A Context-Aware Framework for Detecting Unfair Ratings in an Unknown Real Environment

Cheng Wan, Jie Zhang* and Athirai A. Irissappane*
School of Computer Science and Engineering, Southeast University, China.{chengwan@njmu.edu.cn}
Department of Healthcare Information System, Nanjing Medical University, China.
*School of Computer Engineering, Nanyang Technological University, Singapore.

***Abstract*—Reputation systems are highly prone to unfair rating attacks. Though many approaches for detecting unfair ratings have been proposed so far, their performance is often affected by the environment where they are applied. For a given unknown real environment, it is difficult to choose the most suitable approach for detecting unfair ratings as the ground truth data necessary to evaluate the accuracy of the detection approaches remains unknown. In this paper, we propose a novel Context-AwaRE (CARE) framework, to choose the most suitable unfair rating detection approach for a given unknown real environment. The framework first identifies simulated environments, closely similar to that of the unknown environment. The detection approaches performing well in the most similar simulated environments are then chosen as the suitable ones for the unknown real environment. Detailed experiments illustrate that the CARE framework can choose the most suitable detection approach to accurately distinguish fair and unfair ratings for any given unknown environment.**

I. Introduction

In electronic marketplaces, a reputation system is designed to measure the reputation of sellers by collecting opinions (*i.e.*, ratings) from buyers who have had experience with the sellers. Reputation systems are particularly useful in large e-marketplace environments in which buyers may often interact with sellers with whom they have no prior experience. In such environments, the buyers can still make informed decisions based on the prior experience of other buyers (advisors[1]). However, advisors may provide unfair ratings to promote some sellers or bad-mouth others.

A lot of detection approaches (trust models) [1] have been proposed to identify the unfair ratings, in order to improve the effectiveness of reputation systems. However, such trust models are highly affected by the environment where they are applied [2]; BRS [3] performs particularly well when buyers do not have much experience in the environment and the majority of ratings are fair; TRAVOS [4] performs well when buyers have sufficient experience but when advisors provide unfair ratings to only some target sellers; Personalized approach [1] fares well in both cases. Also, most of the detection approaches rely on certain tuning parameters which affect their performance significantly. For example, BRS uses $quantile\ (q)$ parameter to filter dishonest advisors. TRAVOS uses N_{bin} parameter to identify previous ratings of an advisor which are similar to the advisor's current rating.

[1]When a buyer evaluates a seller, other buyers are that buyer's advisors who provide opinions about the seller.

For a real environment, it is not easy to obtain ground truth (information needed to distinguish fair and unfair ratings) because 1) it may be prohibitively expensive or time-consuming to hire human subjects to inspect every rating irrespective of whatever interaction is rated by the rating or whoever is involved in the interaction; 2) though ground truth may be available to some people and institutions, they may not be willing to share it. Even if we determine the definite truth for a few such environments, it is impossible to cover the exhaustive list of unknown real environments. Further, the changing behaviour of participants in the environments can hinder the accuracy of the unfair rating detection approaches. Thus, choosing the most suitable unfair rating detection approach for an unknown real environment with no ground truth data becomes a challenging problem.

In this paper, we propose a Context-AwaRE (CARE) framework, to choose the most suitable unfair rating detection approach for a given unknown real environment. Our framework is based on the idea that if an approach performs well in one environment, it should also perform well in another similar environment. The similar environment has similar features (*e.g.*, the ratio of number of buyers versus sellers) as the original environment. We first find out the best approaches with their best parameter values for a set of simulated environments. Given an unknown real environment, we calculate the similarity between each simulated environment and the real environment based on a set of carefully selected features. The approaches performing well in the most similar simulated environments are chosen as the suitable ones for the unknown real environment. Detailed experiments illustrate that our CARE framework can choose the most suitable detection approach for any given unknown real environment.

II. Related Work

A lot of approaches have been proposed to detect unfair ratings in reputation systems. In Whitby *et al.* [3], if the reputation value of a target seller based on the set of honest buyers falls in the rejection area (q quantile or $1-q$ quantile) of the beta distribution of the buyer's ratings to that seller, then the buyer is identified as a dishonest buyer. Teacy *et al.* [4] proposed the TRAVOS model which first determines the accuracy of advisors based on their previous advice and then adjusts the advisors opinions according to their accuracy. The Personalized approach proposed by Zhang and Cohen [1]

models the trustworthiness of an advisor by taking into account both buying agents private experience with the advisor and the public knowledge about the advisor held by the system.

Liu *et al.* [5] proposed an integrated clustering-based approach called iCLUB to filter unfair testimonies using multi-nominal ratings. This approach adopts clustering techniques and uses buyer's local and global information about the seller to filter unfair ratings. Dellarocas [6] used a collaborative filtering technique to divide all ratings into two clusters: one containing lower ratings and one containing higher ratings. The ratings present in the higher ratings cluster are considered as unfair ratings. Weng *et al.* [7] proposed an entropy-based method to measure the quality of the ratings, based on which unfair ratings are filtered. Here, a rater gives high endorsement to other raters who provide similar ratings and low endorsement to raters who provide different ratings. Yang *et al.* [8] used statistical detectors to detect the time intervals in which collaborative unfair ratings are highly likely.

Though the above approaches provide mechanisms to filter unfair ratings, their performance is highly dependent on the environment where they are applied [2], as well as the values set for the parameters in the trust models. In this paper, we focus on the problem of selecting the most suitable unfair rating detection approach and its parameter values for a particular unknown real environment using the CARE framework described in detail in the following section.

III. The CARE Framework

The CARE framework consists of a set of simulated environments and the corresponding detection approaches which are the most suitable for those environments (*i.e.*, *Environment-Approach Pairs EAPs*). The EAPs are previously determined through detailed analysis and experimentation. Given an unknown real environment, the framework first extracts some features of the environment which are considered to be the most influential in detecting unfair ratings. These features are used to calculate the similarity between the given unknown real environment and the simulated environments present in the framework. The most similar simulated environment is identified and the detection approach which performs the best in this environment is determined using the EAPs. This approach is then considered to be the most suitable detection approach for the given unknown real environment. The components of the CARE framework are described in detailed below.

A. Simulated and Unknown Environment

The most important components of the framework are the environments. An environment (e) is defined as a population of all the ratings present in a particular business case.

$$e = \{R_{s,b} \,|s = 1...N_s, b = 1...N_b\} \tag{1}$$

where N_s and N_b are the numbers of sellers and buyers in e respectively. Rating vector $R_{s,b}$ denotes the rating given by a buyer b to a seller s. $R_{s,b}$ is a multi-tuple $(id, sellerId, buyerId, timeSession, val_1, val_2, flag)$, where id, $sellerId$, $buyerId$ denote the current rating id, id of the seller and that of the buyer respectively, $timeSession$ denotes the time when the rating is given and val_1, val_2 denote the value of the rating given. val_1 is a binary value (1 if a seller is trustworthy, 0 otherwise) and val_2 is a real number in the range $[0-1]$. val_1 and val_2 are introduced to accommodate detection approaches for reputation systems with binary and multi-nominal ratings respectively. The ground truth of the rating $R_{s,b}$ is denoted using the $flag$ attribute;$flag = 0$ denotes that the ground truth of the rating is unknown, $flag = 1$ denotes that the rating is fair and $flag = -1$ denotes that the rating is unfair. Any two environments e_1 and e_2 are considered to be independent.

1) Simulated Environments (E): The simulated environments are environments with known ground truth data. Each simulated environment e_i is represented using Equation 1 ($R_{s,b}$: $flag \neq 0$). We simulate a large number of such environments to cover as many scenarios as possible which could closely depict possible real environments. For example, we simulate a very sparse environment where the number of ratings provided by buyers to sellers is small, a very dense environment where each seller is flooded with a large number of ratings, etc. A variety of attack models (*e.g.*, attacks which affect only reputable/disreputable targets, collusion attacks, etc) are also simulated. We denote the set of all simulated environments in the framework as E, where E = (e_1, e_2,....e_n).

2) Unknown Environment (e_u): The unknown environment is an environment for which the ground truth about which ratings are unfair is not known. We represent the unknown environment e_u using Equation 1 ($R_{s,b}$: $flag = 0$).

B. Environment Features

The environments in the framework are represented by a set of well defined features. Features refer to the statistics describing the characteristics of an environment (*e.g.*, the ratio of number of buyers versus sellers, etc). For s features representing an environment in the framework, the *Feature Vector* (F) is given by $(F_1, F_2, \cdots, F_s)$. For a simulated environment e_i, the value of the feature vector F is defined as a vector $f^{e_i} = (f_1^{e_i}, f_2^{e_i}, ...f_s^{e_i})$, where $f_1^{e_i}$ is the value of the feature attribute F_1 in the environment e_i. A feature attribute can have different values for different simulated environments.

We use *Correlation and Regression Analysis* to select only a subset of features ($\widetilde{F}$, $\widetilde{F} \subseteq F$) from the exhaustive list present, to compare the given unknown real environment and the simulated environments in the framework.

The values of the selected s' ($s' <= s$) most influential feature attributes in a simulated environment e_i is given by the vector $\widetilde{f^{e_i}}$, $\widetilde{f^{e_i}} \subseteq f^{e_i}$ and $\widetilde{f^{e_i}} = (\widetilde{f_1^{e_i}}, \widetilde{f_2^{e_i}}, \cdots, \widetilde{f_{s'}^{e_i}})$. For the unknown environment e_u, the values of the s' feature attributes is represented using $\widetilde{f^{e_u}}$, where $\widetilde{f^{e_u}} = (\widetilde{f_1^{e_u}}, \widetilde{f_2^{e_u}}, \cdots, \widetilde{f_{s'}^{e_u}})$.

C. Candidate Approaches

Each *Candidate Approach (CP_j)* is a combination of an unfair rating detection approach (*e.g.*, BRS, TRAVOS, Per-

sonalized, etc) along with its fixed tuning parameters (*e.g.*, quantile q for BRS, N_{bin} parameter for TRAVOS, etc). If A is a detection approach and p is the tuning parameter of A and p can take values v_k $(k = 1, 2, 3...)$, then the candidate approach is given by $CP_j : A(p = v_k)$, $(j = 1, 2, 3, ...; k = 1, 2, 3, ...)$. If p can take values (v_1, v_2, v_3, v_4), we obtain 4 candidate approaches for the approach A; $CP_1 : A\,(p = v_1)$, $CP_2 : A\,(p = v_2)$, $CP_3 : A\,(p = v_3)$ and $CP_4 : A\,(p = v_4)$. When the value v_k is continuous, the range of v_k is divided into equal number of intervals and a value is randomly chosen from each interval to formulate the corresponding CP_j.

D. Best Environment-Approach Pair (EAP)

The candidate approach which is the most suitable for the simulated environment in the framework is identified through detailed experimentation and the EAP set is determined. For every pair of simulated environment and candidate approach (e_i, CP_j), the performance of CP_j in detecting the dishonest advisors in e_i is measured using Matthew's Correlation Coefficient (MCC) [2],

$$MCC(e_i, CP_j) = \frac{(t_p * t_n - f_p * f_n)}{\sqrt{(t_p + f_p)\,(t_p + f_n)\,(t_n + f_p)\,(t_n + f_n)}} \quad (2)$$

where, t_p = true positives, f_p = false positives, t_n = true negatives and f_n = false negatives. A false positive denotes that a dishonest advisor in e_i is incorrectly detected by CP_j as honest. A false negative denotes that a honest advisor is misclassified as dishonest by CP_j. If $MCC(e_i, CP_j)$ is above a threshold (0.8), then CP_j is regarded as the best[2] detection approach for e_i. The EAP is described by the triple, $(e_i, CP_j, MCC(e_i, CP_j))$ and is added to the local EAP set.

E. Most Similar Simulated Environment

Based on the initial assumption that any two environments are independent of each other, the difference in the feature attributes of the environments stands to represent the similarity between the environments themselves. We use the *Distance Correlation Analysis* to calculate the similarity between the given unknown real environment e_u and the simulated environment e_i using the function $D(\widetilde{f^{e_u}}, \widetilde{f^{e_i}})$. The definition of the similarity function is based on the nature of the feature attributes. If the feature attributes are continuous, distance measure is used (*e.g.*, Euclidean Distance, Squared Euclidean Distance, Chebyshev Distance, City Block Distance, etc). Statistics is used otherwise (*e.g.*, λ^2 statistics, and φ^2 statistics). Thereby, the simulated environment e', most similar to the unknown environment e_u is obtained.

F. Most Suitable Approach for the Unknown Real Environment

Using the environment e' which is closely similar to e_u (as described in the above section), we select the candidate approach CP' which is the best approach for the simulated environment e' from the EAP set. This approach is then considered to be the most suitable approach to detect unfair ratings for the unknown real environment e_u.

[2]The number of detection approaches considered the best for a particular simulated environment can be more than 1.

IV. Experiments

To illustrate the performance of the CARE framework, we conduct experiments to show the accuracy of the framework in choosing the most suitable approach to detect unfair ratings for an unknown real environment.

A. Experimental Settings

1) Generation of Simulated Environments (E): We simulate 648 environments using various business cases involving fair ratings and unfair ratings as described below.

Generation of Fair Ratings: A marketplace environment involving 10 sellers with different reputation values $[0.5 - 0.9]$ is considered for creating 9 business cases involving fair ratings. The marketplace operates for 90 days. The number of buyers is chosen depending upon the nature of the simulated environment (*e.g.*, when the behaviour of the buyers is sparse and the total number of ratings is 1000, the number of buyers is 1000). We specifically consider certain simulation parameters to design the business cases as listed below:

- *Total Number of Fair Ratings*: It represents the total number of fair ratings in the marketplace environment and can take values $[20, 100, 1000]$.
- *Behavior of Honest Buyers*: The behaviour of the buyers in the market can be of three kinds: (1) sparse, where each buyer rates a seller at most once; (2) intensive, where the buyers can rate a seller more than one time; (3) mixed, which is a combination of sparse and intensive buyers.

Generation of Unfair Ratings: We choose 5 simulation parameters to design the business cases which involve attacking behaviors. The parameters are listed below:

- *Attack Rate*: It is the ratio of the number of unfair ratings versus the number of fair ratings. $attackRate = 0.2$ denotes that most ratings are fair, 1 denotes equal number of fair and unfair ratings and 2 denotes that most of the ratings are unfair.
- *Attack Type*: It signifies whether the attack generates unfair bad ratings ($attackType = 0$) or unfair good ratings ($attackType = 1$). Here, unfair bad ratings denote the unfair ratings with $val_1 = 0$ and unfair good ratings denote the unfair ratings with $val_1 = 1$.
- *Time Session of the Attack*: It represents the time session when the unfair ratings are given. A $timeSession$ of 7 denotes a concentrated attack. $timeSession = 90$ represents a distributed attack for 3 months.
- *Attack Behaviour*: The behaviour of the attackers in the market can be of three kinds: (1) sparse, where the number of unfair ratings provided by each attacker is at most 1; (2) intensive, where an attacker can provide more than 1 unfair rating; (3) mixed, which is a combination of sparse and intensive.
- *Attack Object*: It represents the seller being attacked. It can take two values, high or low. The value is high when the $attackObject$ is the seller with the highest reputation

in the market and low when $attackObject$ is the seller with the lowest reputation in the market.

After combining these 5 parameters, we generate 72 business cases with different types of attacks using randomly generated unfair ratings. Using the 9 business cases which incorporate fair ratings and the 72 business cases with unfair ratings, we generate 648 different simulated environments.

2) Selection of the Most Influential Features: From a wide range of environmental features, we select 3 most influential features using correlation and regression analysis: (1) variance of rating rate per seller; (2) average number of ratings for each (buyer, seller) pair and (3) ratio of number of buyers versus sellers. Only these features are used to calculate the similarity between the unknown real environment and the simulated environments.

3) Selection of Candidate Approaches (CP_j): We consider 3 unfair rating detection approaches, BRS, TRAVOS, and Personalized approach. We select several tuning parameters for each detection approach and generate 60 candidate approaches. The detailed steps involved in designing the candidate approaches are presented below.

- BRS: The selected tuning parameters are (1) $time\ weight$ (λ) which is used to model the changing behaviour of an agent with time. It can take values $[0.95-1]$; (2) $quantile$ (q) parameter which is used to filter the dishonest buyers. It can take values $[0.2, 0.1, 0.05, 0.02]$.
- TRAVOS: The $number\ of\ bins$ (N_{bin}) is the tuning parameter. It can take values $[5, 8, 10, 20]$.
- Personalized: The tuning parameters are (1) N_{min}, which is the minimum number of rating pairs needed for a buyer to be confident about the private reputation of an advisor. N_{min} can take values $[11, 24, 42, 51, 95, 115]$; (2) $gamma$ (γ), which is the level of confidence the buyer would like to attain. γ can take values $[0.5, 0.6, 0.7, 0.8]$.

4) Generation of Best Environment-Approach Pairs: Based on the set of simulated environments E and the candidate approaches CP_j's, the MCC values of every (e_i, CP_j) pair is calculated and only pairs with MCC value > 0.8 are selected. 6666 such pairs are obtained.

5) Generation of Unknown Environment (e_u): Two categories of unknown environments are generated.

- *Category 1*: Here, the unknown environments have similar simulation parameters as that of the simulated environments in the CARE framework. 100 such unknown environments are generated.
- *Category 2*: The unknown environments have different simulation parameters than that of the simulated environments. We change the range set from which the simulation parameters take their values as follows, (1) range for the reputation values of the sellers is changed to $[0.75, 0.95]$; (2) total number of fair ratings $[40, 150, 800]$; (3) number of sellers 20; (4) ratio of number of buyers versus sellers $[2, 1, 0.5]$; (5) ratio of number of unfair ratings versus fair ratings $[0.5, 1, 3]$ and (6) the time session of the attack is changed to [7, 30]. We generate 100 unknown environments in this category.

B. Experimental Results

Experiments to evaluate the performance of the CARE framework, comparing it with BRS, TRAVOS, and Personalized approach are conducted. We choose the following standard[3] parameter settings for the detection approaches for comparison, BRS: $\lambda = 0.98$ and $q = 0.95$; TRAVOS: $N_{bin} = 5$; Personalized Approach: $N_{min} = 51$ and $\gamma = 0.5$.

1) Unknown Environments with Similar Parameters: Figure 1 shows the results for the 100 unknown environments of *Category 1*. We find that the mean MCC value obtained by using the detection approach proposed by the CARE framework is above 0.5 for most of the unknown environments and falls in the range $(0.9-1)$ for 58 of them. For BRS, the mean MCC value is < 0 for 21 unknown environments and is in the range $(0.9-1)$ for 12 environments. For TRAVOS and Personalized approach, the mean MCC values for most of the unknown environments fall between $(0.2-0.5)$. Thus, we find that the detection approach proposed by the CARE framework has high mean MCC values for most of the unknown environments when compared to the other approaches.

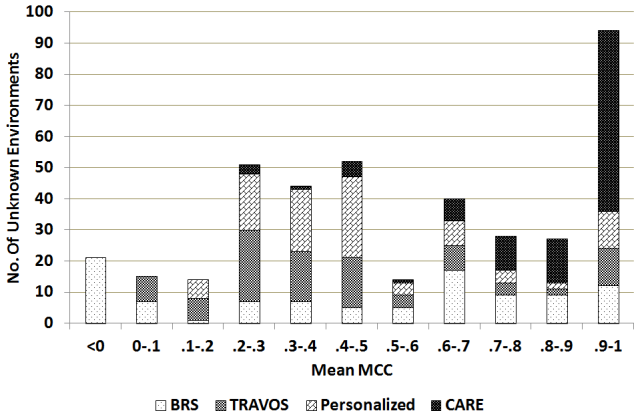

Fig. 1. Mean MCC for Unknown Environments with Similar Parameters

2) Unknown Environments with Different Parameters: Figure 2 shows the results for the unknown environments with different parameters. We find that the mean MCC obtained using the CARE framework is better than the other approaches and falls in the range $(0.9-1)$ for 34 unknown environments.

3) Unknown Environments with Special Attacks: We simulate 4 kinds of scenarios. Each scenario is evaluated based on 3 cases: (1) attack ratio < 0.5; (2) attack ratio is between $[0.5-1]$ and (3) attack ratio > 1.

- *Sparse buyers with unfair positive ratings attack*: Here, the sellers are new to the market. The number of ratings

[3]The settings are the standard settings used by the authors of the corresponding trust models.

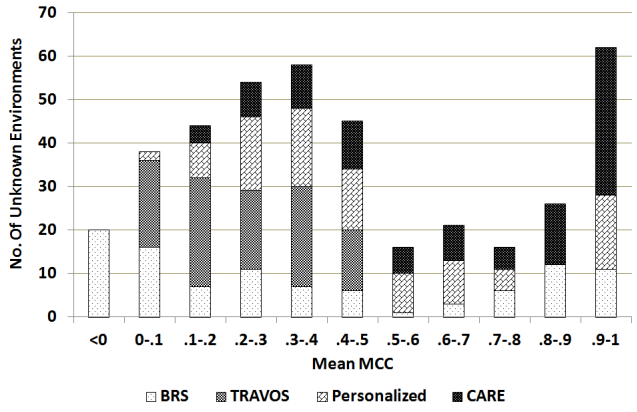

Fig. 2. Mean MCC for Unknown Environments with Different Parameters

per seller is < 10. We simulate *unfair positive ratings attack*, where some sellers improve their reputation and gain priority over the others. The results are shown in Figure 3 (a).

- *Sparse buyers with unfair negative ratings attack*: Here, we simulate a scenario which is nearly opposite to the one mentioned in the previous case. *Unfair negative ratings attacks*, which happen when some sellers want to purposely reduce other sellers' reputation are simulated. The results are shown in Figure 3 (b).

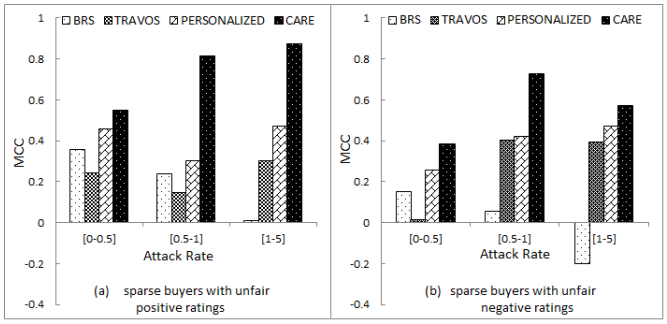

Fig. 3. MCC values for Sparse Buyer Market

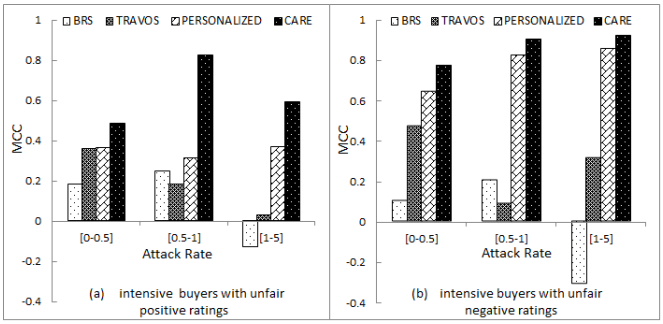

Fig. 4. MCC values for Intensive Buyer Market

- *Intensive buyers with unfair positive ratings attack*: We simulate a mature market where sellers have established good reputation scores. The rating time and number of ratings per seller is high (> 30). *Unfair positive ratings attack* is simulated. The results are shown in Figure 4 (a).
- *Intensive buyers with unfair negative ratings attack*: Here, we simulate the opposite scenario as that of the previous case. *Unfair negative ratings attack* is simulated. The results are shown in Figure 4 (b).

We see that in all the cases the detection approach proposed by the framework has high MCC values when compared to BRS, TRAVOS and Personalized approach. This shows the ability of the framework to detect unfair ratings in any given environment, including special attacking scenarios.

V. Conclusion and Future Work

Existing unfair rating detection approaches are highly affected by the environment in which they are applied. It therefore becomes difficult to choose the most suitable unfair rating detection approach for a given unknown real environment. In this paper, we propose a Context-AwaRE (CARE) framework which finds a set of simulated environments, closely similar to the given unknown real environment and uses the best Environment-Approach Pairs to determine the most suitable detection approach for the unknown real environment. Experimental results show that the CARE framework can handle various kinds of unknown environments with a variety of unfair rating behaviour. In the future, we plan to improve the learning ability of the CARE framework by using the *Reinforcement Learning Model* and compare the performance of the framework with other existing approaches (iCLUB [5], WMA [9], Entropy-based approach [7], etc).

Acknowledgment

This work is partially supported by the NTU Start-up (M4080096.020) and MOE AcRF Tier 1 (M4010265.020) Grant awarded to Dr. Jie Zhang.

References

[1] J. Zhang and R. Cohen, "Evaluating the trustworthiness of advice about seller agents in e-marketplaces: A personalized approach," *Electronic Commerce Research and Applications*, vol. 7, no. 3, pp. 330–340, 2008.

[2] J. Zhang, "Extensive experimental validation of a personalized approach for coping with unfair ratings in reputation systems," *Journal of theoretical and applied electronic commerce research*, vol. 6, no. 3, pp. 43–64, 2011.

[3] A. Whitby, A. Jøsang, and J. Indulska, "Filtering out unfair ratings in bayesian reputation systems," in *Proceedings of the 7th International Workshop on Trust in Agent Societies*, 2004.

[4] 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*, vol. 12, no. 2, pp. 183–198, 2006.

[5] 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 *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), Taipei, Taiwan*, 2011, pp. 1151–1152.

[6] C. Dellarocas, "Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior," in *Proceedings of the 2nd ACM conference on Electronic commerce*, 2000, pp. 150–157.

[7] W. Jianshu, M. Chunyan, and G. Angela, "An entropy-based approach to protecting rating systems from unfair testimonies," *IEICE TRANSACTIONS on Information and Systems*, vol. 89, no. 9, pp. 2502–2511, 2006.

[8] Y. Yang, Y. Sun, S. Kay, and Q. Yang, "Securing rating aggregation systems using statistical detectors and trust," *IEEE Transactions on Information Forensics and Security*, vol. 4, no. 4, pp. 883–898, 2009.

[9] B. Yu and M. Singh, "Detecting deception in reputation management," in *Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. ACM, 2003, pp. 73–80.