Robustness of Trust Models and Combinations for Handling Unfair Ratings

Lizi Zhang, Siwei Jiang, Jie Zhang, Wee Keong Ng

Robustness of Trust Models and Combinations for Handling Unfair Ratings

Lizi Zhang, Siwei Jiang, Jie Zhang, Wee Keong Ng

School of Computer Engineering
Nanyang Technological University, Singapore
{y080077, sjiang1, zhangj, awkng}@ntu.edu.sg

Abstract. In electronic marketplaces, after each transaction buyers will rate the products provided by the sellers. To decide the most trustworthy sellers to transact with, buyers rely on trust models to leverage these ratings to evaluate the reputation of sellers. Although the high effectiveness of different trust models for handling unfair ratings have been claimed by their designers, recently it is argued that these models are vulnerable to more intelligent attacks, and there is an urgent demand that the robustness of the existing trust models has to be evaluated in a more comprehensive way. In this work, we classify the existing trust models into two broad categories and propose an extendable e-marketplace testbed to evaluate their robustness against different unfair rating attacks comprehensively. On top of highlighting the robustness of the existing trust models for handling unfair ratings is far from what they were claimed to be, we further propose and validate a novel combination mechanism for the existing trust models, Discount-then-Filter, to notably enhance their robustness against the investigated attacks.

Key words: Trust models, Unfair ratings, Robustness, Multi-agent system, Electronic marketplaces

1 Introduction

Nowadays, electronic marketplaces (*e.g.*, eBay) have greatly facilitated the transaction processes among different people. However, unlike traditional face-to-face transaction experiences, it is hardly possible for buyers to evaluate the products provided by sellers before they decide whether to buy from a potential seller. Current e-commerce systems like eBay, allow buyers to rate their sellers according to the quality of their delivered products after each transaction is completed.

In the context of the multiagent-based e-marketplace, when a buyer agent evaluates the reputation of a potential seller agent, he may need to ask for other buyers' opinions (advisor[1] agents' ratings) towards that seller agent. We define the following terms discussed in the remaining paper:

[1] When a buyer evaluates a seller, other buyers are that buyer's advisors. The terms *advisor* and *buyer* are used interchangeably in this paper

- *Honest seller*: A seller that delivers his product as specified in the contract.
- *Dishonest seller*: A seller that does not deliver his product as specified in the contract.
- *Reputation*: A value calculated by trust models to indicate whether a seller will behave honestly in the future: the higher reputation, the higher probability that the seller will behave honestly.
- *Positive rating*: A rating given by a buyer/advisor to a seller indicating a seller is an honest seller.
- *Negative rating*: A rating given by a buyer/advisor to a seller indicating a seller is a dishonest seller.
- *Honest buyer/advisor*: A buyer that always provides positive ratings to honest sellers or negative ratings to dishonest sellers.
- *Dishonest buyer/advisor or Attacker*: A buyer that provides negative ratings to honest sellers or positive ratings to dishonest sellers. Exception: some special attacker (*e.g.* Camouflage Attacker) may strategically behave like an honest buyer.
- *Trust or Trustworthiness*[2]: A value calculated by trust models to indicate whether an advisor is honest or not: the higher trustworthiness, the higher probability that the advisor is honest.

Cheating behaviors from sellers, such as not performing the due obligations according to the transaction contract, are still possible to be sanctioned by law if trust models fail to take effect. However, advisors' cheating behaviors, especially providing *unfair ratings* to sellers, are more difficult to be dealt with. Dellarocas distinguished unfair ratings as unfairly high ratings ("ballot stuffing") and unfairly low ratings ("bad-mouthing") [1]. Advisors may collude with certain sellers to boost their reputation by providing unfairly positive ratings while bad-mouthing their competitors' reputation with unfairly negative ratings. An example is that three colluded men positively rated each other several times and later sold a fake painting for a very high price [10].

To address this challenge, researchers in the multiagent-based e-marketplace have designed various trust models to handle unfair ratings to assist buyers to evaluate the reputation of sellers more accurately. Recently it is argued that the robustness analysis of these trust models is mostly done through simple simulated scenarios implemented by the model designers themselves, and this cannot be considered as reliable evidence for how these systems would perform in a realistic environment [4]. If a trust model is not robust against, or vulnerable to, certain unfair rating attack, mostly it will inaccurately rate a dishonest seller's reputation higher than that of an honest seller; thus, it will suggest honest buyers to transact with a dishonest seller, and sellers can gain higher transaction volumes by behaving dishonestly. Therefore, there is an urgent demand to evaluate the robustness of the existing trust models under more comprehensive unfair rating

[2] Generally, the terms *reputation*, *trust* and *trustworthiness* are used interchangeably in many works. To avoid confusion, in this work we use them to model behaviors of sellers and buyers/advisors separately

attack environment before deploying them in the real market. The "Agent Reputation and Trust Testbed (ART) [3] is an example of a testbed that has been specified and implemented by a group of researchers. However, it is currently not flexible enough for carrying out realistic simulations and robustness evaluations for many of the proposed trust models [4].

In this work, we select and investigate four well-known existing trust models (BRS, iCLUB, TRAVOS and Personalized) and six unfair rating attack strategies (Constant, Camouflage, Whitewashing, Sybil, Sybil Camouflage, and Sybil Whitewashing Attack). We classify these trust models into two broad categories: *Filtering-based* and *Discounting-based*, and propose an extendable e-marketplace testbed to evaluate their robustness against different attacks comprehensively and comparatively. To the best of our knowledge, we for the first time experimentally substantiate the presence of their multiple vulnerabilities under the investigated unfair rating attacks. On top of highlighting the robustness of the existing trust models is far from what they were claimed to be—none of the investigated single trust model is robust against all the six investigated attacks, we further propose and validate a novel combination approach, *Discount-then-Filter*, for the existing trust models. This combination notably enhances their robustness against all the attacks: our experiments show most of Discount-then-Filter combined trust models are robust against all the six attacks.

2 Related Work

2.1 Cheating Behavior from Advisors—Unfair Rating Attack

Typical cheating behaviors from sellers, such as *Reputation Lag*, *Value Imbalance*, *Re-entry*, *Initial Window*, and *Exit*, have been studied by Kerr and Cohen [6]. They assumed maximal cheating in their paper: a cheating seller does not ship out his product thus no cost is incurred, and the buyer will learn the results only after the lag has lapsed. Recent work by Jøsang and Golbeck identified more seller attack strategies and reduced all types of advisor cheating behaviors to Unfair Rating Attack [4]. Particularly, Kerr and Cohen found combined seller attacks are able to defeat every investigated trust model. Researchers, especially those models' designers, might be tempted to argue that, cheating behaviors from sellers are possible to be handled by law and their models are still robust against advisors' unfair rating attack rather than sellers' attack strategies.

We argue that even though cheating behaviors from sellers are possible to be sanctioned by law, advisors' cheating behaviors are still able to defeat the existing trust models; thus, improving the robustness of the existing trust models for handling unfair ratings is urgently demanded. To begin with, online transactions are essentially contracts: sellers are obliged to deliver products as specified by themselves and buyers are obliged to pay the specified amount of money. Therefore, most sellers' cheating behaviors can be considered as unlegal: in the real life, it is very common that buyers may sue their sellers if the delivered products are not as good as specified by the sellers according to the contract law. Although sellers' cheating behaviors can be sanctioned by law, advisors' unfair

ratings can only be considered as unethical rather than unlegal [4], therefore there is an urgent demand to address the unfair rating problem. Our paper focuses on cheating behaviors from advisors and below are a list of typical unfair rating attacks[3] that may threaten the existing trust models in e-marketplaces.

Constant Attack The simplest strategy from dishonest advisors is, constantly providing unfairly positive ratings to dishonest sellers while providing unfairly negative ratings to honest sellers. This simple attack is a baseline to test the basic effectiveness of different trust models in dealing with unfair ratings.

Camouflage Attack Dishonest advisors may camouflage themselves as honest ones by providing fair ratings strategically. For example, advisors may provide fair ratings to build up their trustworthiness (according to certain trust models) at the early stage before providing unfair ratings. Intuitively, trust models assuming attackers' behaviors are constant and stable may be vulnerable to it.

Whitewashing Attack In e-marketplaces, it is hard to establish buyers' identities: users can freely create a new account as a buyer. This presents an opportunity for a dishonest buyer to *whitewash* his low trustworthiness (according to certain trust models) by starting a new account with the default initial trustworthiness value (0.5 in our investigated trust models).

Sybil Attack When evaluating the robustness of trust models, it is usually assumed that the majority of buyers are honest. In our experiments, the aforementioned three types of attackers are minority compared with the remaining honest buyers. However, it is possible that dishonest buyers (unfair rating attackers) may form the majority of all the buyers in e-marketplaces. In this paper, we use the term *Sybil Attack*, which was initially proposed by Douceur, to describe the scenario where dishonest buyers have obtained larger amount of resources (buyer accounts) than honest buyers to constantly provide unfair ratings to sellers [2]. This attack can be considered as, dishonest buyers are more than honest buyers and they perform Constant Attack together.

Sybil Camouflage Attack As the name suggests, this attack combines both Camouflage Attack and Sybil Attack: dishonest buyers are more than honest buyers and perform Camouflage Attack together.

Sybil Whitewashing Attack Similar to Sybil Camouflage Attack: dishonest buyers are more than honest buyers and perform Whitewashing Attack together.

Non-Sybil-based and Sybil-based Attack Obviously, under the Constant, Camouflage and Whitewashing Attack, the number of dishonest buyers is less than half of all the buyers in the market (minority). We refer to them as the *Non-Sybil-based Attack*. On the contrary, the number of Sybil, Sybil Camouflage, and

[3] Some attack names are used interchangeably in both seller attacks and advisors' unfair rating attacks (*e.g.*, Sybil Attack), in this paper we refer to the latter

Sybil Whitewashing Attackers is greater than half of all the buyers (majority), and these attacks are referred to as the *Sybil-based Attack*.

2.2 Trust Models for Handling Unfair Rating—Defense Mechanisms

Various trust models have been proposed to deal with different attacks. In the interest of fairness, we select four representative models proposed during the year 2002—2011 that self-identified as applicable to e-marketplaces and robust against unfair rating attacks. In this section, we also classify them into two broad categories: *Filtering-based* and *Discounting-based*.

Beta Reputation System (BRS) The Beta Reputation system (BRS) was proposed by Jøsang and Ismail to predict a seller's behavior in the next transaction based on the number of honest and dishonest transactions (the two events in the beta distribution: $[p, n]$, where p and n denote the number of received positive and negative ratings) he has conducted in the past [5]. Whitby *et al.* further proposed an iterative approach to filter out unfair ratings based on the *majority rule* [9]. According to this approach, if the calculated reputation of a seller based on the set of honest buyers (initially all buyers) falls in the rejection area (q quantile or $1 - q$ quantile) of the beta distribution of a buyer's ratings to that seller, this buyer will be filtered out from the set of honest buyers and all his ratings will be considered as unfair ratings since his opinions (ratings) are not consistent with the majority of the other buyers' opinions (the majority rule). Then the seller's reputation will be re-calculated based on the updated set of honest buyers, and the filtering process continues until the set of honest buyers eventually remains unchanged. Obviously, the majority rule renders BRS vulnerable to Sybil-based Attack because the majority of buyers are dishonest and the other honest buyers' (the minority) ratings will be filtered out.

iCLUB iCLUB is a recently proposed trust model in handling multi-nominal ratings [7]. It adopts the clustering approach and considers buyers' local and global knowledge about sellers to filter out unfair ratings. For local knowledge, the buyer compares his ratings with advisors' ratings (normalized rating vectors) towards the *target seller* (the seller under evaluation) by clustering. If an advisor's ratings are not in the cluster containing the buyer's ratings, they will be considered as not consistent with the buyer's opinions, and will be filtered out as unfair ratings. Obviously, comparing advisors' ratings with the buyer's own opinions is reliable since the buyer never lies to himself. If transactions between the buyer and the target seller are too few (few evidence), the buyer will not be confident to rely on his local knowledge, and global knowledge will be used. The buyer will compare his and the advisors' ratings towards all the sellers excluding the target seller by performing clustering. A set of advisors who always have similar ratings with the buyer (in the same cluster) towards every seller are identified. Eventually, these advisors are used to filtered out the other untrustworthy advisors' ratings when evaluating all advisors' ratings to the target seller. In general, buyers' local knowledge is more reliable than his global knowledge.

This is because when the set of advisors whose opinions are always similar to the buyer's cannot be found, the global knowledge will use the majority rule to filter out unfair ratings; this may be vulnerable to Sybil-based Attack.

Filtering-based Trust Models BRS and iCLUB filter out unfair ratings before aggregating the remaining fair ratings in evaluating a seller's reputation, therefore, we classify them as **Filtering-based**. The reputation of the seller S, $\Gamma(S)$, is calculated as:

$$\Gamma(S) = \frac{\sum p_i + 1}{\sum p_i + \sum n_i + 2} \tag{1}$$

where p_i and n_i are the number of positive and negative ratings from each advisor i to the seller S after unfair ratings are filtered out. When S does not receive any ratings, his initial reputation is 0.5.

TRAVOS Teacy *et al.* proposed TRAVOS to evaluate the trustworthiness of advisors, τ_i, and use τ_i to discount their ratings before aggregating these ratings to evaluate the target seller's reputation [8]. To evaluate an advisor's trustworthiness, first, a set of reference sellers are identified if these sellers' reputation are similar to the target seller's reputation as calculated by using this advisor's ratings towards them. Then the buyer will use the cumulative distribution function of beta distribution based on the total number of his positive and negative ratings to each reference seller to compute the trustworthiness of that advisor. Compared with BRS, TRAVOS incorporates a buyer's personal transaction experiences with the target seller in the process of evaluating his advisors' trustworthiness. However, TRAVOS assumes the advisors' behaviors are constant; thus, this model may be vulnerable if the attackers camouflage themselves by giving fair ratings strategically before providing unfair ratings.

Personalized Zhang and Cohen proposed a personalized approach to evaluate an advisor's trustworthiness τ_i in two aspects: private and public trust [10]. To evaluate the private trust of an advisor, the buyer compares his ratings with the advisor's ratings to their commonly rated sellers. Greater disparity in the comparison indicates discounting of the advisor's trustworthiness to a larger extent. Similarly, the public trust of an advisor is estimated by comparing the advisor's ratings with the majority of the other advisors' ratings towards their commonly rated sellers. Obviously, public trust adopts the majority rule in evaluating an advisor's trustworthiness and therefore may be vulnerable to Sybil-based Attack. Since private trust is more reliable, when aggregating both private and public trust of an advisor, this model will allocate higher weightage to private trust if the buyer has more commonly rated sellers with the advisor (more evidence). When the number of such commonly rated sellers exceeds a certain threshold value (enough evidence), the buyer will only use the private trust to evaluate the advisor's trustworthiness more accurately.

Discounting-based Trust Models TRAVOS and Personalized calculate advisors' trustworthiness and use their trustworthiness to discount their ratings before aggregating them to evaluate a seller's reputation. Thus, we classify them as **Discounting-based**. The reputation of the seller S, $\Gamma(S)$, is calculated as:

$$\Gamma(S) = \frac{\sum \tau_i \times p_i + 1}{\sum \tau_i \times p_i + \sum \tau_i \times n_i + 2} \tag{2}$$

where p_i and n_i are the number of positive and negative ratings from each advisor i to the seller S, and τ_i is the trustworthiness of the advisor i. When S does not receive any ratings, his initial reputation is 0.5.

3 Evaluation Method

3.1 The E-marketplace Testbed

Our experiments are performed by simulating the transaction activities in the e-marketplace. As mentioned in Section 1, the existing ART testbed is not suitable for carrying out experiments to compare robustness of trust models under different unfair rating attacks. In light of its limitations, we design and develop an e-marketplace testbed, which is extendable via incorporating new trust or attack models.

In our e-marketplace testbed, there are 10 dishonest sellers and 10 honest sellers. To make the comparison more obvious, we consider a "Duopoly Market": there are two sellers in the market that take up a large portion of the total transaction volume in the market. We assume a reasonable competition scenario: one duopoly seller (*dishonest duopoly seller*) tries to beat his competitor (*honest duopoly seller*) in the transaction volume by hiring or collaborating with dishonest buyers to perform unfair rating attacks. We refer to the remaining sellers (excluding the duopoly sellers) as *common sellers*. Typically, trust models are most effective when 30% of buyers are dishonest [9]. To ensure the best case for the trust models, we added 6 dishonest buyers (attackers) and 14 honest buyers in the market for Non-Sybil-based Attack, and switch their values for Sybil-based Attack. The entire simulation will last for 100 days. On each day, each buyer chooses to transact with one seller once. Since most trust models are more effective when every advisor has transaction experiences with many different sellers, we assume that there is a probability of 0.5 that buyers will transact with the duopoly sellers while there is another probability of 0.5 that buyers will transact with each common seller randomly. The value of 0.5 also implies the duopoly sellers take up half of all the transactions in the market. When deciding on which duopoly seller to transact with, honest buyers use trust models to calculate their reputation values and transact with the one with the higher value, while dishonest buyers choose sellers according to their attacking strategies. After each transaction, honest buyers provide fair ratings, whereas dishonest buyers provide ratings according to their attacking strategies.

The key parameters with their values in the e-marketplace testbed are summarized as follows:

- *Number of honest duopoly seller*: 1
- *Number of dishonest duopoly seller*: 1
- *Number of honest common seller*: 9
- *Number of dishonest common seller*: 9
- *Number of honest buyer/advisor ($|B^H|$)*: 14 (Non-Sybil-based Attack) or 6 (Sybil-based Attack)
- *Number of dishonest buyer/advisor or attacker ($|B^D|$)*: 6 (Non-Sybil-based Attack) or 14 (Sybil-based Attack)
- *Number of simulation days (L)*: 100
- *The ratio of duopoly sellers' transactions to all transactions (r)*: 0.5

3.2 The Trust Model Robustness Metric

To evaluate the robustness of different trust models, we compare the transaction volumes of the duopoly sellers. Obviously, the more robust the trust model, the larger the transaction volume difference between the honest and dishonest duopoly seller. The robustness of a trust model (defense, Def) against an attack model (Atk) is defined as:

$$\Re(Def, Atk) = \frac{|Tran(S^H)| - |Tran(S^D)|}{|B^H| \times L \times r} \tag{3}$$

where $|Tran(S^H)|$ and $|Tran(S^D)|$ denote the total transaction volume of the honest and dishonest duopoly seller, and the values of key parameters in the e-marketplace testbed $|B^H|$, L, and r are given in Section 3.1.

If a trust model Def is *completely robust* against a certain attack Atk, $\Re(Def, Atk) = 1$. It means the reputation of the honest duopoly seller is always higher than that of the dishonest duopoly seller as calculated by the trust model, so honest buyers will always transact with the honest duopoly seller. On the contrary, $\Re(Def, Atk) = -1$ indicates, the trust model always suggests honest buyers to transact with the dishonest duopoly seller, and Def is *completely vulnerable* to Atk. When $\Re(Def, Atk) > 0$, the greater the value is, the more robust Def is against Atk. When $\Re(Def, Atk) < 0$, the greater the absolute value is, the more vulnerable Def is to Atk[4].

In Eq. 3, the denominator denotes the transaction volume difference between the honest and dishonest duopoly seller when the trust model (Def) is completely robust against or vulnerable to a certain attack (Atk): all the honest buyers (B^H) always transact with the duopoly honest seller (S^H, when completely robust) or duopoly dishonest seller (S^D, when completely vulnerable) in the 100 days with a probability of 0.5 to transact with the duopoly sellers. In our experiment, the denominator is 700 ($14 \times 100 \times 0.5$) if Atk is Non-Sybil-based Attack, or 300 ($6 \times 100 \times 0.5$) if Atk is Sybil-based Attack.

[4] When Def is completely robust against or vulnerable to Atk, in our experiments $\Re(Def, Atk)$ can be slightly around 1 or -1 because the probability to transact with the duopoly sellers may not be exactly 0.5 in the actual simulation process.

Table 1. Robustness of single trust models against attacks. Every entry denotes the mean and standard deviation of the robustness values of trust model against attack

	Constant	Camouflage	Whitewashing	Sybil	Sybil Cam	Sybil WW
BRS	0.84±0.03	0.87±0.04	-0.48±0.08	-0.98±0.09	-0.63±0.08	-0.60±0.10
iCLUB	1.00±0.04	0.98±0.03	0.81±0.10	-0.09±0.33	0.95±0.11	-0.16±0.26
TRAVOS	0.96±0.04	0.88±0.04	0.98±0.04	0.66±0.10	-0.60±0.09	-1.00±0.08
Personalized	0.99±0.04	1.01±0.03	0.99±0.04	0.84±0.12	0.67±0.09	-1.00±0.11

*Sybil Cam: Sybil Camouflage Attack; Sybil WW: Sybil Whitewashing Attack

4 Robustness of Single Trust Models

In this section, we evaluate the robustness of all the trust models against all the attack strategies covered in Section 2 with the e-marketplace testbed described in Section 3. In our experiments, when models require parameters we have used values provided by the authors in their own works wherever possible. The experiments are performed 50 times, and the mean and standard deviation of the 50 results are shown in Table 1 in the form of $(mean \pm std)$. We discuss the robustness of all the single trust models against each attack.

Constant Attack All the trust models are robust against this baseline attack. Consistent with Whitby *et al.*'s experimental results, our experiment also shows BRS is not completely robust against Constant Attack [9]. Fig. 1 and Fig. 2 depict under Constant Attack, how the transactions of the duopoly sellers grow day after day when BRS and iCLUB are used by honest buyers to decide which duopoly seller to transact with. The transaction volume difference between the honest and dishonest duopoly seller on Day 100 (around 700) indicates that iCLUB is completely robust against Constant Attack. Space prevents the inclusion of such figures for every trust model; throughout this paper, all key data are presented in Table 1—2 and we use charts where illustration is informative.

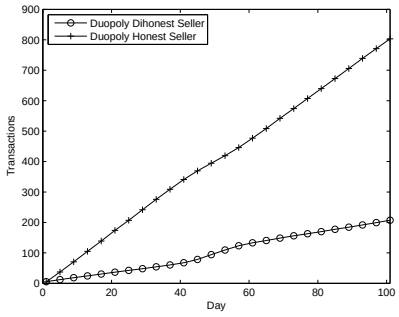

Fig. 1. BRS vs. Constant

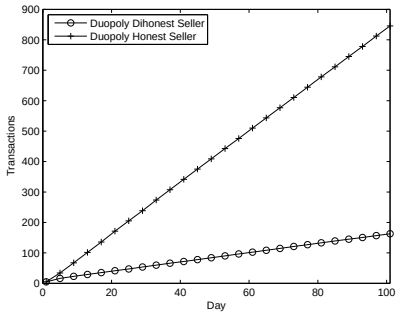

Fig. 2. iCLUB vs. Constant

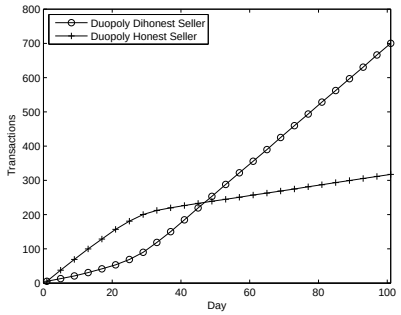

Fig. 3. BRS vs. Whitewashing

Camouflage Attack In this experiment, Camouflage Attackers give fair ratings to all the common sellers to establish their trustworthiness before giving unfair ratings to all sellers (with a probability of 0.5 to transact with the duopoly sellers). From the results of Table 1, without enough attackers, Camouflage Attack does not threaten the trust models very much.

Whitewashing Attack In our experiment, each Whitewashing Attacker provides one unfair rating on one day and starts with a new buyer account on the next day. The value $\Re(BRS, Whitewashing) = -0.48$ in Table 1 shows BRS is vulnerable to this attack. According to Fig. 3, the honest duopoly seller has more transactions than the dishonest one at the beginning. However, after some time (around Day 45) the dishonest duopoly seller's transaction volume exceeds his competitor. In fact, after some time the calculated reputation of a seller will more easily fall in the rejection area of the beta distribution of an honest buyer's single accumulated ratings (single $[p, 0]$ to an honest seller and single $[0, n]$ to a dishonest seller, where p and n become very large as transaction experiences accumulate) rather than Whitewashing Attackers' multiple one-transaction ratings (multiple $[0, 1]$ to an honest sellers and multiple $[1, 0]$ to a dishonest seller). The other trust models are robust against Whitewashing Attack.

Sybil Attack As described in Section 2, BRS is completely vulnerable to Sybil Attack due to its employed majority-rule. The robustness of iCLUB is not stable as indicated by its standard deviation of 0.33. To explain, an honest buyer can rely on his local knowledge to always transact with one duopoly seller while using the global knowledge, which is wrong when majority of advisors are attackers, to evaluate the reputation of the other duopoly seller. The duopoly seller to always transact with can be either honest or dishonest as long as his reputation is always higher than that of his competitor, which is possible in either case. Besides, TRAVOS and Personalized are not completely robust against Sybil Attack. This is due to the lack of transactions among different buyers and sellers at the beginning. For TRAVOS, at the beginning it is hard to find common reference sellers for the buyer and the advisor so the discounting is not effective (we refer to this

phenomenon as *soft punishment*). When majority are dishonest buyers, their aggregated ratings will overweigh honest buyers' opinions. For instance, if the trustworthiness of each dishonest and honest buyer are 0.4 and 0.6, and all buyers provides only one rating to a particular seller, according to Eq. 2, the reputation of an honest seller is $0.41 < 0.5$ ($0.41 = (0.6 \times 6 + 1)/(0.4 \times 14 + 0.6 \times 6 + 2)$) and that of a dishonest seller is $0.59 > 0.5$ ($0.59 = (0.4 \times 14 + 1)/(0.4 \times 14 + 0.6 \times 6 + 2)$); both suggest inaccurate decisions. However, if a Discounting-based model is able to discount the trustworthiness of a dishonest buyer to a larger extent, say 0.1, while promote that of an honest buyer to a larger extent, say 0.9, the evaluation of sellers' reputation will become accurate. For Personalized, at the beginning the buyer will more rely on public trust to evaluate the trustworthiness of an advisor, which is inaccurate when majority of buyers are dishonest. Fig. 4 and Fig. 5 show that, as transactions among different buyers and sellers grow, TRAVOS becomes more effective in discounting advisors' trustworthiness and Personalized tends to use private trust to accurately evaluate advisors' trustworthiness.

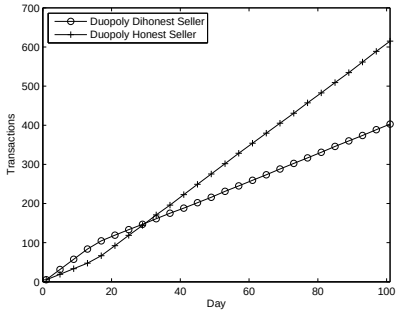

Fig. 4. TRAVOS vs. Sybil

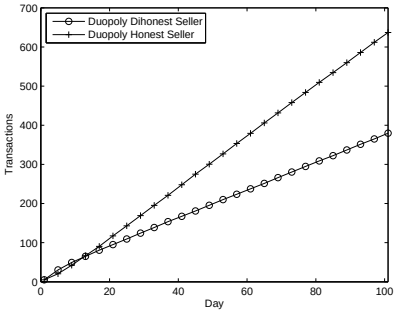

Fig. 5. Personalized vs. Sybil

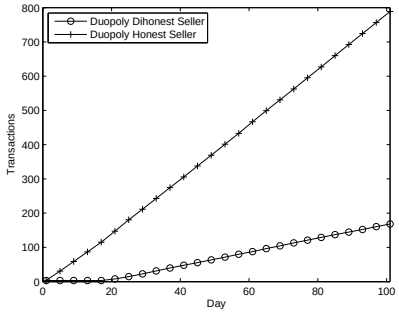

Fig. 6. TRAVOS vs. Camouflage

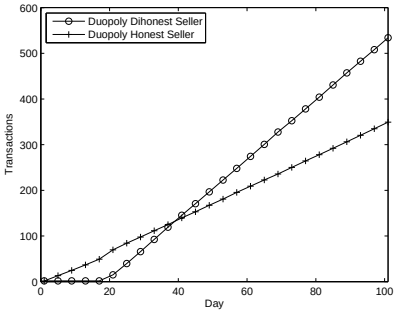

Fig. 7. TRAVOS vs. Sybil Camouflage

Sybil Camouflage Attack Unlike Sybil Attack, Sybil Camouflage Attack is unable to render BRS completely vulnerable. This is because at the beginning

attackers camouflage themselves as honest ones by providing fair ratings, where BRS is always effective. After attackers stop camouflaging, the duopoly dishonest seller's transaction volume will soon exceed his competitor. For iCLUB, during the camouflaging stage, the honest duopoly seller will only transact with honest buyers. After attackers stop camouflaging, only the reliable local knowledge will be used by honest buyers to evaluate the trustworthiness of the honest duopoly seller (of high value), and honest buyers will continue to transact with him. Compared with Camouflage and Sybil Attack, Personalized becomes less robust against Sybil Camouflage Attack. This is because the public and private trust of attackers have not been discounted to a large extent right after they complete the camouflaging stage (soft punishment). When the majority are attackers, their aggregated ratings will overweigh honest buyers' opinions. After attackers stop camouflaging, their private trust will continue to drop and Personalized will be effective. Compared with Camouflage Attack, TRAVOS becomes vulnerable to Sybil Camouflage Attack: although TRAVOS will inaccurately promote the trustworthiness of a Camouflage Attacker (most are slightly larger than 0.5), when majority are honest buyers, the aggregated ratings from attackers are still not able to overweigh honest buyers' opinions. However, under Sybil Camouflage Attack, when majority are dishonest buyers, these attackers' aggregated ratings will easily overweigh honest buyers' opinions and render TRAVOS vulnerable. Fig. 6 and Fig. 7 clearly show the difference of the robustness of TRAVOS against Camouflage Attack and Sybil Camouflage Attack.

Sybil Whitewashing Attack This is the strongest attack: it can defeat every single trust model as observed from Table 1. Similar to Sybil Attack, the robustness of iCLUB against Sybil Whitewashing Attack is still not stable. Compared with Whitewashing Attack, BRS is still vulnerable to Sybil Whitewashing Attack while TRAVOS and Personalized change dramatically from completely robust to completely vulnerable. For TRAVOS, since every whitewashing attacker provides only one rating to a duopoly seller, buyer cannot find reference seller to effectively discount the trustworthiness of whitewashing attackers to a large extent. When majority are soft punished dishonest buyers, TRAVOS will always suggest honest buyers to transact with the dishonest duopoly seller. For Personalized, since every whitewashing attacker provides only one rating to a duopoly seller, the buyer cannot find enough commonly rated sellers and will heavily rely on public trust to evaluate the trustworthiness of an advisor, which is inaccurate when majority of buyers are dishonest. Therefore, similar to TRAVOS, the trustworthiness of whitewashing attacker cannot be discounted to a large extent and the soft punishment renders Personalized completely vulnerable.

It is also noted that although discounting-based TRAVOS and Personalized are robust against Whitewashing, Camouflage, and Sybil Attack, their robustness drops to different extents when facing Sybil Whitewashing and Sybil Camouflage Attack. Based on our results demonstrated in Table 1, we conclude that, none of our investigated single trust models is robust against all the six attacks. Therefore, there is a demand to address the threats from all these attacks.

Table 2. Robustness of combined trust models against attacks. Every entry denotes the mean and standard deviation of the robustness values of trust model against attack

	Constant	Camouflage	Whitewashing	Sybil	Sybil Cam	Sybil WW
Filter-then-Discount						
BRS + TRAVOS	0.89±0.06	0.87±0.03	-0.55±0.10	-1.01±0.11	-0.55±0.09	-0.59±0.11
BRS + Personalized	0.89±0.06	0.88±0.03	-0.34±0.05	-0.96±0.07	-0.53±0.08	-0.58±0.08
iCLUB + TRAVOS	0.96±0.03	0.98±0.04	0.95±0.04	0.85±0.08	0.97±0.10	0.70±0.12
iCLUB + Personalized	0.98±0.03	0.99±0.03	0.92±0.06	0.88±0.13	0.98±0.09	0.67±0.13
Discount-then-Filter						
TRAVOS + BRS	0.95±0.03	0.86±0.06	0.98±0.04	0.91±0.06	-0.57±0.12	0.98±0.10
TRAVOS + iCLUB	0.95±0.04	0.92±0.03	0.93±0.03	0.91±0.12	0.91±0.10	0.94±0.12
Personalized + BRS	0.99±0.03	0.98±0.03	1.01±0.03	0.96±0.11	0.87±0.08	1.00±0.10
Personalized + iCLUB	0.97±0.04	0.95±0.02	0.98±0.04	0.92±0.09	0.94±0.09	0.93±0.07

*Sybil Cam: Sybil Camouflage Attack; Sybil WW: Sybil Whitewashing Attack

5 Robustness of Combined Trust Models

5.1 Combining Trust Models

Based on the results of Table 1, Discounting-based trust models may change from vulnerable to robust if some attackers' ratings can be filtered out by Filtering-based models to reduce the effect of Sybil-based Attack to that of Non-Sybil-based Attack. On the other hand, based on analysis in Section 4, under most attacks Discounting-based models are still able to discount the trustworthiness of dishonest buyers to lower than 0.5 (although only slightly). Intuitively, filtering out ratings from advisors with lower trustworthiness may be a promising pre-filtering step before using Filtering-based models. Therefore, we combine trust models from different categories to evaluate their new robustness to the same set of attacks. Generally, there are two approaches for combination: **Filter-then-Discount** and **Discount-then-Filter**. Details are given below.

Approach 1—Filter-then-Discount:

1. Use a Filtering-based trust model to filter out unfair ratings;
2. Use a Discounting-based trust model to aggregate discounted ratings to calculate sellers' reputation.

Approach 2—Discount-then-Filter:

1. Use a Discounting-based trust model to calculate each advisor i's trustworthiness τ_i;
2. If $\tau_i < \epsilon$, remove i's all ratings ($\epsilon = 0.5$ in our experiment);
3. Use a Filtering-based trust model to filter out unfair ratings before aggregating the remaining ratings to calculate sellers' reputation.

5.2 Robustness Evaluation

Eight possible combinations of trust models are obtained and their robustness against all the attacks have been evaluated. Notice that the new model name follows the order of using the two different models. We will discuss the robustness enhancement of each combined model against all attacks based on the experimental results presented in Table 2.

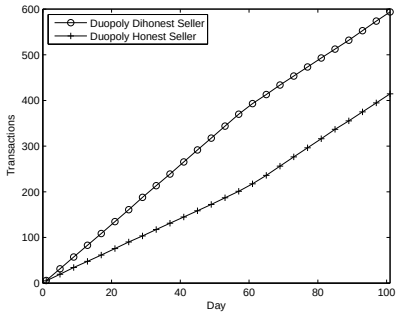

Fig. 8. BRS vs. Sybil WW

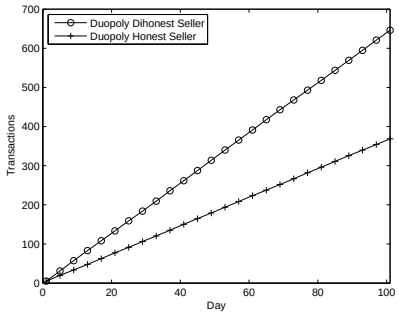

Fig. 9. Personalized vs. Sybil WW

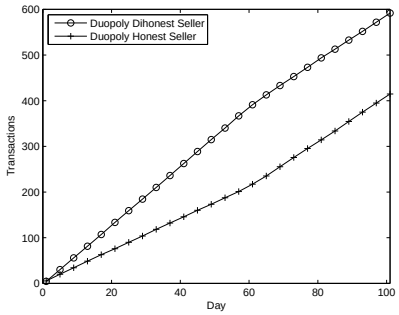

Fig. 10. BRS + Personalized vs. Sybil WW

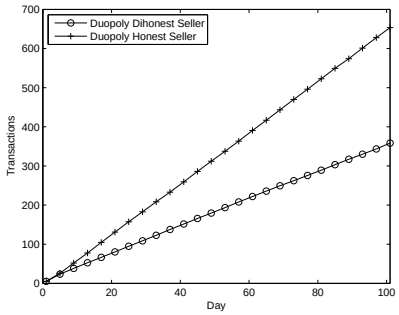

Fig. 11. Personalized + BRS vs. Sybil WW

BRS + TRAVOS and BRS + Personalized Similar to BRS, they are still vulnerable to many attacks such as Whitewashing, Sybil, Sybil Whitewashing, and Sybil Camouflage Attack. The reason is, under these attacks BRS will inaccurately filter out some honest buyers' ratings and keep some dishonest buyers' ratings after the first step of Approach 1; the remaining unfair ratings will be used by Discounting-based trust models to inaccurately suggest honest buyers to transact with the dishonest duopoly seller.

iCLUB + TRAVOS and iCLUB + Personalized Contrary to BRS, iCLUB is robust against Whitewashing and Sybil Camouflage Attack. Therefore, iCLUB + TRAVOS and iCLUB + Personalized are also able to effectively filter out unfair ratings at the first step of Approach 1, and are robust against these attacks. However, due to the instability of the robustness of iCLUB against Sybil and Sybil Whitewashing Attack, iCLUB + TRAVOS and iCLUB + Personalized are still not completely robust against these attacks.

Discount-then-Filter The complete robustness of TRAVOS and Personalized against Whitewashing Attack ensures all the attackers' ratings will be filtered out

at the first step of Approach 2. As described in Section 4, although TRAVOS and Personalized are unable to discount the trustworthiness of a Sybil, Sybil Whitewashing or Sybil Camouflage Attacker to a large extent (soft punishment: only slightly lower than 0.5), the threshold value we choose ($\epsilon = 0.5$) is able to filter out all these attackers' ratings at the second step of Approach 2. Therefore, Personalized + BRS and Personalized + iCLUB are completely robust against Sybil, Sybil Whitewashing and Sybil Camouflage Attack. Likewise, TRAVOS + BRS and TRAVOS + iCLUB are completely robust against most attacks. One exception is that, TRAVOS + BRS is still vulnerable to Sybil Camouflage Attack. This is because TRAVOS inaccurately promotes attackers' trustworthiness (most are slightly higher than 0.5) and their ratings are unable to be filtered out at the second step of Approach 2. Unlike iCLUB, which is robust against Sybil Camouflage Attack, BRS is vulnerable to it.

Based on the results in Table 1—2, we conclude that, robustness of single trust models can be enhanced by combining different categories, and Discount-then-Filter is most robust. Particularly, TRAVOS + iCLUB, Personalized + BRS, and Personalized + iCLUB are robust against all the investigated attacks. Fig. 8-11 show how the robustness of the trust models is enhanced with the Discount-then-Filter approach, while Filter-then-Discount is still vulnerable.

6 Conclusion and Future Work

Trust models can benefit us in choosing trustworthy sellers to transact with in the e-marketplace only when they are robust against external unfair rating attacks. Recently it is argued some trust models are vulnerable to certain attacks and they are not as robust as what their designers claimed to be. Therefore, robustness of trust models for handling unfair ratings have to be evaluated under a comprehensive attack environment to make the results more credible.

In this paper, we designed an extendable e-marketplace testbed to incorporate each existing trust model under a comprehensive set of attack models to evaluate the robustness of trust models. To the best of our knowledge, this is the first demonstration that multiple vulnerabilities of trust models for handling unfair ratings do exist. We conclude that, in our experiments there is no single trust model that is robust against all the investigated attacks. While we have selected a small number of trust models for this initial study, we can hardly believe that other trust model will not have these vulnerabilities. We argue that, in the future any newly proposed trust model at least has to demonstrate robustness (or even complete robustness) to these attacks before being claimed as effective in handling unfair ratings. To address the challenge of existing trust models' multiple vulnerabilities, we classified existing trust models into two categories: Filtering-based and Discounting-based, and further proposed two approaches to combining existing trust models from different categories: Filter-then-Discount and Discount-then-Filter. We for the first time proved that most of the Discount-then-Filter combinations are robust against all the investigated attacks.

Although our work focused on unfair rating attacks, we plan to combine sellers' cheating behaviors with advisors' unfair ratings, and evaluate their threats to the existing trust models. We are also interested in re-designing new trust models to be completely robust against all the investigated attacks without combining existing ones. Since Sybil-based unfair ratings attacks are more effective than Non-Sybil-based, we also want to design more effective unfair rating attacks with limited buyer account resources. We believe these directions inspired by this work will yield further important insights in the trust management area.

7 Acknowledgement

We wish to acknowledge the funding support for this project from Nanyang Technological University under the Undergraduate Research Experience on CAmpus (URECA) programme.

References

1. C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proceedings of the 2nd ACM Conference on Electroic Commerce*, pages 150–157. ACM, 2000.
2. J. Douceur. The sybil attack. *Peer-to-peer Systems*, pages 251–260, 2002.
3. K. Fullam, T. Klos, G. Muller, J. Sabater, A. Schlosser, Z. Topol, K. Barber, J. Rosenschein, L. Vercouter, 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 4th International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 512–518. ACM, 2005.
4. A. Jøsang and J. Golbeck. Challenges for robust of trust and reputation systems. In *Proceedings of the 5th International Workshop on Security and Trust Management (SMT 2009), Saint Malo, France*, 2009.
5. A. Jøsang and R. Ismail. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, pages 41–55, 2002.
6. R. Kerr and R. Cohen. Smart cheaters do prosper: Defeating trust and reputation systems. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 993–1000, 2009.
7. S. Liu, J. Zhang, C. Miao, Y. Theng, and A. Kot. iclub: an integrated clustering-based approach to improve the robustness of reputation systems. In *The 10th International Conference on Autonomous Agents and Multiagent Systems*, volume 3, pages 1151–1152, 2011.
8. W. Teacy, J. Patel, N. Jennings, and M. Luck. Travos: Trust and reputation in the context of inaccurate information sources. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.
9. A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. In *Proc. 7th Int. Workshop on Trust in Agent Societies*, 2004.
10. J. Zhang and R. 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.