solving a subproblem of*



(a) $n = 3, p = 0.2, r = 0$ (b) $n = 3, p = 0.8, r = 0$

(c) $n = 10, p = 0.2, r = 2$ (d) $n = 10, p = 0.8, r = 8$

Figure 4.5: Comparing predictions on distributions of O

defending the worst-case unfair rating attack to make a trust model more robust. Hence, we simply assume that p is already accurately estimated by the trust models. In fact, we also demonstrate through simulation in Section 5 that even when p is not entirely accurately estimated by the trust models (e.g., TRAVOS and MET), their robustness can still be improved by our ITC method.

In this way, by being aware of the worst-case strategies in advance, the buyer gains the initiative to derive accurate trust opinions under the worst case.

4.2 Robustness Analysis

As surveyed in Chapter 2, TRAVOS [10], BLADE [1] and MET [2] are the three typical trust models to address the unfair rating problem, where TRAVOS and BLADE assume

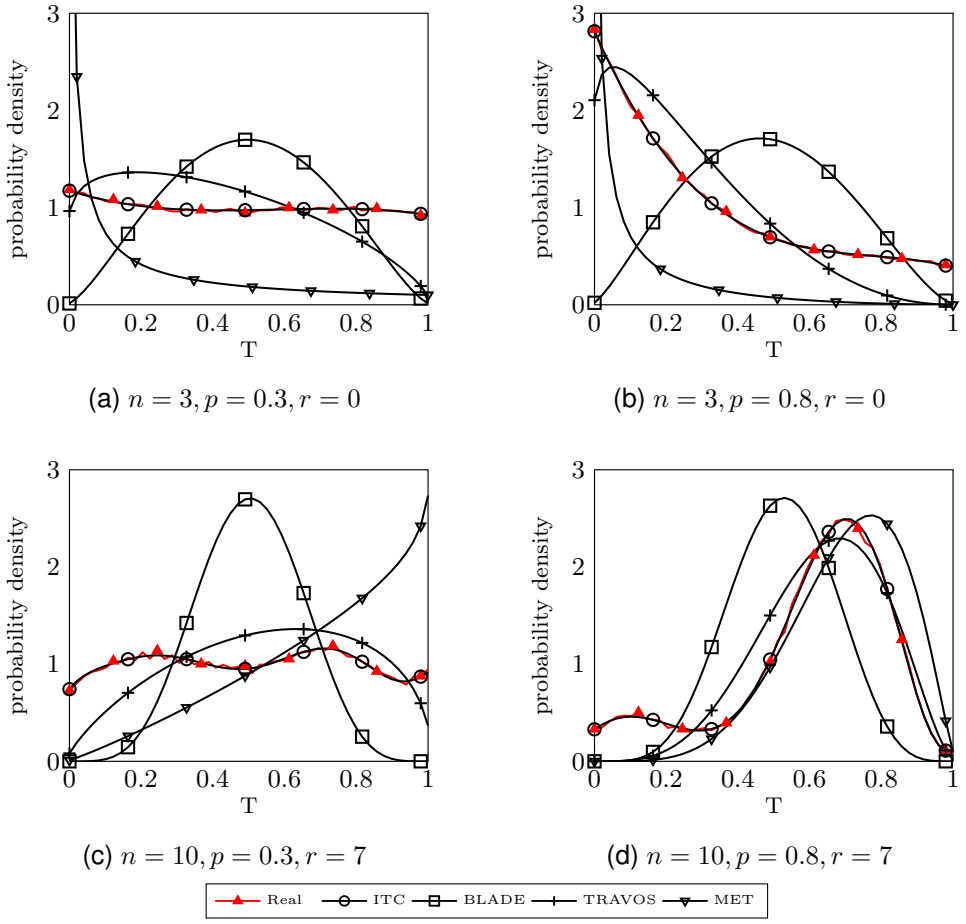


Figure 4.6: Comparing predictions on distributions of T

some simple attack strategies for advisors but MET tries to cope with some typical attacks and their combinations. In this section, we evaluate the robustness of these trust models, and more importantly to demonstrate that our induced trust computation (ITC) method can further improve the robustness of these trust models.

More specifically, we conduct a set of simulations[7]. In the first simulation, we compare the trust opinions about sellers that the trust models and ITC construct, under two types of the worst-case attacks: advisors hiding true observations (O) and hiding seller integrity (T). Because ITC always assumes the worst case, to have more fair comparison, in the second simulation, we compare the accuracy of trust opinions given by ITC and the three models, under other random attack strategies which are not the worst case. Modeling the honesty of advisors accurately is not the focus of this work.

[7]We did not use existing testbeds such as the ART testbed [86] because they are often only used to study the quality of expectations about trust evaluation.

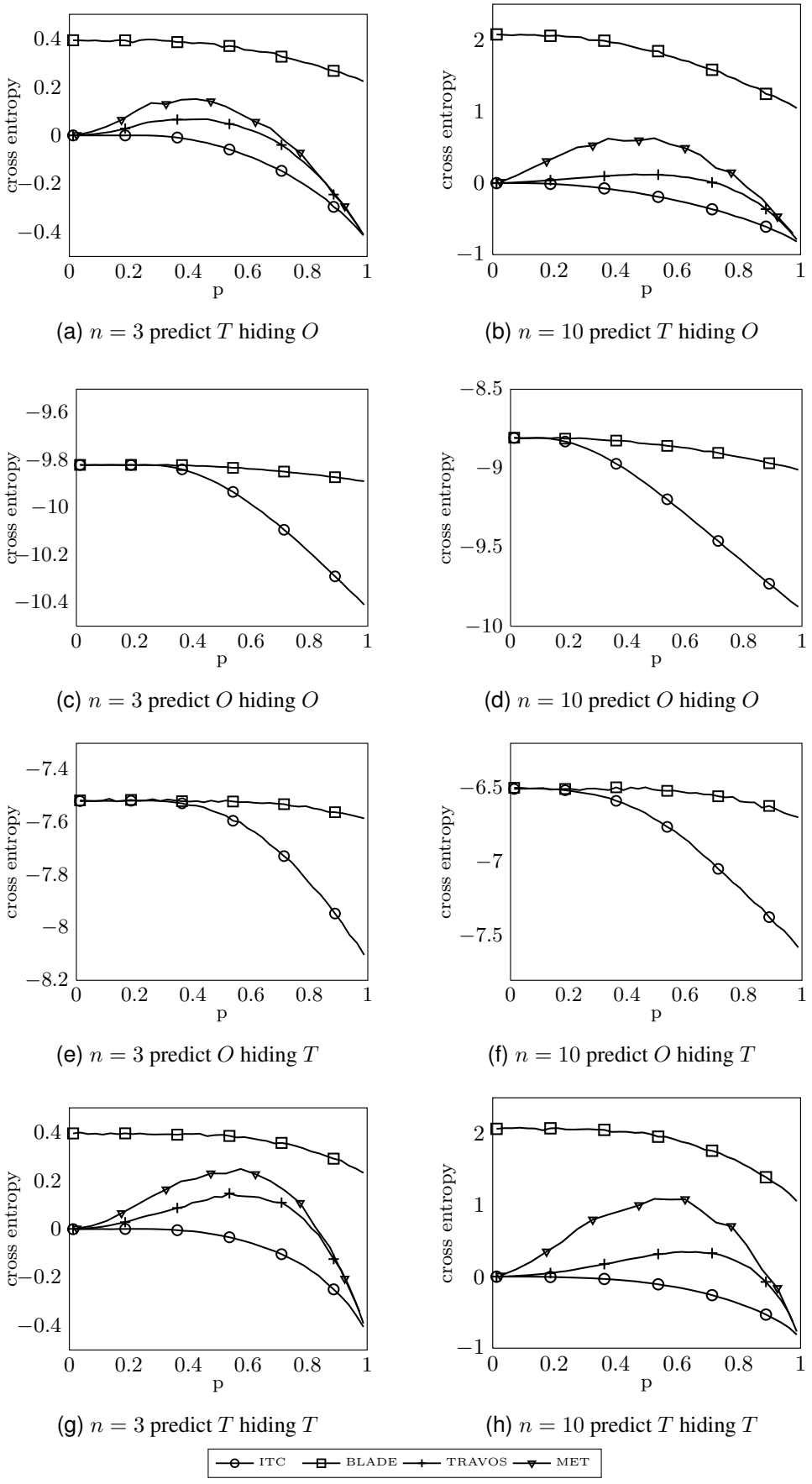
Hence in the first two simulations, we simply assume that the honesty of advisors is accurately estimated by all trust models, which is set as a same parameter (p) to all models. In the third simulation, we further study whether our ITC method can improve the robustness of the trust models given whatever output of advisor honesty by these trust models.

All of the simulations above rely on the true behavior of the seller. To address this, we run the Monte Carlo simulation. In each run, $t \in [0, 1]$ is uniformly randomly chosen as a sample of T for the seller. Then, n Bernoulli samples are drawn with the probability t, which provides us an o as the true value of O. Based on o and an advisor's strategy a, a rating r is generated as the true value of R (ratings of the advisor). The trust models are provided with rating r, which is used to construct the trust opinion about the seller.

4.2.1 Under the Worst Case

In the first simulation, we compare the predictions on T and O against the truth, under the worst-case strategies of hiding T and O respectively. The values for parameters n, p, r^* are manually chosen as the number of transactions, probability of advisor honesty, and rating. We then run the simulation, but reject the sample of T (and the corresponding sample of O) if the resulting $R{\neq}r^*$. In this way, we get the true probability distributions of T and O: $P(T|R{=}r^*)$ and $P(O|R{=}r^*)$, which are used to compare to that predicted by TRAVOS, BLADE, MET and ITC.

For comparison about O, we select four groups of values for n, p, r^*: $(3, 0.8, 0)$, $(3, 0.2, 2)$, $(10, 0.8, 8)$, $(10, 0.2, 2)$. Figure 4.5 presents the results. TRAVOS and MET are not considered here, as they do not generate the prediction of O. The predictions of ITC have much smaller difference with the real distributions compared with BLADE. Larger p leads to more converged predictions. Comparing Figures 5(a) and 5(c), although $p{=}0.2, s{=}2$ are the same, prediction of ITC given $n{=}10$ is converged on $O{=}2$ while that given $n{=}3$ is uniformly distributed. According to the theoretical proof in the former section, when $n{=}3$, $p{<}\frac{1}{n+1} = 0.25$, there is no information leakage of O under the worst-case attack. Hence, ITC predicts maximum uncertainty of O.



(a) $n = 3$ predict T hiding O

(b) $n = 10$ predict T hiding O

(c) $n = 3$ predict O hiding O

(d) $n = 10$ predict O hiding O

(e) $n = 3$ predict O hiding T

(f) $n = 10$ predict O hiding T

(g) $n = 3$ predict T hiding T

(h) $n = 10$ predict T hiding T

Figure 4.7: Comparing accuracy of predicting T (or O) under the worst-case of hiding T (or O)

For comparison about R, we select four groups of values for parameters n, p, r^*: $(3, 0.8, 0)$, $(3, 0.3, 0)$, $(10, 0.8, 7)$, $(10, 0.3, 7)$. Figure 4.6 presents the results. The probability distributions of T predicted by ITC are much closer to the real distributions than that of TRAVOS, BLADE and MET. For ITC, TRAVOS and MET, the shapes of predicted distributions are mainly decided by p and r, while BLADE is largely influenced by n instead. Comparing Figures 6(a,c) with 6(b,d), larger p leads predictions of ITC and TRAVOS to be more converged and aligned with the ratings, because the buyer tends to believe the advisor more.

Figures 4.5 and 4.6 are restricted to a fixed r, n, p and a. To make more meaningful comparisons, we use cross entropy (Definition 3.6 in Chapter 3) to measure the quality of a prediction so that we can compare a multitude of outcomes simultaneously. In a good prediction, cross entropy is low. We generate a true integrity of a seller t, a true observation o and a rating r in each run, and r is used as input for the models to yield a trust opinion about the seller. To generate the graphs, we let $n = 3$ and $n = 10$, and let $0 < p < 1$ be the x-axis. We study four scenarios: predicting O (T) under the worst-case strategies of hiding O (T), and predicting T (O) under the worst-case strategy of hiding O (T). Because TRAVOS and MET do not output predictions of O, they do not appear in Figure 4.7 (c-f). Figure 4.7 provides the following information.

First, when $p \leq \frac{1}{n+1}$, all the ITC graph segments are flat, meaning that uniform distribution is predicted. This corroborates our proofs: when $p \leq \frac{1}{n+1}$, there is no information leakage about T (O) given R in the worst case, thus $H(T|R)$ (or $H(O|R)$) reaches the maximum, which implies uniform distribution. Note that for continuous distributions, the uniform distribution has entropy zero, explaining why ITC has cross entropy of 0 for small p, in Figures 4.7 (a, b, g, h). In Figures 4.7 (c-f), the uniform distribution is over discrete variables, meaning that the entropy depends on n, which explains the difference in cross entropy for value p near 0.

Second, when $p > \frac{1}{n+1}$, ITC shows lower cross entropy than BLADE, TRAVOS and MET, for equal p and n. Moreover, we can identify the trends that BLADE and ITC have decreasing cross entropy over p (and n), whereas for TRAVOS and MET, the cross entropy increases before it decreases, over p. The reason that ITC is decreasing, is simply because $H(T|R)$ (and $H(O|R)$) are decreasing over p. Recall (Definition 3.6)

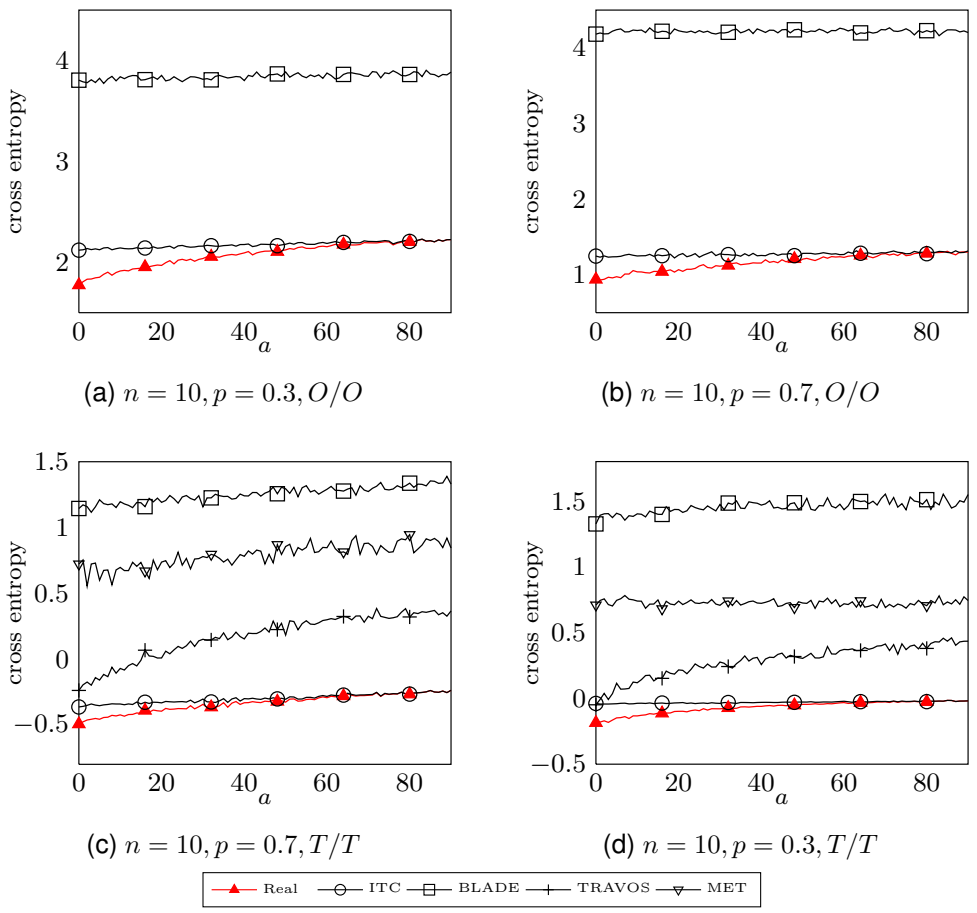
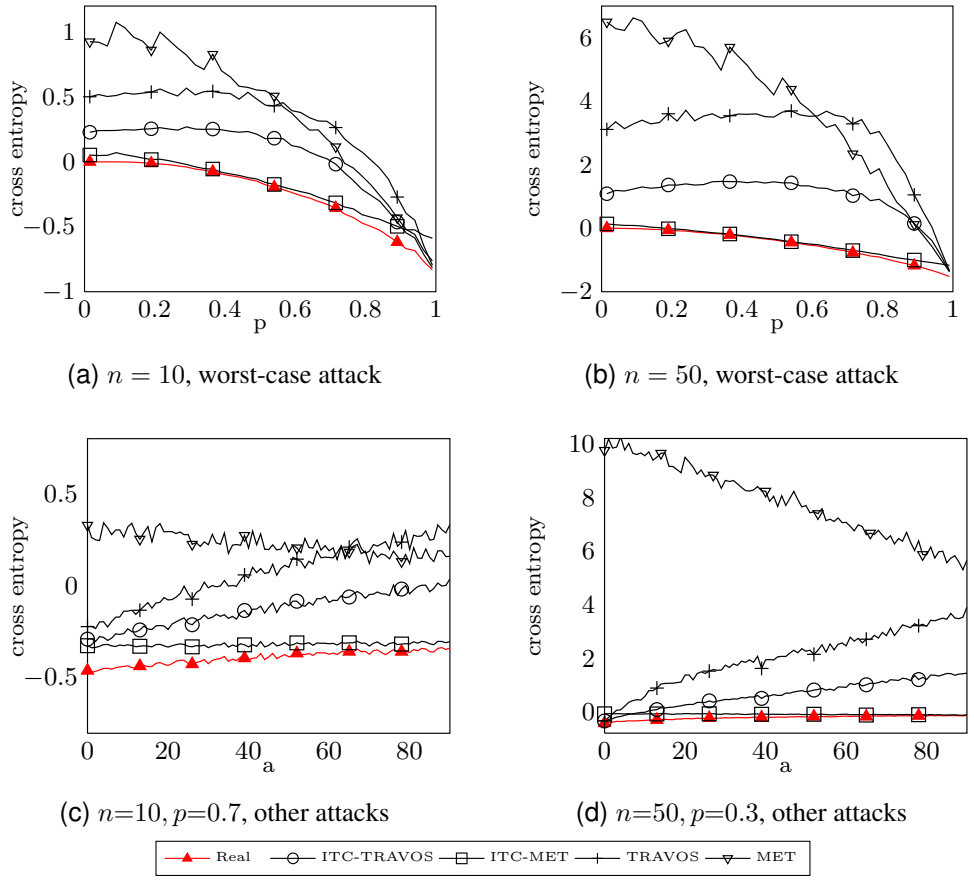


Figure 4.8: Under various other types of attacks. O/O: predict O by hiding O; T/T: predict T by hiding T

that cross entropy is the entropy of the truth plus KL-divergence, and that ITC has KL-divergence of 0, because it computes correct $H(T|R)$ (and $H(O|R)$) by knowing p and the worst-case strategies of advisors. TRAVOS and MET first increase because they over-predict – causing to assign unreasonably low probability to unlikely events (as shown in Figure 4.6). As p tends to 1, their over-predictions start to match the true distribution. BLADE suffers the same problem of over-predicting. However, its over-predicting is not linked with p. Therefore, we observe a decreasing cross-entropy, as reality tends towards more polarised outcomes. Note that using the same real p value, the accuracy of TRAVOS is higher than MET, indicating that the method of aggregating ratings in TRAVOS is better than that of MET under the worst-case attack. In fact, MET adopts a simple weighted average method to aggregate advisors' ratings.

Third, when p is close to 1, the curves of TRAVOS, MET and ITC with the same n



(a) $n = 10$, worst-case attack

(b) $n = 50$, worst-case attack

(c) n=10, p=0.7, other attacks

(d) n=50, p=0.3, other attacks

Figure 4.9: Using p estimated by other trust models under the worst-case and other types of attacks

get to converge at a same point. With p being close to 1, nearly all of advisors report the truth. The predictions of TRAVOS and MET thus get closer to the truth, which is the prediction of ITC.

From the analysis above, it is obvious that our prediction of T (O) is much more accurate than TRAVOS, BLADE and MET under the worst case. Using our ITC method could improve the robustness of these trust models.

4.2.2 Under other Attacks

The real strategies of advisors cannot be known. To always assume the worst case is a safe choice, but may not be the most accurate choice. Hence, we investigate the performance of ITC, which assumes the worst-case strategies, under other types of attacks. Recall that a strategy of an advisor is represented as a matrix $a_{i,j}$ where $0 \leq i, j \leq n$

(see Section 4.1.1). We randomly generate ninety such strategies. Then, these strategies are combined with the worst-case strategy by assigning the worst case a weight varying from 0 to 1. In so doing, the strength of the resulting strategies approximately increases. We then compare the cross entropy regarding the predictions of T (O) given by ITC, TRAVOS, BLADE and MET, under all of these strategies. Figure 4.8 presents the results.

For the truth (the red line), the cross entropy is equal to the entropy of true distribution of T (O) given R because KL-divergence is 0. As the rating strategy tends to be worse, the entropy of T (O) given R increases towards the maximum, which is exactly the worst case. In Figure 4.8, ITC has much smaller cross entropy with the truth, compared to the three models, indicating that ITC predicts much closer to the truth. And as the generated attack strategy gets closer to the worst case, ITC predicts more and more accurately. Notice that there is little variance in the cross entropy of BLADE and MET as the attack strategies change, implying that their performance does not change much for all those strategies. On the other hand, the cross entropy of TRAVOS increases as the attack strategy gets closer to the worst case, showing that the performance of TRAVOS gets worse as the attacks become stronger.

From this simulation, even always assuming the worst case, our ITC method can still improve the robustness of the trust models against various other types of attacks.

4.2.3 Inaccurate Estimation of Advisor Honesty

The above simulations are conducted by assuming the accurate estimation of advisor honesty (i.e., true p). In this simulation, we investigate how ITC performs when p is predicted by other trust models, which may not be completely accurate. BLADE does not estimate p, so we only compare the accuracy of ITC (ITC-TRAVOS and ITC-MET) with TRAVOS and MET, based on their predicted p respectively. We consider the prediction of seller integrity T under two scenarios: 1) the worst-case strategy of hiding O, with the real p value varying from 0 to 1 (Figure 4.9 (a-b)); 2) other types of attacks with $p = 0.7$ and $p = 0.3$ (Figure 4.9 (c-d)).

Based on p predicted by the corresponding trust models, ITC still has much higher accuracy indicated by the lower cross entropy of ITC-TRAVOS and ITC-MET as shown in the figure, confirming that ITC can effectively improve the robustness of TRAVOS and MET even when the estimation of p may not be entirely accurate, and when the advisor attack strategies may not be the worst case.

Similar as the results in Figure 4.7, larger p leads to more accurate prediction because the advisor is more trustworthy. In addition, when the estimation of p is more accurate, the prediction of seller integrity T should also be more accurate. With this, compare ITC-TRAVOS and ITC-MET. ITC performs better when using p output by MET than when using p from TRAVOS, indicating that MET predicts the honesty of advisors more accurately than TRAVOS. This is also supported by the results in Jiang et al. [2]. However, with the p value from MET, ITC cannot accurately predict the truth even under the worst case (see Figure 4.9 (a-b)), indicating that advisor honesty estimated by MET is not completely accurate.

On the other hand, TRAVOS performs better than MET when $p<0.6$ in Figure 4.9 (a-b,d). Also, recall the results in Figure 4.7 where given the same true p, the predictions of TRAVOS are more accurate than MET. These results indicate that TRAVOS has a nice method for aggregating ratings from the advisors. However, when $p>0.6$, MET outperforms TRAVOS, indicating that when the advisors are more trustworthy, the effect of that method becomes less important. This can also be observed from Figure 4.9 (c) that when $p=0.7$ and under the worst-case attack (attack #90), MET provides more accurate prediction than TRAVOS. In fact, for other types of attacks that are close to the worst case, MET also outperforms TRAVOS.

4.3 Summary

In this work, we used information theory to measure how helpful ratings are to advisees that receive them. A fraction of advisors giving ratings is dishonest: attackers. We identified and analyzed which attack strategies reduce the overall helpfulness of ratings.

Our techniques and results can increase the robustness of existing trust models against unfair rating attacks.

We introduced two information theoretic measures for the quality of a rating, concerning how much a rating by an advisor reveals about the true observations of that advisor and about the true integrity of the trustee, respectively. We find that the two measures coincide iff ratings reveal nothing; that the ratings cannot always reveal nothing, even with more attackers than honest advisors; and that it may be rational for an attacker to report the truth, to obscure the truth.

We derived how to compute trust opinions, assuming the worst-case attack strategies. The results of our simulations show that our method's predictions are more accurate than TRAVOS, BLADE and MET, meaning our method is more robust, and more importantly that our method complements the trust models in improving their robustness.

Chapter 5

Collusive and Static Unfair Rating Attacks

In Chapter 4, we studied the type of unfair rating attacks where attackers are assumed to be independent, based on information theory. In this chapter, we study collusive unfair rating attacks (CUR) attacks [87]. In CUR attacks, attackers coordinate on a shared strategy to achieve a same goal. There can be various ways of colluding. Trust systems designers sometimes verify the robustness only under specific CUR attacks [20, 26–28]. 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 favor of the proposed system.

We argue that if a system cannot function well under the strongest attack, then it is not robust. Otherwise, if a group of systems is merely robust to some given CUR attacks, which they are tested against, then we need to be able to compare the strength of these attacks. In both the cases, we need to measure the strength of CUR attacks.

In Chapter 4, we introduced information theory to identify and measure the worst-case (strongest) attacks by independent attackers. Here, we focus on how strong arbitrary attacks are (also using information theory), which allows us to reason about the

$O \backslash R$	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 5.1: Strategy matrix of the colluders from Example 5.1

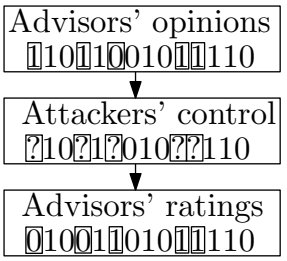


Figure 5.1: Rating modeled as channel

quality of the validation of a trust system. Moreover, we shift focus from independent attackers to colluding attackers (including Sybil attackers). The formalism from Chapter 4 must be fundamentally altered to allow for measurements of coalitions of attackers.

We extend the methodology from Chapter 4 to: (1) quantify and compare CUR attacks found in the literature (Section 5.2), (2) quantify types of CUR attacks (Section 5.3), and (3) identify the strongest possible CUR attacks.

Doing this, we found (1) attacks designed in the literature are not suitable to stress-test the robustness, (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. We consider the results from [82] explicitly as a special case of CUR attacks.

5.1 Modeling CUR Attacks

From an information theoretic perspective, advisors can be seen as a (noisy) channel. See Figure 5.1. The opinions of the advisors are being outputted 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 behavior. The difference in noise-per-advisor is often ignored in the literature, potentially skewing analysis of the attacker-resistance.

In Chapter 4, we proposed to quantify how much an advisee can learn as the information leakage of the ratings, which can then be used to measure the strength of the corresponding attacks. However, only the strongest independent attackers are considered there. In this chapter, we extend the quantification method to cover 1) collusive attackers, and 2) non-strongest attacks. To measure the strength of a CUR 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 5.1. *Consider a rating scenario with* 4 *advisors,* 2 *of which are colluding attackers*[8]*. Hence we make the reasonable assumption in this work that the number or percentage of colluders is known. Take the perspective of an advisee; 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 advisee received four ratings, three positive and one negative.*

We model the attackers' strategy in this example with the matrix in Table 5.1. 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 advisee 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 advisee has about the observations can be represented using joint entropy (Definition 3.5): $H(O_1, O_2, O_3, O_4)$. The

[8]Note that there exist many methods which can estimate the number of colluders in a system [88, 89].

joint entropy expresses the uncertainty associated with these four observations. Given the ratings, the information the advisee 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:

$$H(O_1, O_2, O_3, O_4) - H(O_1, O_2, O_3, O_4|R_1, R_2, R_3, R_4) \tag{5.1}$$

The conditional entropy of observations given the ratings is:

$$H(O_1, O_2, O_3, O_4|R_1{=}1, R_2{=}1, R_3{=}1, R_4{=}0) \tag{5.2}$$

$$= -\sum_{o_1,o_2,o_3,o_4} \mathbf{f}(p(o_1, o_2, o_3, o_4|R_1{=}1, R_2{=}1, R_3{=}1, R_4{=}0)) \tag{5.3}$$

$$= -\sum_{o_1,o_2,o_3,o_4} \mathbf{f}(\sum_{i,j} p(C_2{=}(i, j)|R_1{=}1, R_2{=}1, R_3{=}1, R_4{=}0) \tag{5.4}$$

$$\cdot\, p(o_1, o_2, o_3, o_4|R_1{=}1, R_2{=}1, R_3{=}1, R_4{=}0, C_2{=}(i, j))) \tag{5.5}$$

$$= -\frac{1}{2}\log(\frac{1}{12}) \approx 1.79 \tag{5.6}$$

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. Therefore, by Definition 3.7, 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 advisees. 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 who have observed $\overline{o}$ report 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

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

TABLE 5.2: Strategy matrix of general collusion attacks

Table 5.2. 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

$$H(\overline{O}) - H(\overline{O}|\overline{R}). \tag{5.7}$$

The conditional entropy of observations given ratings is as follows:

$$H(\overline{O}|\overline{R}) = -\sum_{\overline{r}} p(\overline{r}) \cdot H(\overline{O}|\overline{r}) \tag{5.8}$$

$$= -\sum_{\overline{r}} p(\overline{r}) \sum_{\overline{o}} \mathbf{f}(p(\overline{o}|\overline{r})) \tag{5.9}$$

$$= -\sum_{\overline{r}} p(\overline{r}) \sum_{\overline{o}} \mathbf{f}(\sum_{c} p(c|\overline{r}) \cdot p(\overline{o}|\overline{r}, c)) \tag{5.10}$$

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

5.2 Quantifying CUR Attacks

Collusive unfair rating attacks have been studied by many researchers [2, 20, 27, 28, 90]. They propose various methods to counter such attacks, e.g., detection based approaches [26], incentive based approaches [90], and defense-mechanism design based approaches [2, 28]. Based on simulations or experiments, they verify how effective their methods are in minimizing the impact of collusion attacks, namely how robust their models or systems are. Sometimes such verification only covers specific attacks that are assumed or designed by the researchers themselves [20, 26–28]. To choose

specific attacks to compare the robustness of two models is not convincing. One that fails under these attacks may behave better for some other attacks.

We argue that to equitably compare the robustness of two systems, we need to compare the strength of attacks that they are tested against. A system should be considered more robust if it can resist 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 [20] propose to mitigate the influence of unfair ratings by helping advisees 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 with 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[9].

The "FIRE+" trust model is proposed in [26]. FIRE+ aims to detect and prevent CUR attacks. It considers two kinds of CUR attacks: in the first type, advisors collude with the target under evaluation, providing false positive ratings to promote the target as reputable. 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[10].

The authors in [27] design a SocialTrust mechanism to counter the suspicious collusion attacks, the patterns of which are learned from an online e-commerce website.

[9]The computation of these values is omitted, as their generalization is provided in Theorem 5.2.
[10]The computation of these values is omitted, as their generalization is provided in Theorem 5.3.

Instead of filtering collusive unfair ratings or preventing collusion behaviors, 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 trusted nodes and 30 attackers. Based on Theorem 5.2, the information leakage is 5.96 bits[1].

The authors in [28] aim to design a reliable reputation system 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 5.2, the information leakage is between 31.33 bits (for 50% attackers) to 75.97 bits (for 10% attackers)[1].

In summary, the CUR attacks in the literature above are basically ballot-stuffing and lying. From the quantification, we get following results. First, referring to [20, 28], the information leakage of ballot-stuffing CUR attack decreases with the increase of the percentage of attackers. Second, 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 to 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.

5.3 Quantifying Types of CUR Attacks

In the papers we discussed above, only specific CUR attacks are considered in the verification of a system. 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 type of attacks (and if it is a range, we identify the strongest attack within the type). To verify the robustness of a system to a type of attacks, we propose to test it against the strongest attack in that type.

We summarize various types of CUR 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 CUR 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 (bad-mouthing).

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 type of attacks are a special case of the collusion attacks, namely those that coincide with the independent attacks. The second type of attacks are commonly found in the literature, where all attackers are either ballot-stuffing or bad-mouthing (see, e.g., [20, 27, 28]). There are also papers considering the other two types of attacks [90]. The last type of attacks are interesting, since it covers the three preceding types, as well as all other conceivable types.

We use the general CUR attack model from Section 5.1 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 5.1. *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 5.2, 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 $\alpha_{k,h,y,-} = \frac{\binom{h}{k} \cdot y}{\binom{m}{k} \cdot 2^{m-k}}$ 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 5.2. *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 3.3), 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 5.3. *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. □

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 5.4. *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 3.9):

$$\sum_{\overline{o}:\delta(\overline{o},\overline{r})} \mathbf{f}(p(\overline{o}|\overline{r})) \geq \zeta \cdot \mathbf{f}(\frac{\sum_{\overline{o}} p(\overline{o}|\overline{r})}{\zeta}) \tag{5.11}$$

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:

$$\sigma_{o_k,r_k} = \frac{1}{\zeta \cdot \binom{m-i}{k-i} \cdot p(c_k)}, \tag{5.12}$$

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 5.3. Naively, one may expect

$O \setminus R$	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 5.3: Example. strongest collusion attack's strategy matrix

the attackers to always lie, to ensure the probability that a given rating is truthful is half. However, each attacker in Table 5.3 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 5.5. *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 5.4 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, $p(\overline{o}|\overline{r}) = 1/\zeta$. □

5.4 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 5.1). 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 5.2) 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 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, under-approximated attacks are 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 under-approximate 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. Muller and Schweitzer [85] provide a general formula that allows mathematically correcting trust evaluations, given the attackers' behavior. We propose to use computations in the formula, based on the assumption that the attackers' behavior 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 a 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 5.5, 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 5.1, 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.

5.5 Summary

We quantify 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/bad-mouthing are 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.

Chapter 6

Dynamic Unfair Rating Attacks

There are various formulations of the unfair rating attacks in the literature [18, 91, 92]. Typically, there are three types: ballot-stuffing – for some targets, attackers always provide positive ratings (e.g., a number of buyers in eBay are bribed to rate an unreliable seller highly), bad-mouthing – for some targets, attackers always provide negative ratings – and lying – for some targets, attackers always report the opposite of their true observations [87]. These attacks are all static, meaning attackers' behaviors are independent of the rating history. We have studied static attacks in the chapters before.

We refer to unfair rating attacks where time (or history) influences attackers' behavior as dynamic. A robust trust system must also function properly under dynamic unfair rating attacks – where attackers' strategies are closely related to their rating histories. In this chapter, we study dynamic unfair rating attacks [93]. The commonly studied form of dynamic attacks is camouflage attacks – where attackers pretend to be honest by rating strategically [2, 94]. In the literature, camouflage attacks are usually defined very specifically. For example, in both [2, 95], attackers are assumed to provide fair ratings to common sellers (not duopoly sellers) before a time threshold, and then provide unfair ratings to all sellers in a simulated e-marketplace system. To only be able to defend against camouflage attacks cannot guarantee the security of a system, however. In reality, it is hard to predict what kind of dynamic attacks would happen. Hence, it is important to study whether there are more harmful dynamic attacks.

In this chapter, we focus on modeling dynamic attacks and measuring their harm. We apply the information-theoretic measurement of harm (*information leakage*)which we introduced in 4. The harm of an attack depends on how much information an advisee can gain about the real observations of the attacker, which influences the difficulty. As we will discuss later, only static attacks are considered in Wang et al. [82], which have fundamentally different assumptions.

For dynamic unfair rating attacks, we introduce a stochastic process-based model. The hindsight of an advisee about past ratings influences the information it can gain from a dynamic attack. Hence, we distinguish three cases: (1) the advisee cannot determine whether the attacker reported the truth (*blind advisee*), (2) the advisee can accurately determine whether the attacker reported the truth (*aware advisee*), and (3) the advisee can determine whether the attacker reported the truth with limited accuracy (*general advisee*). For each type of advisees, we derive the harm of dynamic unfair rating attacks, and relate it to the harm of merely pretending to be honest.

We found that initially purely pretending to be honest is not the most harmful against any types of advisees. For blind advisees who have no hindsight, camouflage attacks are no more harmful than static attacks. Even for aware advisees, who can perceive honest behavior, it is not the worst to be cheated by disguised honesty. We found that, for all advisees, to cause the maximum harm, attackers must be honest with smaller probability than to lie, even initially. Furthermore, we found that a non-blind advisee can always gain some information, provided the percentage of attackers is below an exponentially large threshold. The most harmful dynamic attacks have not been identified in the literature.

6.1 Modeling Dynamic Attacks

Our usage of the term camouflage attacks refers to advisors. We do not consider camouflaging targets (e.g. sellers), which are considered in Oram [96] or "karma suicide attacks" in Glynos et al. [97]. E.g. in e-commerce, some raters first provide reliable suggestions regarding arbitrary sellers, to gain the trust of a buyer, then they cheat the

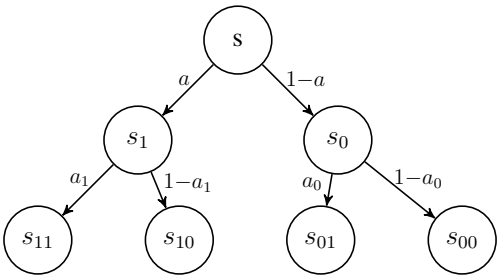


Figure 6.1: Attacker strategies for two ratings.

buyer by unfairly recommending colluding sellers. Below, we call an attack with k honest ratings followed by lies the k-camouflage attack. In dynamic unfair rating attacks, attackers' strategies have memory, hence we need states in attack modeling.

The random variables O_i and R_i represent the observation and rating of an advisor in the i^{th} iteration (or about i^{th} target). O_i only depends on the target's integrity, but R_i depends on both the honesty and strategies of the advisor. We introduce a random variable P to represent honesty of the advisor. The probability $p = p(P{=}1)$ represents the probability that the advisor is honest, and $1{-}p = p(P{=}0)$ that he is dishonest. Hence, $p(R_i = O_i|P = 1) = 1$, for all i. However, $p(R_i = O_i|P = 0)$ depends on the strategies of the attacker.

Let B_i be sets of binary strings of length i, B_i^1 (B_i^0) the subset of strings ending in 1 (0), and $\hat{b}_i$ the string consisting of i 1's. We denote concatenation of b and c as bc. Finally, $b(i)$ refers to the i^{th} bit of b.

Let $\mathcal{S} = \{s_b|b \in B\}$ be a set of states. We introduce a random process $\{S_i : i \in n\}$ to model a dynamic attack of size n. An outcome s_b of random variable S_i represents the state of the attacker after the i^{th} rating, which is modeled as its behavior history after the rating. Formally, $p(O_i{=}R_i|S_i{=}s_b) = b(i)$. The transition probability $p(S_{i+1} = s_c|S_i = s_b)$ is a_b if $c{=}b1$, $1{-}a_b$ if $c{=}b0$, and 0 otherwise. The random process $\{S_i : i \in n\}$ is a Markov chain.

As an example, the model of attacks of size 2 is shown in Figure 6.1. In the initial state, s, no rating occurred. In the state s_{01}, for instance, the attacker lied in the first rating and told the truth in the second rating. The probability of reaching s_{01} is $(1{-}a)a_0$, which we shorthand to α_{01}. Formally, letting $\pi(x, 1) = x$ and $\pi(x, 0) = 1 - x$, we set

$\alpha_b = \pi(a, b(1)) \cdot \pi(a_{b(1)}, b(2)) \cdot \ldots \cdot \pi(a_{b(1)b(2)\ldots b(i-1)}, b(i))$. Simple algebra shows that $p(s_b|P{=}0) = \alpha_b$.

As we focus on the behavior of malicious advisors but not targets, we assume maximum uncertainty (entropy) for targets' integrity (or $\overrightarrow{O_n}$), meaning O_i are independent, and $p(o_i){=}1/2$, $H(O_i){=}1$ for all i. Under this assumption, targets' integrity is invariant and will not influence the computation of information leakage over time. Furthermore, $p(r_i){=}\sum_{o_i} p(o_i)p(r_i|o_i){=}1/2$ for $r_i{\in}\{0,1\}$. As r_i and r_j ($i{\neq}j$) are independent without any observations, $p(\overrightarrow{r_i}){=}1/2^i$.

6.1.1 Define the Strength of Dynamic Attacks

In Chapter 4, we propose to measure the harm of static unfair rating attacks based on information theory. Specifically, the harm of attacks depends on the information leakage of advisors' ratings about their observations. An attack with more information leakage is less harmful to an advisee, since it makes it easier for decision making (more intuitions can be found in 4). We apply such a measurement of harm to dynamic attacks.

Strategies of dynamic attackers are closely related to their rating history, implying their ratings over time may be correlated, which may provide extra information to an advisee. Whereas in a static attack, ratings given in different iterations can be treated as independent of each other, for dynamic attacks, we must measure information leakage over the sequence of ratings, to capture this extra information. We use I_i to represent the information leakage of the rating in i^{th} iteration. The computation of I_i is influenced by ratings received in the past ($\overrightarrow{R}_{i-1}$) and the advisee's hindsight, as will be presented and explained in the following sections. The information leakage of an attack of size n is $\sum_{1\leqslant i\leqslant n} I_i$.

6.2 Measuring Dynamic Attacks

In this section, we study the harm of dynamic attacks based on the model and the measurement introduced above. Especially, we will figure out whether commonly studied

camouflage attacks are the most harmful, and if not, what would be the most harmful attacks.

In dynamic attacks, an advisee's ability to judge an attackers' past behavior influences its perception of the future behavior. From an information theoretical perspective, the information an advisee can gain from future ratings is influenced by its ability to judge past ratings. Hence, the harm of an attack is closely related to the judging ability of the advisee. For example, an attack may cause different harm (information leakage) for advisees that can accurately judge the past and those that cannot.

We study three types of advisees in total: blind advisees, aware advisees, and general advisees. They differ only in abilities in judging the truthfulness of past ratings. Blind advisees are those who cannot distinguish truth from lies. Aware advisees are those who can accurately distinguish truth from lies. General advisees are those with limited judging abilities. The first two types are extreme cases of general advisees.

We study each type of advisees in a subsection. In each subsection, we start with dynamic attacks of size 2 as a running example. We present the information leakage for the appropriate advisees. Based on the definition, we then prove what would be the most harmful attacks of size 2, and some theoretical properties. Then, we generalize all those results to dynamic attacks of arbitrary size.

6.2.1 Attacks against Blind Advisees

In this section, we study cases where an advisee is completely unable to judge whether the attacker has told the truth in the past. This happens when the advisee does not or cannot verify the ratings. We call these advisees *blind advisees*. Below we study the impact of dynamic attacks on blind advisees.

6.2.1.1 Two iterations

We start with the attacks modeled in the example from Figure 6.1. The information leakage of the first rating R_1 with regard to O_1 is $H(O_1)-H(O_1|R_1)$. After receiving R_1 but before receiving R_2, the advisee's uncertainty about O_2 is $H(O_2|R_1)$.

Upon receiving R_2, the uncertainty changes to $H(O_2|R_2, R_1)$. The information leakage of the second rating is the reduction of the advisee's uncertainty about O_2, i.e., $H(O_2|R_1)-H(O_2|R_2, R_1)$. The total information leakage is simply the sum of the information leakage of the different ratings:

$$H(O_1)-H(O_1|R_1) + H(O_2|R_1)-H(O_2|R_1, R_2) \tag{6.1}$$

As the observations of the attacker only depend on the integrity of the target (which is assumed of the maximum uncertainty), O_2 is independent of R_1, and $H(O_2|R_1)=H(O_2)=1$. Similar independence can be proved between any O_i and $\overrightarrow{R_{i-1}}$, $(\overrightarrow{O_{i-1}}, \overrightarrow{R_{i-1}})$, i.e., $H(O_i|\overrightarrow{R_{i-1}})=H(O_i|\overrightarrow{O_{i-1}}, \overrightarrow{R_{i-1}})=H(O_i)=1$.

Proposition 6.1. *For dynamic attacks of size* 2*, the information leakage is* 0 *iff* $p\leqslant\frac{1}{2}$, $a=\frac{1-2p}{2(1-p)}$ *and* $(1-2p)a_1+a_0=1-2p$.

Proof. Formula (6.1) equals 0 iff $H(O_1)=H(O_1|R_1)$ and $H(O_2)=H(O_1|R_1, R_2)$. These happen iff O_1 is independent of R_1, and O_2 is independent of (R_1, R_2), which means $p(o_1|r_1) = p(o_1)$ and $p(o_2|r_1, r_2)=p(o_2)$ must hold for all o_1, o_2, r_1, r_2. The probabilities can be rewritten via a, a_0, a_1, p. Basic algebra suffices to prove the proposition. □

Proposition 6.1 gives strategies that induce the most harmful dynamic attacks of size 2. For $p=1/2$, we get $a=a_0=0$. This implies when there are equal numbers of honest advisors and attackers, the way of absolutely hiding information is to always lie. Intuitively, this is because a randomly selected advisor in the system is equally likely to tell the truth and to lie. For $p>1/2$, where honest advisors outnumber attackers, there must be information leakage.

6.2.1.2 Formula for n iterations

We generalize formula (6.1) to dynamic attacks of size n. Similar to the example, $\sum_i H(O_i|\overrightarrow{R_{i-1}}) = n$, which we use to simplify the definition of information leakage:

$$n - \sum_i H(O_i|\overrightarrow{R_i}) \tag{6.2}$$

The conditional entropies can be represented in terms of p, n and the strategies of the attacker (i.e. the collection $a_{...}$, using the shorthand α):

Theorem 6.2. *Let B_i^0 (B_i^1) be the set of all bitstrings of length i ending with 0 (1). The information leakage of n iterations is:* $n+\sum_{1\le i\le n}\left(\mathbf{f}((1-p)\sum_{b\in B_i^0}\alpha_b)+\mathbf{f}(p+(1-p)\sum_{b\in B_i^1}\alpha_b)\right)$

Proof. Unfold the conditional entropy, apply the law of total probability over the possible states, and substitute the resulting terms with α. □

6.2.1.3 Theoretical results for n iterations

Blind advisees cannot learn an attacker's honesty from its past ratings. Hence, ratings in different iterations are independent to blind advisees. Information gain of a rating is not related to other ratings:

Theorem 6.3. *Blind advisees will not learn from past ratings:*

$$H(O_i)-H(O_i|R_i) = H(O_i)-H(O_i|\overrightarrow{R_i}) \tag{6.3}$$

Proof. The main equation is equivalent to $H(O_i|R_i) = H(O_i|\overrightarrow{R_i})$. Considering $p(o_i|r_i)=p(r_i|o_i)$, we get $p(o_i|r_i) = 2\cdot\sum_{\overrightarrow{o_{i-1}},\overrightarrow{r_{i-1}}} p(\overrightarrow{o_i},\overrightarrow{r_i})$. At the same time, $p(o_i|\overrightarrow{r_i}) = 2^i\cdot\sum_{\overrightarrow{o_{i-1}}} p(\overrightarrow{o_i},\overrightarrow{r_i})$, and $\sum_{\overrightarrow{o_{i-1}}} p(\overrightarrow{o_i},\overrightarrow{r_i})$ remains the same for any value of $\overrightarrow{r_{i-1}}$, hence we get $p(o_i|r_i)=p(o_i|\overrightarrow{r_i})$ hold for all $o_i, r_i, \overrightarrow{r_i}$. Furthermore, $p(r_i)=2^{i-1}p(\overrightarrow{r_i})$, thus the equality between two conditional entropy can be derived. □

Theorem 6.3 implies that for a blind advisee, a rating of a dynamic attacker can be treated as independent of the iteration in which it happens. And the total information leakage is the sum of the information leakage of individual ratings, i.e.,

$\sum_{0<i\leqslant n} H(O_i)-H(O_i|R_i)$. Dynamic attacks on blind advisees are equivalent to repeated static attacks. The maximum harm of static attacks has already been studied in Wang et al. [82]. Applying the results to dynamic attacks:

Corollary 6.4. *Zero information leakage (corresponding to ultimate attacks) occurs iff* $0 \leqslant p \leqslant 1/2$ *and* $a_b = \frac{1-2p}{2-2p}$*; whereas for* $1/2 \leqslant p \leqslant 1$*, information leakage is minimized when* $a_b = 0$*.*

Proof. Combining Theorem 6.3 with Theorems 3 and 4 in [82] suffices to prove the corollary. □

6.2.1.4 Camouflage attacks

Corollary 6.4 implies the most harmful attacks are not camouflage attacks, since $a_b \leq 1/2$ – attackers lie more often than tell the truth, even initially – in fact, there is no relationship between order and probability of lying. Intuitively, accumulating trustworthiness is meaningless against blind advisees, that cannot distinguish truth from lies. Furthermore, Theorem 6.3 implies that camouflage attacks are no more harmful than the repeated strongest static attacks.

Quantitatively, the information leakage of a k-camouflage attack is $n+(n-k)\cdot\mathbf{f}((1-p))$. It is minimized when $k=0$, i.e. always lie. This means the strategy to minimize information is not to do camouflage.

If camouflage attacks are not the strongest, how much weaker are they? Below, we study the harm of pure camouflage attacks. We use variable k ($0\leq k\leqslant n$, $k=0$ is not included because otherwise, camouflage is meaningless) to represent the turning point in the behavior of an attacker: after k^{th} iteration of rating, the attacker starts to always lie, and before that, he always tells the truth. The information leakage of pure camouflage attacks can be computed from the formula in Theorem 6.2:

Proposition 6.5. *The information leakage of a pure camouflage attack to blind advisees is:*

$$n + (n-k) \cdot \mathbf{f}((1-p)) \tag{6.4}$$

The maximum information leakage is n bits, which means all observations (O_i) completely depend on all ratings (R_i) in n iterations. This happens when either $p \in 0, 1$ or $k = n$. If $k = n$, the attackers simply keep telling the truth as honest advisors in all iterations, hence the advisee can get n bits information (the same goes for $p = 1$ where no attacker exist). If $p = 0$, all advisors are attackers and they give same honest (lying) ratings, in both of which cases the advisee gets maximal information.

The minimal information leakage is $0.47n+0.53k$, which corresponds to the maximal harm that pure camouflage attacks can cause, and it happens when $p \approx 0.63$. The first thing to note is that $0.47n+0.53k \gg 0$, which implies there is no little information under the strongest pure camouflage attacks. The condition of leading to maximal harm is only decided by p (which represents approximately the percentage of attackers), but not related with k (which decides attackers' strategies). This is because to blind advisees, dynamic attacks equal repeated static attacks, therefore when to start lying does not matter anymore. The quantity of maximal harm depends on k, however. With the decrease of k, meaning attackers give less honest ratings to gain trust, they cause more harm.

What if advisees are not blind, but smarter? Perhaps the camouflage attacks are more harmful to smarter advisees, which may recognize honest behavior, allowing attackers to accumulate trust. Next, we study advisees who can accurately tell whether advisors told the truth or not (i.e., *aware advisees*).

6.2.2 Attacks against Aware Advisees

Aware advisees can tell with perfect accuracy whether a previous rating was honest. For now, we ignore the issue of subjectivity (assuming there is no disagreement in honest ratings). Aware advisees are the polar opposite of blind advisees. In this section, we study what would be the most harmful dynamic attacks for aware advisees.

6.2.2.1 Two iterations

Again, we start with the example in Figure 6.1. The information leakage of the first rating remains $H(O_1)-H(O_1|R_1)$. Aware advisees know whether R_1 coincides with O_1, therefore before the second rating, the uncertainty about O_2 is $H(O_2|O_1, R_1)$. Upon receiving R_2, the uncertainty becomes $H(O_2|O_1, R_1, R_2)$. Hence, the information leakage for two ratings is:

$$H(O_1)-H(O_1|R_1)+H(O_2|R_1, O_1)-H(O_2|O_1, R_1, R_2) \tag{6.5}$$

Proposition 6.6. *For dynamic attacks of size* 2*, the information leakage is* 0 *iff* $p\leqslant\frac{1}{4}$, $a=\frac{1-2p}{2(1-p)}$, $a_1=\frac{1-4p}{2(1-2p)}$, $a_0=\frac{1}{2}$.

Proof. Formula (6.5) equals 0 iff $H(O_1)=H(O_1|R_1)$ and $H(O_2)=H(O_2|O_1, R_1, R_2)$. The condition for the first equality is already proved in Proposition 6.1. The second equality holds iff O_2 is independent of (O_1, R_1, R_2), which means $p(o_2)=p(o_2|o_1, r_1, r_2)$. Rewrite the equation using $a_{...}, p$, we can derive their values. □

Proposition 6.6 presents the ultimate attacks that cause 0 information leakage for aware advisees. Note that $a_1<a<\frac{1}{2}$, which implies the attacker has higher chance to lie than to tell the truth in the first rating, and the chance of lying increases if the attacker just told the truth. To get more general results, we extend the attack size from 2 to n ($n\geqslant 2$).

6.2.2.2 Formula for n iterations

The information leakage for n iterations is

$$n - \sum_{1\leqslant i\leqslant n} H(O_i|\overrightarrow{R_i}, \overrightarrow{O_{i-1}}) \tag{6.6}$$

Based on the chain rule of conditional entropy, and the conditional independence between O_i and R_j, $(j>i)$ given $\overrightarrow{O_{i-1}}, \overrightarrow{R_i}$, we can prove the information leakage formula

above is equal to

$$H(\overrightarrow{O_n}) - H(\overrightarrow{O_n}|\overrightarrow{R_n}) \tag{6.7}$$

As before, the conditional entropy can be represented in terms of p, n and the strategy parameters $a_{...}$:

Theorem 6.7. *The information leakage of n iterations is:*

$$n + \mathbf{f}(p + (1-p)\alpha_{\hat{b}_n}) + \sum_{b \in B_n, b \neq \hat{b}_n} \mathbf{f}((1-p)\alpha_b) \tag{6.8}$$

Proof. Modify the proof of Theorem 6.2 by adding a case distinction for the left-most branch. □

6.2.2.3 Theoretical results for n iterations

The attacker can render ratings useless, only if $p \leqslant 1/2^n$.

Theorem 6.8. *For aware advisees and dynamic attacks of size n, attacks with 0 information leakage occur when $0 \leqslant p \leqslant 1/2^n$, $a_{\hat{b}_i} = \frac{1}{2} - \frac{p}{2(1-p)\prod_{j=0}^{j=i-1} a_{\hat{b}_i}}$, and $a_b = 1/2$ for $b \neq \hat{b}_i$.*

Proof. Considering $H(\overrightarrow{O_n}) = n$, to minimize formula (6.7) we only need to maximize the subtracter. $H(\overrightarrow{O_n}|\overrightarrow{R_n}) = \sum_{\overrightarrow{r_n}} p(\overrightarrow{r_n}) \cdot H(\overrightarrow{O_n}|\overrightarrow{r_n})$, and $p(\overrightarrow{r_n})$ constantly equal to $\frac{1}{2^n}$. Based on the Jensen's inequality [98], the maximum of the conditional entropy happens when $p(\overrightarrow{O_n}|\overrightarrow{r_n})$ are equal for any value of $\overrightarrow{O_n}$ and $\overrightarrow{r_n}$, and the maximum value is n. The probabilities can be rewritten using all transition probabilities, and basic algebra suffices to prove the theorem. □

When $p > 1/2^n$, the strategy that causes the most harm is independent of p:

Theorem 6.9. *For aware advisees, the most harmful attack of size n, given $p > 1/2^n$, occurs when $a_{\hat{b}_i} = \frac{2^{n-i-1}-1}{2^{n-i}-1}$ and $a_b = 1/2$ for $b \neq \hat{b}_i$.*

Proof. Take the formula from Theorem 6.7. Since $p{>}1/2^n$, the minimum has $\alpha_{\hat{b}_n} = 0$. Using Jensen's inequality, it suffices to set the remaining α_{b_n} equal. Then, α_{b_n} must equal $\frac{1}{2^n-1}$, which happens when all a_b are set as in the theorem. □

Note that for blind advisees, 0 information leakage can be achieved for any $p{\leqslant}\frac{1}{2}$. But for aware advisees, the required value ranges of p are much smaller and get narrower as rating iterations increase ($p \leqslant 1/2^n$ for n iterations). This is in line with our intuition, that when advisees get smarter, it should be more difficult for attackers to hide information.

6.2.2.4 Camouflage attacks

Theorem 6.8 presents the most harmful attacks for aware advisees. Note that $a_{\hat{b}_i}{\geqslant}a_{\hat{b}_j}$ for $i{<}j$, meaning the probability of continuing to tell the truth is non-increasing over time. This is also the case in camouflage attacks, where attackers may gradually reduce the chance to tell the truth before it starts to lie. However, all $a_{\hat{b}_i}, i = 0, \ldots, (n{-}1)$ are below $1/2$, meaning lying is always more probable. Hence, although camouflage attacks are not the most harmful attacks for aware advisees, pretending to be honest with some decreasing probability is more harmful than a fixed probability.

Quantitatively, information leakage for k-camouflage is $n + \mathbf{f}(1-p) + \mathbf{f}(p)$, for $k \neq n$, and n, for $k = n$. Comparing to the blind advisees, camouflage attacks are less harmful. Moreover, provided the attacker lies at some point, it does not matter when he switches. Therefore, always lying is equally harmful as camouflage attacks.

6.2.3 Attacks against General Advisees

Blind advisees and aware advisees are two extreme examples of advisees. In this section, we study the impact of dynamic attacks on advisees in between of the extremes. We introduce random variables Q_i to represent an advisee's hindsight perception of attackers' honesty, with $Q_i{=}1$ ($Q_i{=}0$) denoting the attacker probably told the truth (lied) in the i^{th} rating. The accuracy of the advisee's hindsight depends on how much he

can learn from his own interactions with the target, or from other sources in the system. Subjective ratings are an example with low accuracy, since even when the advisee forms an opinion different from the rating, it remains probable that the advisor was not lying. For instance, if the user thinks the target is good after interacting with it, while the advisor reported the opposite, it is probable that the advisor was lying, but possible that the user and the advisor simply have different opinions about the target.

We use $q, (0 \leqslant q \leqslant 1)$ to describe the accuracy of the advisee's hindsight. With probability q, the advisee's perception is correct: $p(Q_i{=}1|O_i{=}R_i) = p(Q_i{=}0|O_i{\neq}R_i) = q$, and incorrect with probability $1{-}q$: $p(Q_i{=}0|O_i{=}R_i) = p(Q_i{=}1|O_i{\neq}R_i) = (1{-}q)$. For a high degree of subjectivity, we expect q to be close to $1/2$.

6.2.3.1 Two iterations

Q_i expresses the new knowledge of the advisee about the attacker after the i^{th} rating, the amount of which is decided by q. To show how this influences the definition of information leakage, consider again the example in Figure 6.1. The information leakage of the first iteration remains unchanged. After the first iteration, the advisee knows R_1 and Q_1, and the uncertainty about O_2 is $H(O_2|R_1, Q_1)$. Upon receiving R_2, the uncertainty about O_2 changes to $H(O_2|Q_1, R_1, R_2)$. Hence, the information leakage of the second iteration is $H(O_2|R_1, Q_1){-}H(O_2|Q_1, R_1, R_2)$. The total information leakage of the attack is the sum of the first and the second iterations:

$$H(O_1){-}H(O_1|R_1){+}H(O_2|R_1, Q_1){-}H(O_2|Q_1, R_1, R_2) \tag{6.9}$$

Recall that O_2 only depends on the target, $H(O_2|R_1, Q_1){=}H(O_2){=}1$. We first study the maximum impact of attacks of size 2:

Proposition 6.10. *For dynamic attacks of size 2, the information leakage is* 0 *for general advisees iff: (1) for* $q{\neq}\frac{1}{2}$, $p{\leqslant}\frac{1}{4}$, $a{=}\frac{1-2p}{2(1-p)}$, $a_1{=}\frac{1-4p}{2(1-2p)}$, $a_0{=}\frac{1}{2}$. *(2) for* $q{=}\frac{1}{2}$, *refer to strategies in Proposition 6.1.*

Proof. Formula (6.9) is 0 iff $H(O_1){=}H(O_1|R_1)$ and $H(O_2) {=}H(O_2|Q_1, R_1, R_2)$. We have proved in Proposition 6.1 that to get the first equality, $a{=}\frac{1-2p}{2(1-p)}$.

The second equality holds iff O_2 is independent of (Q_1, R_1, R_2), which means $p(o_2|Q_1, r_1, r_2){=}p(o_2){=}1/2$. Rewrite the probabilities using transition probabilities, to obtain a_0, a_1. □

Note that for any $q{\neq}\frac{1}{2}$ – i.e., an advisee has at least some accuracy in hindsight – the strategy to completely hide information remains the same. Thus, as long as an advisee is not blind, a single strategy suffices to completely hide information, regardless of the advisee's accuracy.

6.2.3.2 Formula for n iterations

Generalizing the attack to size n, the information leakage of an entire attack is:

$$n - \textstyle\sum_i H(O_i|\overrightarrow{R_i}, \overrightarrow{Q_{i-1}}) \tag{6.10}$$

Rewriting in terms of p, q, n and $a_{...}$:

Theorem 6.11. *Let* $\beta_{b,c} = \prod_{1\leqslant j<i} \pi(1{-}q, b(j) \oplus c(j))$. *For* $x \in \{0,1\}$, *let* $\gamma_c^x = (1{-}p)\sum_{b\in B_i^x} \alpha_b \beta_{b,c}$. *Let* $\delta_{c,i} = \gamma_c^0 + \gamma_c^1 + p\beta_{\hat{b}_i,c}$. *The information leakage is:*

$$n + \sum_{1\leqslant i\leqslant n} \sum_{c\in B_i} \delta_{c,i} \cdot \left[\mathbf{f}\left(\frac{\gamma_c^1 + p\beta_{\hat{b}_i,c}}{\delta_{c,i}}\right) + \mathbf{f}\left(\frac{\gamma_c^0}{\delta_{c,i}}\right)\right] \tag{6.11}$$

Proof. In addition to the technique used in Theorem 6.2, apply the law of total probability over all Q_i. □

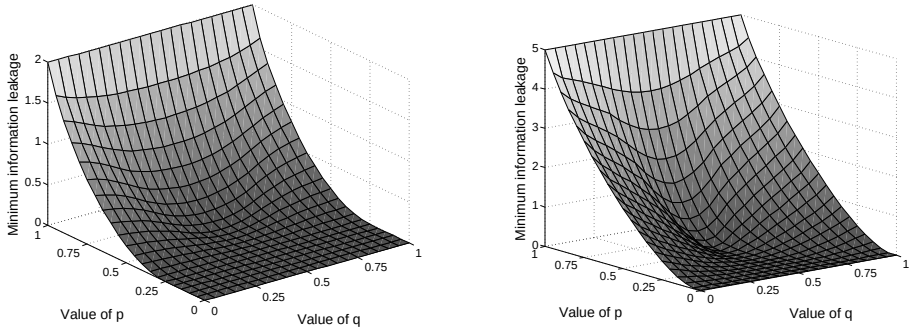
6.2.3.3 Theoretical results for n iterations

The blind and aware advisees are special cases of the general advisees:

Theorem 6.12. *Dynamic attacks on blind advisees (aware advisees) are a special case of attacks on general advisees, where* $q = 1/2$ *(*$q{=}1$*). Specifically:*

$$H(O_i|\overrightarrow{R_i}) = H(O_i|\overrightarrow{R_i}, \overrightarrow{Q_{i-1}}), \qquad q{=}1/2 \tag{6.12}$$

$$H(O_i|\overrightarrow{R_i}, \overrightarrow{O_{i-1}}) = H(O_i|\overrightarrow{R_i}, \overrightarrow{Q_{i-1}}), \qquad q{=}1 \tag{6.13}$$



(a) (b)

Figure 6.2: The information leakage in the most harmful strategies for two (a) or five (b) iterations.

Proof. When $q = 1/2$, $\pi(1-q, c(j) \oplus b(j))$ is a constant $\frac{1}{2}$ for all j, making $\beta_{b,c}$ is a constant, and simplifying γ to fit Theorem 6.2. When $q = 1$, $\beta_{b,c}$ contains a 0-factor, whenever $b \neq c$, meaning $\beta_{b,c} = 1$ iff $b = c$. This implies $\gamma_c^x = \alpha_c$, and the term $p\beta_{\hat{b}_i,c}$ equals p iff $c = \hat{b}_i$. Some formula manipulation shows it fits Theorem 6.7. □

Theorem 6.13. *For dynamic attacks of size n on general advisees, 0 information leakage can be achieved when for all $1 \leqslant i \leqslant n$, $c \in B_i$: $\gamma_c^1 + p\beta_{\hat{b}_i,c} = \gamma_c^0$.*

Proof. To get information leakage be 0, refer to formula (6.10), $H(O_i) = H(O_i|\overrightarrow{R_i}, \overrightarrow{Q_{i-1}})$ must hold for all i. The equalities are achieved iff O_i is independent of $\overrightarrow{R_i}, \overrightarrow{Q_{i-1}}$, which means $p(o_i|\overrightarrow{r_i}, \overrightarrow{q_{i-1}}) = p(o_i) = 1/2$. This implies $p(o_i{=}r_i, \overrightarrow{r_i}, \overrightarrow{q_{i-1}}) = p(o_i{\neq}r_i, \overrightarrow{r_i}, \overrightarrow{q_{i-1}})$, which instantiate $\gamma_c^1 + p\beta_{\hat{b}_i,c}$ and γ_c^0. □

6.2.3.4 Numerical results for n iterations

Using numerical approximation, we found strategies that cause close to maximal harm, given specific p, q and n. Figure 6.2 plots the information leakage w.r.t. p and q, for

n=2 in Figure 6.2 (a), and n=5 in Figure 6.2 (b). Similarly, Figure 6.3 plots the initial probability of telling the truth (a) w.r.t. p and q, for n=2 and n=5.

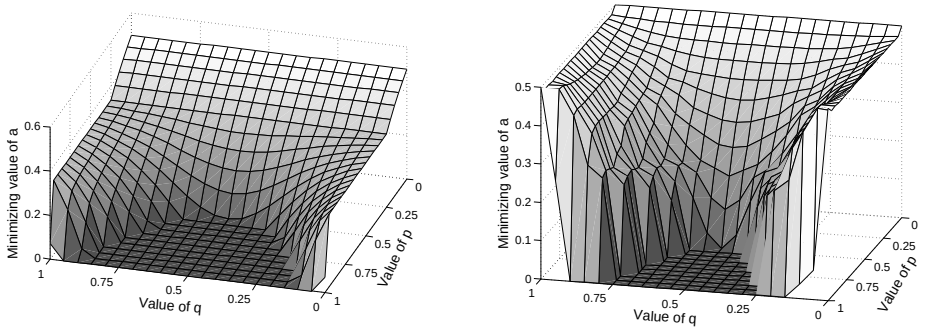
Observe that the graphs are symmetrical over $q = 1/2$, since when $q < 1/2$ we can simply swap the meaning of $Q = 0$ and $Q = 1$. Moreover, note that a is bound by $1/2$ and information leakage by n.

In Figure 6.2, note that if $p = 1$, all information is leaked – since all advisors are honest. Secondly, for two iterations, the information leakage is 0 when $0 \leqslant p \leqslant 1/4$, and for five, when $0 \leqslant p \leqslant 1/32$. This pattern holds for all q, except $q = 1/2$, in which case information leakage is 0 when $0 \leqslant p \leqslant 1/2$. We find that (barring $q = 1/2$) obtaining 0 information leakage requires exponentially more attackers relative to honest advisees, as the number of iterations increases. This extends the theoretical result in Theorem 6.8, where we prove this pattern for aware advisees ($q = 1$).

Regarding Figure 6.3, first observe that there appears to be a phase transition at $p = 1/2$ and $p = 1/32$, in Figures 6.3 (a) and Figures 6.3 (b), respectively. The choice of optimal a is independent of q, for smaller p. Since the same strategy is optimal for blind advisees and aware advisees, when there are many attackers, and there is no point pretending to be honest to blind advisees, there is no point pretending to be honest to any advisees (for small p).

Second, observe that for larger p and q close to $1/2$, $a = 0$. Thus, the attacker initially always lies (Corollary 6.4 proves this for $q = 1/2$). As n increases, the area where the attacker always lies appears to shrink. The strategy of always lying is less often the most harmful, when the number of ratings increases.

Finally, for larger p and q closer to 1 (or 0), $a > 0$. Thus, the attacker sometimes tells the truth, despite the fact that for static scenarios it should always lie. This indicates that pretending to be honest initially (like probabilistic camouflaging) is (the most) harmful. The probability of telling the truth initially should, however, not be too high – it appears to be bounded by $1/2$.



(a) (b)

Figure 6.3: The a value in the most harmful strategies for two (a) or five (b) iterations.

6.2.3.5 Camouflage attacks

Generally, the analysis of camouflage attacks is the same as for aware advisees, with the major exception of an area around $q = 1/2$. The area in the graphs in Figure 6.3 where $a = 0$ denotes combinations of p, q and n where pretending to be honest definitely causes no harm. We see that the area increases in size as p grows, but decreases in size as n grows. Furthermore, as n and p grow, the cut-off (between the cases where pretending to be honest causes harm and cases where it does not) becomes sharper.

Quantitatively, letting $\dot{b}_i$ be k 1's followed by $i-k$ 0's, the information leakage of k-camouflage is $n + \sum_{k<i\leq n} \sum_{c\in B_i} \Big(\mathbf{f}(p\beta_{\hat{b}_i,c}) + \mathbf{f}((1-p)\beta_{\dot{b}_i,c}) - \mathbf{f}(p\beta_{\hat{b}_i,c} + (1-p)\beta_{\dot{b}_i,c})\Big)$. Again, this is minimized when $k = 0$, meaning that even always lying is worse than camouflage.

6.3 Summary

In camouflage attacks, advisors initially pretend to be honest to accumulate their trustworthiness, then they lie to advisees. The motivating question for this chapter, is whether such behavior is harmful. It turns out that the answer to that question depends on the hindsight of the advisees. If advisees have no hindsight (blind advisees), then attackers pretending to be honest can only be beneficial to them. If advisees have perfect or limited hindsight (aware/general advisees), then it may cause harm when attackers initially pretend to be honest with a (small) probability. It never causes harm to (initially) pretend to be honest with probability over $1/2$, let alone probability 1. Therefore, the camouflage attack, where attackers are always honest until a certain point, is not very harmful.

The results are obtained by using a method to measure the strength of unfair rating attacks applied in the literature [82, 87]. The method is not suitable for multiple ratings, so we had to generalize the methodology. Moreover, we needed to construct a formal model of all dynamic rating behavior. We let the advisor be a stochastic process that generates the ratings according to some strategy. We are able to derive explicit formulas expressing the harm of any dynamic attack against any of the three advisees.

The theory not only allowed us to answer the main question, but also to prove interesting properties about the harm of dynamic attacks. For example, attackers can render ratings completely useless to blind advisees, whenever they are in the majority. But for all other advisees, the attackers need to greatly outnumber the honest advisees (exponential in the number of ratings) to render ratings useless. Another interesting result is that against aware advisees, the most harmful attacks do not depend on how many attackers there are (unless they greatly outnumber the honest advisees).

By this chapter, the approach of applying information theory to quantify the strength of attacks becomes general and adaptable.

Chapter 7

The Impact of Subjectivity on Robustness

In real rating, there is typically a degree of subjectivity. Different honest advisors[11] may generate divergent opinions for a same observation, thus providing different ratings. Also, an advisee may hold different opinion with an honest advisor, whose rating may seem to be misleading. For example, if a buyer has much higher expectation for a product than another, then he may give lower rating.

Both subjectivity difference and dishonesty reduce the quality of ratings. In literature, typically, subjectivity and dishonesty are studied orthogonally. Interestingly, subjectivity may change unfair rating attacks, e.g., introducing an attack where dishonest advisors camouflage as honest-but-subjective, thus changing robustness of trust systems. Therefore, the interplay of dishonesty and subjectivity needs to be studied. We are the first to formally analyzes the interference of subjectivity and robustness.

We first study the impact of subjectivity on robustness. Three types of advisors' subjectivity are considered in rating a target: different emphasis on features of the target, different expectations and preferences on a feature. We propose a probabilistic model of rating, where an advisor could be honest (and being subjective or objective), or dishonest. The amount of information leakage of a rating under an attack [82] measures the damage of the attack – less information means being worse off, and conversely, the

[11] Malicious or dishonest advisors strategically rates, where we think there is no subjectivity.

robustness of a trust system – more information means being more robust. We compare the difficulty of achieving ultimate attacks (i.e. attacks that prevent any information leakage) in rating with and without subjectivity, and find the existence of subjectivity makes ultimate attacks easier. We introduce an ordering of subjectivity, with which we show that higher degree of subjectivity in rating means less robustness.

Methods to mitigate the misleading effect of subjective ratings have been proposed (Section 2). Since subjectivity decreases robustness of trust systems, we study whether these methods improve the robustness. The first method is for advisors to rate each feature of a target (feature-based rating) instead of only providing a global rating, which is widely adopted in real online trust systems like Taobao, Amazon. We find that compared with global rating, feature-based rating may be less robust – relaxing restrictions on achieving ultimate attacks and may be even more subjective.

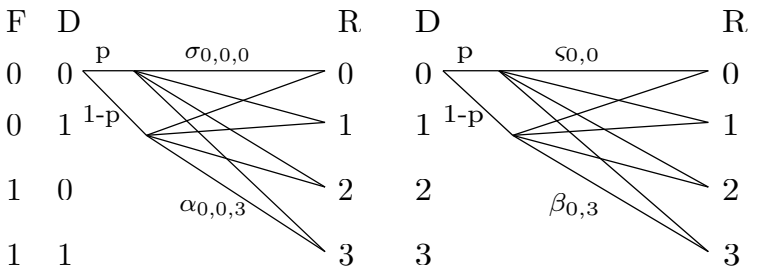
Clustering advisors based on their rating behavior is another way to distinguish subjectivity difference. We alter the rating model to include clustering, and find that clustering, either based on advisors' honesty or subjectivity, increases expected information leakage, regardless of attackers' strategies or the accuracy of clustering. We prove that finer clustering (i.e., splitting existing clusters, and in the limit, considering advisors' behaviors individually) increases robustness.

There exist different ways to deal with clusters, regarding whether to exclude seemingly dishonest clusters or exploiting all clusters. We prove excluding clusters loses information, unless the excluded cluster is under ultimate attacks. Hence, to accurately judge the property of clusters is important before handling clusters.

In this chapter, Section 7.1 introduces the modeling and ordering of subjective rating. Section 7.2 analyzes the robustness of subjective rating. Section 7.3 studies the robustness of two types of methods to mitigate the effect of subjectivity.

7.1 Modeling Subjective Rating under Attacks

The advisee's decision based on ratings is kept abstract. We do not specify the purpose of a trust system; it can range from evaluating products (e.g., e-commerce) to improving



(a) Example of global subjective rating under attacks

(b) Example of feature-based subjective rating under attacks

Figure 7.1: Examples of global subjective rating (a) and feature-based subjective rating (b) under attacks

security (trust-based security systems). We model subjectivity of honest advisors in rating, meanwhile, we consider the possibility that an arbitrary advisor being dishonest (an attacker), and we define an ordering of subjective rating.

7.1.1 Subjectivity in Rating

An undetermined advisor is either honest (non-strategic) or dishonest (strategic). An honest advisor may be subjective in rating, making his ratings deviate from what an advisee chooses or from other honest advisors' ratings. Subjectivity means "based on or influenced by personal feelings, tastes, or opinions" (Oxford Dictionary). We consider three causes of subjectivity: advisors' differences in emphasis, expectation, and disposition. Different advisors may emphasize different features, either when they grade a target, or when they suggest an option. For example, one honest advisor may rate a site unsafe due to excessive advertisements, whereas another honest advisor rates it safe, since it does not operate malware. Even when emphasizing on the same features, ratings may still be different if advisors have different expectations or dispositions. When two honest advisors both take "amount of advertisements" as criterion for safety, one may find it excessive and rate it unsafe, whereas the other may find it acceptable and rate it safe. Take another example, given a feature "location" in rating a hotel, some honest advisors may prefer downtown while some others may prefer countryside.

7.1.2 Modeling Subjective Rating

An advisee cares about a finite set of features, denoted $D_1 \times \cdots \times D_{n_d}$; but not about $F_1 \times \cdots \times F_{n_f}$. Let $D = D_1 \times \cdots \times D_{n_d}$ and $F = F_1 \times \cdots \times F_{n_f}$. Outcomes of D or F could be, e.g., the scores of relevant features, or options for decision making considering relevant features. The advisee wants to estimate D, but advisors may care about F[12], so F may influence ratings. Since outcomes of D (F) are finite, we can label them $0, \ldots, n_d{-}1$ ($0, \ldots, n_f{-}1$). We use R to signify the rating given by an advisor, with its outcomes being a finite set of choices labelled as $0, \ldots, n_r{-}1$. Without ratings, we assume maximum uncertainty about D and F, meaning their prior distributions are uniform.

An advisor's behavior is either honest (denoted H, with $p = p(H)$), or strategic (denoted $\neg H$, with $1 - p = p(\neg H)$). Honest behavior can be characterised by a 3D *subjectivity matrix* σ, where $\sigma_{f,d,r}$ expresses the probability that an honest advisor provides rating r, given the outcome of F, D is f, d; or $\sigma_{f,d,r}{=}p(r|f, d, H)$. Our goal is to analyze subjectivity, so we treat the honest behavior σ as given.

For strategic behavior, we assume the worst possible behavior for the advisee (since the robustness of a system should be measured under the worst-case rating behavior [82]). Whatever that strategic behavior is, we can capture it in another 3D matrix α, with $\alpha_{f,d,r}{=}p(r|f, d, \neg H)$. Now, an arbitrary advisor's rating behavior can be represented by a matrix $\mu{=}p{\cdot}\sigma{+}(1 - p){\cdot}\alpha$, which describes how the advisor's properties p, σ, α decide the link between its rating R and D, F. We formally define subjective rating as follows:

Definition 7.1. (n_f, n_d, σ)-*subjective rating* is a rating function with $f_{n_f,n_d,\sigma}(f, d, \alpha)(r){=}p\sigma_{f,d,r}{+}(1{-}p)\alpha_{f,d,r}$, $d{\in}\{0, \ldots, n_d{-}1\}$, $f{\in}\{0, \ldots, n_f{-}1\}$, $r{\in}\{0, \ldots, n_r{-}1\}$.

[12]Note that from Chapter 4 to Chapter 6, we assumed that there is no subjectivity in ratings. Hence, we did not distinguish features that an advisee cares from that of an advisor. And we use a single variable O to model the observation.

Note that the rating model in Chapter 4 is a special case of $f_{n_f,n_d,\sigma}$: σ is an identity matrix I and there is no F. We refer to it as $f_{n_d,I}$, which is an objective rating function. As a rating R is based on all features D, F, we name $f_{n_f,n_d,\sigma}$ as a global rating function.

A modeling example is presented in Figure 7.1a. D, F have binary outcomes $0, 1$, resulting in four combinations corresponding to the outcomes of R. $\sigma_{0,0,0}$ means an honest advisor observes $0, 0$ for D, F and reports 0. $\alpha_{0,0,3}$ means under the same observation, a dishonest advisor reports 3.

7.1.3 Ordering of Subjective Rating

We create an ordering of subjectivity, to be able to say one advisor is at most as subjective as another ($\trianglelefteq$); or that one advisor is less subjective than another ($\triangleleft$). The ordering $\trianglelefteq$ is not complete (it is a preorder), 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.

The subjectivity in ratings may result from advisors' different emphasis on features. However, an advisee is not interested in features F, only on how F affects interesting features D[13].

To construct the ordering, we first average over the F dimension of matrix σ and label its two-dimensional version as ς: with $\varsigma_{d,r}{=}p(r|d,H){=}\frac{1}{n_f}\sum_f \sigma_{d,f,r}$. Similarly, let $\beta_{d,r}{=}\frac{1}{n_f}\sum_f \alpha_{d,f,r}$ and $\nu_{d,r}{=}\frac{1}{n_f}\sum_f \mu_{d,f,r}$. We define the ordering on these 2D subjectivity matrices, and simply let $\varsigma{\trianglelefteq}\varsigma'$ imply $f_{n_f,n_d,\sigma}{\trianglelefteq}f_{n_f,n_d,\sigma'}$.

There are some notions that any reasonable subjectivity ordering of matrices must have: 1) The relation must be reflexive and transitive (i.e. subjectivity is a preorder), and no antisymmetry, since two different matrices may be equally subjective. 2) The identity matrix I is a minimum element: $\nexists_\varsigma \varsigma \triangleleft I$; since reporting with I is objective rating. 3) The uniform matrix U is a maximal element: $\nexists_\varsigma U \triangleleft \varsigma$; since ratings from U are unrelated to the truth.

[13]For example, take a general product rating that depends on product strength (more objective) and aesthetics (more subjective), and we are interested in strength only. The fact that aesthetics affect the general score matters, since they will not perfectly accurately report the strength. We are interested, therefore in the subjectivity in the relation between strength and general rating.

If a specific rating implies a high probability for a small set of values (or a single value) of D, then the subjectivity is low. If, on the other hand, a rating results in roughly equal probabilities for all values of D, then the subjectivity is high. To capture this, we need to be able to reason how mixed $p(d|r), 0{\leq}d{<}n_d$ are given a rating r – being more mixed implies being more subjective under the rating.

Let x, y be vectors of n elements. They have equal sums: $|x|{=}\sum_{0\leq i<n} x_i{=}1{=}|y|$. We expect $0.2, 0.7, 0.1$ to be more mixed than $0, 1, 0$. So for x to be less mixed, its largest value must exceed y's largest value. But also, $0.8, 0.1, 0.1$ seems more mixed than $0.8, 0.2, 0$, so we must take the second value into account (and indeed the i^{th} value, $0{\leq}i{<}n$). If, for any i, the i most probable values are more probable, then it must be less mixed. This is captured by the preorder majorisation [99]:

Definition 7.2. Let $x^{\downarrow}$ and $y^{\downarrow}$ be n-sized vectors x and y sorted non-increasingly and have same sums. Vector x is majorised by y: $x{\prec}y$, if $\forall 1{\leq}j{\leq}n$, $\sum_{0\leq i<j} x_i^{\downarrow} \leq \sum_{0\leq i<j} y_i^{\downarrow}$.

Antisymmetry does not hold here: if components of x, y are equal but not in the same order, then $x{\prec}y, y{\prec}x$ and $x{\neq}y$.

We write $\overrightarrow{\varsigma_j}$ to mean the vector $(\varsigma_{0,j}, \ldots, \varsigma_{n_d-1,j})$. As $p(D{=}i|R{=}j){=}\varsigma_{i,j}/|\overrightarrow{\varsigma_j}|$, normalized vector $\overrightarrow{\varsigma_j}/|\overrightarrow{\varsigma_j}|$ represents the distribution of $p(D{=}i|R{=}j), 0{\leq}i{<}n_d$. Relation $\overrightarrow{\varsigma_j}/|\overrightarrow{\varsigma_j}|{\prec}\overrightarrow{\varsigma'_j}/|\overrightarrow{\varsigma'_j}|$ implies that when reporting the same rating $R{=}j$, advisor with ς is more subjective than advisor with ς'.

The naive generalization from specific ratings (a column) to an expected rating (a matrix) is to do direct comparison. However, arguably, relabeling ratings (i.e. switching columns) should not affect the subjectivity (switching feature values in a column does not impact due to majorisation). E.g., if society collectively decides that 1 star is perfect, and 5 stars is horrible, the star ratings would be mirrored. Yet, this does not affect the subjectivity of the said ratings. We should allow subjectivity comparison under different ratings.

Note that the prior probability of a rating also matters. Namely, for each individual rating, advisor Amy may be less subjective than Bob, but Bob's least subjective rating

may be most likely to occur. The simple way to avoid this issue is to demand two columns under comparison to have the same weights: require both $\frac{\overrightarrow{\varsigma_j}}{|\overrightarrow{\varsigma_j}|} \prec \frac{\overrightarrow{\varsigma_l}}{|\overrightarrow{\varsigma_l}|}$ and $|\overrightarrow{\varsigma_j}| = |\overrightarrow{\varsigma_l}|$. Note that if $|\overrightarrow{\varsigma_j}| = |\overrightarrow{\varsigma_l}|$, then $\frac{\overrightarrow{\varsigma_j}}{|\overrightarrow{\varsigma_j}|} \prec \frac{\overrightarrow{\varsigma_l}}{|\overrightarrow{\varsigma_l}|}$ iff $\overrightarrow{\varsigma_j} \prec \overrightarrow{\varsigma_l}$.

We formally define subjectivity ordering as:

Definition 7.3. ς is at most as subjective as ς' ($\varsigma \trianglelefteq \varsigma'$) when, there exists a one-to-one mapping $m : [0, n_r) \to [0, n_r)$, such that $|\overrightarrow{\varsigma_j}| = |\overrightarrow{\varsigma'_{m(j)}}|$ and $\overrightarrow{\varsigma_j} \prec \overrightarrow{\varsigma'_{m(j)}}$.

We say that ς is less subjective than ς', $\varsigma \triangleleft \varsigma'$, if $\varsigma \trianglelefteq \varsigma'$ but not $\varsigma' \trianglelefteq \varsigma$; i.e. when they are not equally subjective.

The hard requirements can be straightforwardly proven:

Proposition 7.4. *The relationship $\trianglelefteq$ is reflexive and transitive, and $\nexists \varsigma, U \triangleleft \varsigma$ or $\varsigma \triangleleft I$.*

Proof. Relationship $\trianglelefteq$ is based on majorisation, which is reflexive and transitive. For all n-sized vectors x: $(\frac{1}{n}, \ldots, \frac{1}{n}) \prec x \prec (1, 0, \ldots, 0)$. Simple application of Definition 7.3 suffices to prove the proposition. □

7.2 Robustness Analysis of Subjective Rating

Subjectivity difference induces biased ratings, and system designers have already been addressing this problem (see Section 7.3). However, it is an open question whether subjectivity also influences the robustness of trust systems, and whether the influence is positive or negative. We formally study this issue in this section. We start from measuring unfair rating attacks under subjective rating, based on which we measure the robustness of trust systems.

7.2.1 Measuring Attacks

We apply the measure of unfair rating attacks introduced in Chapter 4 – information leakage. Information leakage of a rating R about a truth D quantifies how much information (uncertainty) R reveals (reduces) about D through $f_{n_f, n_d, \sigma}$. Information leakage

of ratings under an attack decides the strength of the attack – less meaning attack being stronger.

Proposition 7.5. *Given strategic behavior α, the information leakage of $f_{n_f,n_d,\sigma}$ is* $I(D;R) = \frac{1}{n_d n_f} \sum_{r,d,f} \mu_{f,d,r} \log \frac{n_d \sum_f \mu_{f,d,r}}{\sum_{d,f} \mu_{f,d,r}}$

Proof. Apply the definition of information leakage. Using the law of total probability, $p(d,r) = p(d) \cdot p(r|d) = p(d) \cdot \sum_f p(f|d) \cdot p(r|d,f)$. Considering independence between D, F, $p(d,r) = \frac{1}{n_d n_f} \sum_f \mu_{f,d,r}$. $p(r) = \sum_{f,d} p(f,d) \cdot p(r|f,d) = \frac{1}{n_d n_f} \sum_{d,f} \mu_{f,d,r}$. □

Proposition 7.5 shows honest advisors' subjectivity (σ) influences strength of a given attack. Hence, subjectivity may influence robustness of a trust system.

In Section 7.2.2, we study a qualitative notion of robustness: "For which p, n_d, n_f, n_r, σ, can the attacker select a strategy α such that no information is leaked?" – we call this an ultimate attack. In Section 7.2.3 we study how the degree of subjectivity affects the amount of information leakage – quantitative robustness.

7.2.2 Ultimate Attacks

If the attacker can select a strategy such that there is zero information leakage of features D given the ratings (ultimate attacks), then the attacker has succeeded in fully disrupting the system. No matter how sophisticated a trust system is, the ratings are useless. 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.

For some values of $p<1, \sigma$, we can select α to get 0 information leakage ($p=1$ is not interesting as no attacker exists):

Theorem 7.6. *Given $f_{n_f,n_d,\sigma}$, $\forall r$, let $\varsigma_{d^*,r} = \max_d \varsigma_{d,r}$, $\exists \alpha$, such that $I(D;R)=0$ iff $p \leq \frac{1}{\sum_r \varsigma_{d^*,r}}$.*

Proof. The information leakage is zero iff D and R are independent, which holds iff $\forall d, r, p(r|d)=p(r|d^*)$. The equation can be represented by $p\varsigma_{d,r} + (1-p)\beta_{d,r} = p\varsigma_{d^*,r} +$

$(1-p)\beta_{d^*,r}$. Sum over r on both sides, we get $p\sum_r \varsigma_{d^*,r} + (1-p)\sum_r \beta_{d^*,r} = 1$. For "if", note that since $\sum_r \beta_{d^*,r} \geq 0$, we can rewrite to $p\sum_r \varsigma_{d^*,r} \leq 1$. For "only if", note that we can set $(1-p)(\beta_{d^*,r} - \beta_{d,r}) = p(\varsigma_{d,r} - \varsigma_{d^*,r})$. □

Intuitively, increasing n_r typically increases $\sum_r \varsigma_{d^*,r}$ (as e.g. it does for σ being the identity matrix). However, if increasing n_r values does not significantly increase $\sum_r \varsigma_{d^*,r}$ (as e.g. it does for σ being the uniform matrix), then increasing n_r does not mitigate ultimate attacks. Therefore, if honest subjective advisors cannot reliably distinguish reporting levels (e.g. 77% versus 78%), then the additional reporting levels are not helpful.

For objective rating $f_{n_d,I}$, the constraint on p to achieve 0 information leakage is $p \leq \frac{1}{n_r}$ [82]. Note that $\frac{1}{\sum_r \varsigma_{d^*,r}} \geq \frac{1}{n_r}$, with equality only if all $\varsigma_{d^*,r}$ equal 1. Even if there are not enough attackers to perform an ultimate attack on an objective-trust system ($\frac{1}{n_r} < p$), there may be sufficiently many attackers to do so on a system with subjectivity, namely when $p \leq \frac{1}{\sum_r \varsigma_{d^*,r}}$. The introduction of subjectivity makes it easier for attackers to completely hide information, thus, leaving a trust system less robust.

Based on the ordering of subjectivity in Definition 7.3, we can formalise the notion that increasingly subjective rating makes it easier for an attacker to achieve ultimate attacks:

Theorem 7.7. *For rating functions with subjectivity ordering being* $f_{n_f,n_d,\sigma} \trianglelefteq f'_{n_f,n_d,\sigma'}$, $\sum_r \varsigma_{d^*,r} \geq \sum_r \varsigma'_{d^*,r}$.

Proof. $f_{n_f,n_d,\sigma} \trianglelefteq f'_{n_f,n_d,\sigma'}$ means $\varsigma \trianglelefteq \varsigma'$. Hence, each column of ς' is majorised by a column of ς. According to Definition 7.2, $\sum_r \max_d \varsigma'_{d,r} \leq \sum_r \max_d \varsigma_{d,r}$. □

Referring to Theorem 7.6, Theorem 7.7 implies that with more subjective rating, the minimal percent of attackers required for ultimate attacks becomes less. Note that $\sum_r \varsigma_{d^*,r} \geq \sum_r \varsigma'_{d^*,r}$ is a necessary but not sufficient condition for $f_{n_f,n_d,\sigma} \trianglelefteq f'_{n_f,n_d,\sigma'}$. A rating function relaxing the condition for ultimate attacks may not necessarily be more subjective, but the function must not be less subjective. Intuitively, subjectivity influences the robustness against ultimate attacks, but there may be other factors.

7.2.3 Quantitative Robustness Comparison

Under non-ultimate attacks (with non-zero information leakage), increasingly subjective rating also facilitates attackers' decreasing information leakage (R' corresponds to $f'(F, D, \alpha')$):

Theorem 7.8. *For any ratings* $f \trianglelefteq f'$*, for any* $f_{n_f,n_d,\sigma}(F, D, \alpha)$*, there* $\exists\alpha'$ *such that* $I(D; R') \leq I(D; R)$.

Proof. From Definition 7.2, each column of ς' is majorised by a column of ς. If $\overrightarrow{\varsigma'_l} \prec \overrightarrow{\varsigma_j}$, let τ_l, τ_j denote the descending order of $\overrightarrow{\varsigma'_l}, \overrightarrow{\varsigma_j}$ respectively, and choose $\beta'_{\tau_l(i),l} = \beta_{\tau_j(i),j}, 0 \leq i < n_d$. As $\nu = p\varsigma + (1-p)\beta$, $\nu' = p\varsigma' + (1-p)\beta'$, it can be easily proved that $\overrightarrow{\nu'_l} \prec \overrightarrow{\nu_j}$. Since $x \log x$ is a convex function, we get $-H(D|R{=}j) \geq -H(D|R{=}l)$ (Olkin and Marshall [99]). Further, via Definition 7.3, $p(l) = p(j) = {}^1\!/_{n_d}|\overrightarrow{\nu'_j}|$, and $H(D)$ equals for f, f'. Hence, we get $I(D; R) \geq I(D; R')$. □

According to the standard notion of information-theoretic robustness, a trust system that permits a strategy that induces lower information leakage is less robust. Using that notion, Theorem 7.8 implies that increasingly subjective rating decreases its robustness. Finally, considering $\forall\varsigma$, $I \trianglelefteq \varsigma$, and that $f_{n_d,I}$ is objective rating, it follows that subjective rating is less robust than objective rating.

7.3 Robustness of Existing Approaches to Deal with Subjectivity

Given a target, subjectivity difference may result in biased ratings, with reduced reference value to advisees. 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. We consider two types of existing ways to treat subjectivity: feature-based rating, which is popularly applied in reality to help to resolve conflicting emphasis on features in global rating, and clustering advisors, which is proposed in literature to

distinguish advisors with different subjectivity. These approaches aim to mitigate subjectivity impacts, so it is interesting to study whether these approaches also improve the robustness of trust systems.

7.3.1 Feature-based Rating

Feature-based rating is a popularly applied rating setting where advisors need to rate each (or a group of) features of a target, instead of providing a global rating for all features. For example, in Booking.com or Expedia.com, consumers can rate multiple features of a hotel after their use, such as cleanliness, comfort, location, facilities, staff. Compared to receiving a global rating, advisees get a more comprehensive view of a target with feature-based rating. Moreover, distinguishing features helps avoid subjectivity induced by advisors' different emphasis on features. We study whether feature-based rating can improve robustness compared with global rating.

An advisee and an advisor care same features in feature-based rating. We alter the model of global rating in Section 7.1 towards feature-based rating, by excluding variable F which denotes features an advisee does not care. A feature-based rating process is how an advisor's properties determine the mapping from a truth D to a rating R.

Definition 7.9. (D, σ)*-feature-based subjective rating* is a rating function with $f_{n_d,\sigma}(d, \alpha)(r) = p\sigma_{d,r} + (1-p)\alpha_{d,r}$, $d{\in}\{0, \ldots, n_d{-}1\}, r{\in}\{0, \ldots, n_r{-}1\}$.

Figure 7.1b presents an example of feature-based rating. There are four outcomes for both D and R. $\sigma_{0,0}$ ($\alpha_{0,3}$) means an honest (dishonest) advisor observes 0 and reports 0 (3).

Below, we look for restrictions on achieving ultimate attacks under feature-based rating, and compare it to that under global rating in Section 7.2.2.

Theorem 7.10. *Given* $f_{n_d,\sigma}(d, \alpha)$ *and* $p{<}1$, $\forall r$, *let* $\sigma_{d^*,r} = \max_d \sigma_{d,r}$, $\exists \alpha$, *such that* $I(D; R){=}0$ *iff* $p \leq \frac{1}{\sum_r \sigma_{d^*,r}}$.

Proof. Proof is similar to that of Theorem 7.6. □

Similarly as in Theorem 7.7, a more subjective $f_{n_d,\sigma}$ has smaller $\sum_r \sigma_{d^*,r}$, meaning it requires less minimal percentage of dishonest advisors to get 0 information leakage – ultimate attacks. Hence, like global rating, more subjective feature-based rating decreases robustness.

Comparing Theorem 7.10 to Theorem 7.6, the constraint on p for ultimate attacks remains the same format in feature-based rating as in global rating. They only differ in signs ($\varsigma_{d^*,r}$ vs. $\sigma_{d^*,r}$), both of which can take any possible values and do not have a fixed magnitude relationship. If $\sum_r \varsigma_{d^*,r} > \sum_r \sigma_{d^*,r}$, the corresponding global rating is more robust than feature-based rating. Moreover, feature-based rating is not necessarily less subjective than global rating, and may be even more subjective, depending on the values of their subjectivity matrixes. For example, for $f_{n_f,n_d,\sigma'}$ and $f_{n_d,\sigma}$, it is possible that $\varsigma' \trianglelefteq \sigma$.

Intuitively, although subjectivity by different emphasis on features is reduced in feature-based rating, subjectivity induced by other factors like various expectations 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.

7.3.2 Clustering Advisors

Thus far, we have proposed 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 7.1.2. The approach is also sufficient for two other cases. The first case is rather unrealistic, where all advisors have the same subjective disposition. The second case is when we have no knowledge about advisors' subjective disposition. That is, each advisor is a complete stranger to the advisee, and therefore indistinguishable and thus equal.

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

7.3.2.1 Modeling

To deal with the general case, we base our analysis on the popular clustering approach. Therein, advisors are assigned to clusters based on their (past) behavior, and each cluster has a different behavior model. Not only can the subjectivity matrices differ from cluster to cluster, we also allow the p-value to differ, to model clustering based on degree of advisors' honesty. Finally, we assume that attackers behave in the worst way – minimising total information leakage.

Assume there are k advisors in total, with R_i representing the rating of i^{th} advisor, and $\overline{R}{=}R_0, \ldots, R_{k-1}$. Let there be clusters $c', c^{\dagger}, \ldots$, with symbols $', {}^{\dagger}, {}^{\S}$ denote association to a cluster. $\overline{R}'$ refers 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 $\overline{C}{=}C_0, \ldots, C_{k-1}$. 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'. Furthermore, we assume that clustering is unrelated to the properties of the target, so $\overline{C}$ is independent of D, F. +We also assume ratings of two advisors are conditionally independent given D, F, regardless of whether they belong to a same cluster: $p(r_i, r_j|D, F){=}p(r_i|D, F) \cdot p(r_i|D, F)$, thus $(\overline{r}', \overline{r}^{\dagger}|D, F){=}p(\overline{r}'|D, F) \cdot p(\overline{r}^{\dagger}|D, F)$.

7.3.2.2 Robustness of clustering

The crucial question is whether clustering increases the robustness of subjective rating. In other words, whether clustering increases the expected information leakage. Clustering indeed increases expected information leakage, no matter what the attacker's behavior is, or the way and the accuracy of clustering are, except in the case where $\overline{C}$ has no impact on the relationship between D and $\overline{R}$:

Theorem 7.11. *$I(D; \overline{R}|\overline{C}) \geq I(D; \overline{R})$, with equality iff $\overline{C}$ and D are conditionally independent under $\overline{R}$.*

Proof. First, note: $I(D; \overline{R}|\overline{C}){=}H(D|\overline{C}){-}H(D|\overline{R}, \overline{C})$, and $I(D; \overline{R}){=}H(D) - H(D|\overline{R})$. Since D and $\overline{C}$ are independent, it suffices to prove $H(D|\overline{R}, \overline{C}) \leq H(D|\overline{R})$.

This is a known property of conditional entropy. Furthermore, equality holds only if $\overline{C}$ and D are conditionally independent under $\overline{R}$. □

Recursive application of Theorem 7.11 shows that if we split existing clusters (cluster again for each individual cluster), we will get more expected information leakage. If we keep splitting existing clusters, until each combination of parameter values has its own cluster, we obtain maximum expected information leakage. Then, not strictly speaking, we are tracking each individual advisor. In Regan et al. [1] and Teacy et al. [11], correlation between the ratings of an advisor and the truth is learned to reinterpret his ratings. Hence, from a information-theoretic view, these approaches are the most robust.

Note that the property of conditional entropy always holds for the condition being a variable but not being any of its outcomes. The specific-conditional entropy $H(D|\overline{R}, \overline{C}{=}\overline{c})$ can be either greater or less than $H(D|\overline{R})$. Therefore, clustering increases the expected information leakage, but a specific clustering may not increase information leakage.

7.3.2.3 Dealing with clusters

We discussed the robustness of clustering advisors. There exist various ways to deal with clusters. Some choose to exclude clusters where the advisors are considered dishonest [29], while some others propose to learn from clusters where advisors are strategic [75]. We study the impact of these different ways on robustness.

Theorem 7.12 states if a cluster has no information leakage about D, then it does not impact the correlation between D and the other clusters.

Theorem 7.12. *If $\overline{R}'$ is independent of D, then $I(\overline{R}^{\dagger}; D|\overline{R}') = I(\overline{R}^{\dagger}; D)$, also $I(\overline{R}^{\dagger}; D|\overline{R}', \overline{R}^{\S}) = I(\overline{R}^{\dagger}; D|\overline{R}^{\S})$.*

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

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

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

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

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

The independence between $\overline{R}'$ and D means $I(D; \overline{R}')=0$. Following immediately from Theorem 7.12, we get Corollary 7.13.

Corollary 7.13. *If* $I(D; \overline{R}')=0$*, then* $I(D; \overline{R}', \overline{R}^{\dagger}, \overline{R}^{\S}) = I(D; \overline{R}^{\dagger}, \overline{R}^{\S})$.

Proof. Based on the chain rule of mutual information, $I(D; \overline{R}', \overline{R}^{\dagger}, \overline{R}^{\S})=I(D; \overline{R}')+I(D; \overline{R}^{\dagger}|\overline{R}')+I(D; \overline{R}^{\S}|\overline{R}', \overline{R}^{\dagger})$. Considering $I(D; \overline{R}')=0$ and Theorem 7.12, the corollary follows straightforwardly. □

Corollary 7.13 implies if a cluster has no information about D, then it can be completely excluded without causing an advisee to lose any information.

Remember there are conditions for a cluster to have 0 information leakage (Theorem 7.6 or 7.10): its percentage of honest advisors p needs to be below a threshold. Throughout the thesis, p is an underlying truth about advisors, which is not obvious to an advisee, and clustering mechanisms may not estimate p accurately. In Noorian [100], advisors considered as dishonest are filtered before clustering. If the true p values of these advisors are above the threshold, then the system loses useful information. In Fang [75], ratings from both clusters of honest and dishonest advisors are exploited using alignment mechanism (i.e., adjust ratings of an advisor based on their correlation with the truth). If the information leakage is 0, then there is no gain from trying to exploit the ratings. Exploiting ratings that are subjective and potentially unfair, is subtle.

Recall Theorem 7.6. It implies that whether a cluster has 0 information leakage depends on p and on the subjectivity of the honest advisors. It is, therefore, not sufficient to look at honesty and subjectivity in isolation, to make decisions.

7.4 Summary

In this chapter, we studied how subjectivity influences robustness of trust systems – specifically the strength of unfair rating attacks – via measuring the information leakage of ratings. We saw that it is insufficient to look at subjectivity and robustness separately, an we focused on their interplay. In that context, we studied whether existing methods, like feature-based rating and clustering, improve robustness.

Three types of subjectivity were included: advisors' different emphasis, expectations and dispositions. We built a conditional probabilistic matrix to model an arbitrary advisor's subjective rating. The model distinguishes features an advisee cares and does not care. This allows for ratings to correlate with each other, without correlating with the truth. Most importantly, we introduced an ordering of subjectivity matrices to compare the degree of subjectivity.

For robustness analysis, we proved that the existence of subjectivity allows attackers to more easily achieve ultimate attacks; and more subjectivity makes it even easier. In fact, more subjective rating decreases information leakage – hence robustness. Feature-based rating exists widely in reality, but we show that it may not improve, and even worsen the robustness. Contrarily, clustering advisors (no matter on what) improves robustness. Finer clustering increases robustness, and in the limit, to learn individual advisor's behavior is the most robust.

Chapter 8

Conclusion and Future Work

In this thesis, we study the robustness of trust systems, especially trust systems, against various types of unfair rating attacks. We also study how subjectivity of honest advisors influence the robustness. In this chapter, we will give a summary of our contributions and also point out several important directions for future work.

8.1 Conclusion

We reviewed the existing approaches for robustness in Chapter 2. Most of them only deal with existing or well-known attacks, without exploring whether there are stronger attacks. This is passive and cannot ensure robustness, considering the uncertainty in attackers' future strategies. We take an active approach, by reasoning all probable strategies of attackers. We argue that to be robust, a system should first be prepared for the worst-case attacks, which cause the minimal utility to its users. As far as we know, our work is the first to study the worst-case unfair rating attacks.

Usually, in current approaches, attacks are measured either heuristically or based on its effect on a specific system, both of which are unfair. With diverse backgrounds, people may have different heuristics. Moreover, different systems may perform differently under same attacks. We propose a unified quantification for unfair rating attacks – information leakage. A stronger attack leads to less information leakage of ratings

about true observations. Such a quantification concerns the correlation between ratings and the truth under attacks. Its application is not limited by specific rating platforms, purposes of using ratings and how ratings are used. In this way, it facilitates the direct comparison of attacks and can be applied across different systems.

Based on the information-theoretic quantification, we study unfair rating attacks from two different angles: 1) whether attackers are independent or collusive in rating, 2) whether their behavior varies over time. By minimizing information leakage, we calculate the worst-case strategies for each type of attacks. And under these strongest attacks, the performance of several existing trust models decreases significantly (see the simulation results in Chapter 4). We also quantify a group of attacks found in the literature, which appear to be far less strong to stress-test the robustness. Meanwhile, we got some non-intuitive results. For example, in independent attacks, when over half advisors are attackers, ratings are still useful even under the worst-case attacks; when attackers' percentage exceeds a certain threshold, they need to sometimes report the truth to minimize information (Chapter 4). In dynamic attacks, although to camouflage requires being honest at the beginning, we prove that to best hide information, attackers should lie more often even initially (Chapter 6).

Understanding attacks guides us to better improve robustness. For example, under some strongest attacks, ratings still have information leakage which can be made use of. Hence, we designed a trust computation mechanism ITC, which presents the best trust evaluation accuracy under the strongest attacks, when being integrated into several existing trust models. In Chapter 5, we obtained that coalition does not effectively help minimize information leakage, but may rather help minimize information leakage about the shape and size of the coalition. Therefore, compared with dealing with independent attackers, the only more effort needed for collusion attacks is to analyze their shapes and sizes. Our study also shows that system parameters may influence the strength of an attack, by managing which, system designers can improve robustness. For example, to increase either rating features or levels makes it harder for attackers to completely hide information (namely to cause zero information leakage).

In the literature, subjectivity and unfair rating attacks are treated orthogonally. Based on the information-theoretic measurement, we find that subjectivity may change

the strength of attacks. This implies that subjectivity may change robustness. We modify the modeling for unfair rating attacks to include subjectivity by 1) distinguishing features that an advisee emphasizes, 2) incorporating probabilities an honest advisor reports various ratings given an observation. We also define an ordering of subjectivity to reason how different degrees of subjectivity change robustness. We find that the introduction of subjectivity weakens robustness: 1) it becomes easier to completely hide information, 2) higher degree of subjectivity makes a system less robust. There exist two ways to mitigate the effect of subjectivity, and we find feature-based rating barely changes robustness, while clustering advisors improves robustness.

8.2 Future Work

We have proposed a novel and effective method to quantify and improve robustness. There remain multiple opportunities to extend our work.

8.2.1 Robustness and Performance

The first extension concerns how to deal with the balance or interaction between the robustness and performance of a trust system. This thesis aims to improve robustness against unfair rating attacks. In practice, besides robustness, performance is another important criterion to evaluate a system[11]. The performance of a trust system depends on the accuracy of trust evaluation[12].

Robustness may contradict with performance. For example, to assume the worst-case attacks greatly increases a system's robustness, but it may decrease the system's performance under weaker attacks, where information available is not effectively used (Section 5.4 in Chapter 5). On the other hand, practically, robustness is often defined in close relation with performance. For example, designers evaluate robustness based on how their systems perform under attacks, where they may have different criteria. To achieve both robustness and good performance – namely to perform the best regardless

[11] As used in Chapter 4, the word performance refers to whether a system fulfills its purpose.
[12] For trust-based systems, performance also depends on the quality of trust-based decision making.

of whether attacks exist – is a common goal for designers. In future, there are several issues we want to study: 1) how to achieve robustness, with the minimal loss of performance, 2) how to design a trust system which satisfies the requirements in 1), and 3) how to propose design suggestions for an arbitrary trust system, to achieve 1).

8.2.2 Other Attacks

In this thesis, we study unfair rating attacks regarding whether attackers are collusive and whether their behavior changes over time. There exist other ways in which malicious advisors launch attacks. For example, an advisor creates and controls several fake accounts to increase the impact of its ratings – a form of Sybil attacks[14]. Advisors may collude with targets. It has already been unveiled in Taobao that some sellers bribe buyers to rate high, or they even collude to create fake transactions to boost reputation.

Besides attacks caused by advisors, there exist other types of attacks in trust systems, as we surveyed in Chapter 2. For example, a target may discriminate in interaction, leading to conflict ratings – discrimination attacks. A target who can easily register a new identity may launch whitewasher attacks. In future, we would like to study how to deal with these various attacks, to make our work for robustness more thorough.

8.2.3 More Application Domains for Information-Theoretic Analysis

The development of Internet has been promoting the sharing of information and various resources. Information sharing plays an important role when people are faced with judgments, choices and other decisions to make. In the age of Big Data, people are not satisfied with basing their analysis purely on experience or instinct. Instead, they tend to rely more and more on data, opinions, or experiences from peers, authorities and other sources, for better decision making.

[14]Sybil attacks can also refer to malicious targets acting in multiple identities to distribute punishment.

Rating is intrinsically a type of information sharing, where advisors share information (opinions) to advisees. There exist many other forms of information sharing, e.g, online text reviewing, data sharing. They may also be analyzed from an information-theoretic view, e.g., to measure information uncertainty with entropy, to measure correlation with information leakage, and to measure information gain/loss with KL-divergence (Definition 3.6 in Chapter 3 for details). In future, we want to generalize the information-theoretic methods to different scenarios of information sharing. In so doing, we hope to broaden their contribution.

Appendix A

List of Publications

1. Dongxia Wang, Tim Muller, Yang Liu, and Jie 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*, pages 511–518. IEEE, 2014

2. Dongxia Wang, Tim Muller, Athirai A Irissappane, Jie Zhang, and Yang Liu. Using information theory to improve the robustness of trust systems. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 791–799. IFAAMAS, 2015

3. 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 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 111–117. AAAI Press, 2015

4. Dongxia Wang. Quantifying and improving robustness of trust systems. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1997–1998. IFAAMAS, 2015

5. Dongxia Wang, Tim Muller, Jie Zhang, and Yang Liu. Is it harmful when advisors only pretend to be honest? In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 2551–2557. AAAI Press, 2016

6. Tim Muller, Dongxia Wang, and Audun Jøsang. Information theory for subjective logic. In *Modeling Decisions for Artificial Intelligence*, pages 230–242. Springer, 2015

7. Tim Muller, Dongxia Wang, Yang Liu, and Jie Zhang. How to use information theory to mitigate unfair rating attacks. In *IFIP International Conference on Trust Management*, pages 17–32. Springer, 2016

Bibliography

[1] Kevin Regan, Pascal Poupart, and Robin Cohen. Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, pages 1206–1212. AAAI Press, 2006.

[2] Siwei Jiang, Jie Zhang, and Yew-Soon Ong. An evolutionary model for constructing robust trust networks. In *Proceedings of the 12th international conference on Autonomous agents and multi-agent systems*, pages 813–820. IFAAMAS, 2013.

[3] Athirai A Irissappane, Siwei Jiang, and Jie Zhang. A biclustering-based approach to filter dishonest advisors in multi-criteria e-marketplaces. In *Proceedings of the 13th international conference on Autonomous agents and multi-agent systems*, pages 1385–1386, 2014.

[4] Riaz Ahmed Shaikh, Hassan Jameel, Brian J d'Auriol, Heejo Lee, Sungyoung Lee, and Young-Jae Song. Group-based trust management scheme for clustered wireless sensor networks. *IEEE Transactions on Parallel and Distributed Systems*, 20(11):1698–1712, 2009.

[5] Fenye Bao, Ray Chen, MoonJeong Chang, and Jin-Hee Cho. Hierarchical trust management for wireless sensor networks and its applications to trust-based routing and intrusion detection. *IEEE transactions on network and service management*, 9(2):169–183, 2012.

[6] Hui Xia, Zhiping Jia, Xin Li, Lei Ju, and Edwin H-M Sha. Trust prediction and trust-based source routing in mobile ad hoc networks. *Ad Hoc Networks*, 11(7): 2096–2114, 2013.

[7] Rafae Bhatti, Elisa Bertino, and Arif Ghafoor. A trust-based context-aware access control model for web-services. In *Web Services, 2004. Proceedings. IEEE International Conference on*, pages 184–191. IEEE, 2004.

[8] Min Li, Xiaoxun Sun, Hua Wang, Yanchun Zhang, and Ji Zhang. Privacy-aware access control with trust management in web service. *World Wide Web*, 14(4): 407–430, 2011.

[9] Parikshit N Mahalle, Pravin A Thakre, Neeli Rashmi Prasad, and Ramjee Prasad. A fuzzy approach to trust based access control in internet of things. In *Wireless Communications, Vehicular Technology, Information Theory and Aerospace & Electronic Systems (VITAE), 2013 3rd International Conference on*, pages 1–5. IEEE, 2013.

[10] W. T. Luke Teacy, Jigar Patel, Nicholas R Jennings, and Michael Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.

[11] W. T. Luke Teacy, Michael Luck, Alex Rogers, and Nicholas R. Jennings. An efficient and versatile approach to trust and reputation using hierarchical bayesian modeling. *Artificial Intelligence*, 193:149–185, 2012.

[12] Dong Chen, Guiran Chang, Dawei Sun, Jiajia Li, Jie Jia, and Xingwei Wang. Trm-iot: A trust management model based on fuzzy reputation for internet of things. *Computer Science and Information Systems*, 8(4):1207–1228, 2011.

[13] Zheng Yan, Peng Zhang, and Athanasios V Vasilakos. A survey on trust management for internet of things. *Journal of network and computer applications*, 42:120–134, 2014.

[14] Munindar P Singh. Cybersecurity as an application domain for multiagent systems. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1207–1212. IFAAMAS, 2015.

[15] Dongxia Wang, Tim Muller, Yang Liu, and Jie 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*, pages 511–518. IEEE, 2014.

[16] Gediminas Adomavicius and Alexander 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*, 17(6):734–749, 2005.

[17] Nurit Gal-Oz, Ehud Gudes, and Danny Hendler. A robust and knot-aware trust-based reputation model. In *IFIP International Conference on Trust Management*, pages 167–182. Springer, 2008.

[18] Yan Lindsay Sun, Zhu Han, Wei Yu, and KJ Ray Liu. A trust evaluation framework in distributed networks: Vulnerability analysis and defense against attacks. In *Proceedings of the 25th International Conference on Computer Communications*, pages 1–13. IEEE, 2006.

[19] Jie Zhang and Robin Cohen. Evaluating the trustworthiness of advice about seller agents in e-marketplaces: A personalized approach. *Electronic Commerce Research and Applications*, 7(3):330–340, 2008.

[20] Jianshu Weng, Zhiqi Shen, Chunyan Miao, Angela Goh, and Cyril Leung. Credibility: How agents can handle unfair third-party testimonies in computational trust models. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(9):1286–1298, 2010.

[21] Siyuan Liu, Jie Zhang, Chunyan Miao, Yin-Leng Theng, and Alex 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*, pages 1151–1152, 2011.

[22] Siyuan Liu, Alex C Kot, Chunyan Miao, and Yin-Leng 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*, pages 99–106. ACM, 2012.

[23] Jie Zhang and Robin Cohen. Design of a mechanism for promoting honesty in e-marketplaces. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI)*, pages 1495–1500. AAAI Press, 2007.

[24] Radu Jurca and Boi Faltings. Obtaining reliable feedback for sanctioning reputation mechanisms. *Journal of Artificial Intelligence Research*, 29:391–419, 2007.

[25] Hui Fang, Yang Bao, and Jie Zhang. Misleading opinions provided by advisors: Dishonesty or subjectivity. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence (IJCAI)*, pages 1983–1989. AAAI Press, 2013.

[26] 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.

[27] 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.

[28] 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.

[29] Zeinab Noorian, Stephen Marsh, and Michael Fleming. Multi-layer cognitive filtering by behavioral modeling. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 871–878. IFAAMAS, 2011.

[30] Tyrone Grandison and Morris Sloman. A survey of trust in internet applications. *Commun. Surveys Tuts*, 3(4):2–16, 2000.

[31] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[32] Donovan Artz and Yolanda Gil. A survey of trust in computer science and the semantic web. *J WEB SEMANT*, 5(2):58–71, 2007.

[33] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, and Dusit Niyato. A survey of trust and reputation management systems in wireless communications. *Proceedings of the IEEE*, 98(10):1755–1772, 2010.

[34] Jin-Hee Cho, Ananthram Swami, and Ray Chen. A survey on trust management for mobile ad hoc networks. *IEEE Communications Surveys and Tutorials*, 13 (4):562–583, 2011.

[35] Han Yu, Zhiqi Shen, Cyril Leung, Chunyan Miao, and Victor R Lesser. A survey of multi-agent trust management systems. *IEEE Access*, 1:35–50, 2013.

[36] Audun Jøsang, Roslan Ismail, and Colin Boyd. A survey of trust and reputation systems for online service provision. *Decision Support Systems*, 43(2):618–644, 2007.

[37] Pradip Lamsal. Understanding trust and security. *Department of Computer Science, University of Helsinki, Finland*, 2001.

[38] Hidehito Gomi. An authentication trust metric for federated identity management systems. In *Proceedings of the 6th International Conference on Security and Trust Management*, STM'10, pages 116–131. Springer-Verlag, 2011.

[39] Seong-Soo Park, Jong-Hyouk Lee, and Tai-Myoung Chung. Authentication scheme based on trust and clustering using fuzzy control in wireless ad-hoc networks. In *Proceedings of the International Conference on Computational Science and Its Applications (ICCSA)*, ICCSA, pages 345–360. Springer-Verlag, 2009.

[40] Ming-Chin Chuang and Jeng-Farn Lee. TEAM: Trust-extended authentication mechanism for vehicular ad hoc networks. In *Proceedings of International Conference on Consumer Electronics, Communications and Networks (CECNet)*, pages 1758–1761. IEEE, 2011.

[41] Anas El Husseini, Abdallah M'hamed, Bachar El Hassan, and Mounir Mokhtari. Trust-based authentication scheme with user rating for low-resource devices in smart environments. *Personal Ubiquitous Comput*, 17(5):1013–1023, 2013.

[42] Rekha Bhatia and Manpreet Singh. Trust based privacy preserving access control in web services paradigm. In *Proceedings of the 2nd International Conference on Advanced Computing, Networking and Security (ADCONS)*, pages 243–246. IEEE, 2013.

[43] Min Li, Hua Wang, and David Ross. Trust-based access control for privacy protection in collaborative environment. In *Proceedings of the International Conference on e-Business Engineering (ICEBE)*, pages 425–430. IEEE, 2009.

[44] Bhavna Gupta, Harmeet Kaur, N Namita, and P Bedi. Trust based access control for grid resources. In *Proceedings of the International Conference on Communication Systems and Network Technologies (CSNT)*, pages 678–682. IEEE, 2011.

[45] Malamati Louta and Angelos Michalas. Trust management framework for efficient service provisioning in dynamic distributed computing environments. In *Proceedings of the Third International Conference on Internet and Web Applications and Services*, ICIW '08, pages 518–523. IEEE Computer Society, 2008.

[46] Ching Lin and Vijay Varadharajan. Mobiletrust: a trust enhanced security architecture for mobile agent systems. *International Journal of Information Security*, 9(3):153–178, 2010.

[47] Khaled M Khan and Qutaibah Malluhi. Establishing trust in cloud computing. *IT professional*, 12(5):20–27, 2010.

[48] Jemal Abawajy. Establishing trust in hybrid cloud computing environments. In *Proceedings of the 10th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 118–125. IEEE, 2011.

[49] David B Johnson and David A Maltz. Dynamic source routing in ad hoc wireless networks. In *Mobile computing*, pages 153–181. Springer, 1996.

[50] Rajiv Mahajan, Surender Singh, Akhilesh Kumar Bhardwaj, and Parveen Sharma. Trust based routing for secure wireless networking solutions. *International Journal of Advanced Research in Computer Science and Software Engineering*, 3:14, 2013.

[51] Ming Xiang. *Trust-based energy aware geographical routing for smart grid communications networks*. PhD thesis, AUT University, 2013.

[52] Theodore Zahariadis, Panagiotis Trakadas, Helen C Leligou, Sotiris Maniatis, and Panagiotis Karkazis. A novel trust-aware geographical routing scheme for wireless sensor networks. *Wireless personal communications*, 69(2):805–826, 2013.

[53] Zhongwei Chen, Ruihua Zhang, Lei Ju, and Wei Wang. Multivalued trust routing based on topology level for wireless sensor networks. In *Proceedings of the 12th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pages 1516–1521. IEEE, 2013.

[54] Bao Fenye, Chen Ing-Ray, Chang MoonJeong, and Jin-Hee Cho. Hierarchical trust management for wireless sensor networks and its applications to trust-based routing and intrusion detection. *IEEE Transactions on Network and Service Management*, 9(2):169–183, 2012.

[55] Yang Han, Keiichi Koyanagi, Takeshi Tsuchiya, Tadashi Miyosawa, and Hiroo Hirose. A trust-based routing strategy in structured p2p overlay networks. In *Proceedings of the International Conference on Information Networking (ICOIN)*, pages 77–82. IEEE, 2013.

[56] Lars Rasmusson and Sverker Jansson. Simulated social control for secure internet commerce. In *Proceedings of the workshop on New security paradigms (NSPW)*, pages 18–25. ACM, 1996.

[57] Cesar Ghali, Ali Chehab, and Ayman Kayssi. Catrac: Context-aware trust-and role-based access control for composite web services. In *Proceedings of 10th International Conference on Computer and Information Technology (CIT)*, pages 1085–1089. IEEE, 2010.

[58] Indrajit Ray, Dieudonne Mulamba, Indrakshi Ray, and Keesook J Han. A model for trust-based access control and delegation in mobile clouds. In *Data and Applications Security and Privacy XXVII*, pages 242–257. Springer, 2013.

[59] Audun Jøsang. Robustness of trust and reputation systems: Does it matter? In *Proceedings of the 6th IFIP International Conference on Trust Management (IFIPTM)*, 2012.

[60] Reid Kerr and Robin Cohen. Smart cheaters do prosper: defeating trust and reputation systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 993–1000. IFAAMAS, 2009.

[61] Chrysanthos Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proceedings of the Second ACM Conference on Electronic Commerce (EC)*, 2000.

[62] Jie Zhang, Robin Cohen, and Kate Larson. Combining trust modeling and mechanism design for promoting honesty in e-marketplaces. *Computational Intelligence*, 28(4):549–578, 2012.

[63] Yuan Liu and Jie Zhang. An incentive mechanism designed for e-marketplaces with limited inventory. *Electronic Commerce Research and Applications*, 2013.

[64] Qinyuan Feng, Yan Lindsay Sun, Ling Liu, Yafei Yang, and Yafei Dai. Voting systems with trust mechanisms in cyberspace: Vulnerabilities and defenses. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22(12):1766–1780, 2010.

[65] Siyuan Liu, Han Yu, Chunyan Miao, and Alex 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.

[66] Li Xiong and Ling Liu. Peertrust: Supporting reputation-based trust for peer-to-peer electronic communities. *IEEE transactions on Knowledge and Data Engineering*, 16(7):843–857, 2004.

[67] Jyoti Grover, Manoj Singh Gaur, and Vijay Laxmi. A novel defense mechanism against sybil attacks in vanet. In *Proceedings of the 3rd international conference on Security of information and networks*, pages 249–255. ACM, 2010.

[68] Weverton Luis da Costa Cordeiro, Flávio Roberto Santos, Gustavo Huff Mauch, Marinho Pilla Barcelos, and Luciano Paschoal Gaspary. Identity management based on adaptive puzzles to protect p2p systems from sybil attacks. *Computer Networks*, 56(11):2569–2589, 2012.

[69] Marek Klonowski and MichaB Koza. Countermeasures against sybil attacks in wsn based on proofs-of-work. In *Proceedings of the 6th International conference on Security and privacy in wireless and mobile networks*, pages 179–184. ACM, 2013.

[70] Giseop Noh, Young-myoung Kang, Hayoung Oh, and Chong-kwon Kim. Robust sybil attack defense with information level in online recommender systems. *Expert Systems with Applications*, 41(4):1781–1791, 2014.

[71] Audun Jøsang. Robustness of trust and reputation systems. In *Proceedings of the 4th International Conference on Self-Adaptive and Self-Organizing Systems Workshop*, pages 159–159. IEEE, 2010.

[72] Bhaskar Mehta. Unsupervised shilling detection for collaborative filtering. In *Proceedings of the 22nd Conference on Artificial Intelligence (AAAI)*, pages 1402–1407, 2007.

[73] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web (WWW)*, pages 285–295, 2001.

[74] Gediminas Adomavicius and Alexander 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)*, 17(6):734–749, 2005.

[75] Hui Fang, Jie Zhang, and Nadia Magnenat Thalmann. Subjectivity grouping: Learning from users' rating behavior. In *Proceedings of the 13th international conference on Autonomous agents and multi-agent systems (AAMAS)*, pages 1241–1248. IFAAMAS, 2014.

[76] Myron Tribus. *Thermostatics and thermodynamics*. Center for Advanced Engineering Study, Massachusetts Institute of Technology, 1961.

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

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

[79] Kullback Solomon. Letter to the editor: The kullback-leibler distance. *The American Statistician*, 41(4):338–341, 1987.

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

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

[82] Dongxia Wang, Tim Muller, Athirai A Irissappane, Jie Zhang, and Yang Liu. Using information theory to improve the robustness of trust systems. In *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 791–799. IFAAMAS, 2015.

[83] Michel Goemans. Chernoff bounds, and some applications. URL http://math.mit.edu/~goemans/18310S15/chernoff-notes.pdf.

[84] Audun Josang and Roslan Ismail. The beta reputation system. In *Proceedings of the 15th bled electronic commerce conference*, pages 41–55, 2002.

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

[86] G. Muller J. Sabater A. Schlosser Z. Topol K. S. Barber J. S. Rosenschein L. Vercouter K. K. Fullam, T. B. Klos and M. Voss. A specification of the agent reputation and trust (art) testbed: experimentation and competition for trust in agent societies. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 512–518. IFAAMAS, 2005.

[87] 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 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 111–117. AAAI Press, 2015.

[88] 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.

[89] 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.

[90] 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.

[91] George Vogiatzis, Ian MacGillivray, and Maria Chli. A probabilistic model for trust and reputation. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 225–232, 2010.

[92] Kevin Hoffman, David Zage, and Cristina Nita-Rotaru. A survey of attack and defense techniques for reputation systems. *ACM Computing Surveys*, 42(1):1, 2009.

[93] Dongxia Wang, Tim Muller, Jie Zhang, and Yang Liu. Is it harmful when advisors only pretend to be honest? In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, 2016.

[94] Sepandar D Kamvar, Mario T Schlosser, and Hector Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the 12th international conference on World Wide Web*, pages 640–651. ACM, 2003.

[95] Lizi Zhang, Siwei Jiang, Jie Zhang, and Wee Keong Ng. Robustness of trust models and combinations for handling unfair ratings. In *Trust Management VI*, pages 36–51. Springer, 2012.

[96] Andrew Oram. *Peer-to-peer: harnessing the benefits of a disruptive technology*. O'Reilly Media, Inc., 2001.

[97] Dimitris Glynos, Patroklos Argyroudis, Christos Douligeris, and Donal O Mahony. Twohop: metric-based trust evaluation for peer-to-peer collaboration environments. In *Global Telecommunications Conference*, pages 1–6. IEEE, 2008.

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

[99] Ingram Olkin and Albert W Marshall. *Inequalities: theory of majorization and its applications*, volume 143. Academic press, 2016.

[100] Zeinab Noorian, Stephen Marsh, and Michael Fleming. Prob-cog: An adaptive filtering model for trust evaluation. In *Proceedings of the IFIP International Conference on Trust Management (IFIPTM)*, pages 206–222. 2011.

[101] Dongxia Wang. Quantifying and improving robustness of trust systems. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1997–1998. IFAAMAS, 2015.

[102] Dongxia Wang, Tim Muller, Jie Zhang, and Yang Liu. Is it harmful when advisors only pretend to be honest? In *Thirtieth AAAI Conference on Artificial Intelligence*, pages 2551–2557. AAAI Press, 2016.

[103] Tim Muller, Dongxia Wang, and Audun Jøsang. Information theory for subjective logic. In *Modeling Decisions for Artificial Intelligence*, pages 230–242. Springer, 2015.

[104] Tim Muller, Dongxia Wang, Yang Liu, and Jie Zhang. How to use information theory to mitigate unfair rating attacks. In *IFIP International Conference on Trust Management*, pages 17–32. Springer, 2016.