A Fuzzy Logic Based Reputation Model Against Unfair Ratings

Siyuan Liu
School of Electrical and
Electronic Engineering
NTU, Singapore
lius0036@ntu.edu.sg

Han Yu
School of Computer
Engineering
NTU, Singapore
yuha0008@ntu.edu.sg

Chunyan Miao
School of Computer
Engineering
NTU, Singapore
ascymiao@ntu.edu.sg

Alex C. Kot
School of Electrical and
Electronic Engineering
NTU, Singapore
eackot@ntu.edu.sg

ABSTRACT

Reputation systems have become more and more important in facilitating transactions in online systems. However, the accuracy of reputation systems has always been a concern for the users due to the existence of unfair ratings. Though many approaches have been proposed to mitigate the adverse effect of unfair ratings, most of them use the credibility of the rating provider alone to decide whether the rating is unfair without considering other aspects of the rating itself. Models that do consider multiple aspects often combine them through arbitrarily set weights. Therefore, they cannot work well when the credibility is not evaluated accurately or when the weights are not set properly. To resolve this problem, in this paper, we propose a reputation model which considers and combines the temporal, similarity and quantity aspects of the user ratings based on fuzzy logic to improve the accuracy of reputation evaluation. Experimental results based on a set of real user data from a cyber competition show that the proposed model is more robust against unfair ratings than the existing approaches, especially under Sybil attack conditions.

Categories and Subject Descriptors

I.2.11 [**ARTIFICIAL INTELLIGENCE**]: Distributed Artificial Intelligence – Intelligent agents, Multiagent systems

General Terms

Design, Measurement

Keywords

Reputation, Unfair Rating, Fuzzy Logic

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013), Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.*

1. INTRODUCTION

In today's world where the Internet has permeated into every facet of our daily life, a massive amount of transactions are taking place over various online communities around the clock. For example, in 2011, 53% of the U.S. population bought something online and a total of $200 billion was spent on online shopping. The online transactions involve things of either intangible value (e.g., opinions about a restaurant or movie, etc.) or tangible value (e.g., products, services, payments, etc.). Such transactions often involve stakeholders with potentially conflicting interest and even nefarious motivations. Currently, multi-agent research has come a long way in terms of understanding and modeling the cooperation among agents representing the interest of different stakeholders. It is a consensus in the multi-agent system (MAS) research community that trust and reputation management (TRM) is a useful way for sustaining the healthy operation of such online transaction systems [6]. In particular, users can share opinions in an online transaction system in the form of numerical ratings or text reviews, regarding various items, e.g., movies, products, services and even other users. TRM then collects, analyzes and aggregates the shared ratings to derive reputation scores which are available to the general users. In this paper, to distinguish different types of users, we will refer to these who are evaluating the reputation of items as *trusters*, the specific items whose reputation are under evaluation as *trustees*, those who provide ratings regarding the items as *witnesses*, and the ratings provided regarding the items as *testimonies*[1].

The testimonies may appear to be unfair to a truster due to the following factors:

1. *Deliberate alterations*: a witness may engage in collusion with trustees to disseminate distorted testimonies to unfairly inflate the reputation of these trustees or tarnish the reputation of their competitors. For example, a company named "VideoViralView.com" [2] can provide 100 real user ratings to a piece of music on iTunes in exchange for payment.

2. *Changing situations*: the behavior of a trustee may change over time, thus rendering testimonies based on

[1] We use the terms "rating" and "testimony" interchangeably.

past transactions from some witnesses obsolete. For example, a rating provided for a restaurant two years ago may not apply today as the restaurant can change its chef and menu, or even cease to exist.

3. *Subjective differences*: users may have different personal dispositions when rating a trustee. For example, some users may prefer horror movies, while others may enjoy the romantic genre.

It has been shown that unfair testimonies may negatively impact the accuracy of the reputation evaluations and the resulting transaction decisions [23]. Thus, they need to be discounted before being used to evaluate the reputation of a trustee. Many testimony discounting approaches have been proposed. They are designed based on different combinations of intuitions from observations on how human beings handle unfair testimonies together with a wide range of theories including probabilistic theories [5] [16], information theories [18] [21], statistical theories [19], and learning theories [15] [22]. Nevertheless, these approaches tend to rely on the the credibility calculated based on a single aspect of a testimony to assess whether it is unfair, or a simple combination of multiple aspects through arbitrarily set weights. These approaches are ineffective when the credibility cannot be calculated accurately or the weights are not set properly. They also depend on threshold values set by human experts when making the decision, which introduces subjectivity into their models. In addition, these thresholds are normally a single numeric value which makes the discounting decisions for testimonies around the thresholds abrupt.

In this paper, we propose a novel fuzzy logic based reputation model to address these disadvantages. To cope with the problem of unfair testimonies, when evaluating a trustee's reputation, the proposed model provides a truster with the facilities to handle the temporal, similarity and quantity aspects of the testimonies to adjust the truster's propensity to trust. To fuse the three aspects together, a fuzzy logic system is employed, whose output is used as the discounting weight to aggregate testimonies. By adopting fuzzy logic, the proposed model has the advantage of leveraging on human expert knowledge while not depending solely on crisp preset threshold values to calculate the discounting weights. The proposed model is evaluated against a state-of-the-art approach using real user data collected from a cyber competition [8] [14] with ratings which include a significant proportion from collusive attackers engaged in various attacks. Our experimental results show that compared with other approaches, the proposed model can more effectively resist the attacks and significantly improve the accuracy of reputation evaluation.

The remainder of this paper is structured as follows. A review of related work is given in Section 2. Section 3 presents the proposed fuzzy logic based reputation model. The performance of the proposed model is studied and experimental results are presented in Section 4. Section 5 discusses some potential issues requiring more considerations. Finally, Section 6 concludes the paper.

2. RELATED WORK

Many reputation models for mitigating the adverse effects of unfair testimonies have been proposed in MAS research. According to the technical mechanisms they adopt, they can be classified into the following categories.

2.1 Probabilistic Approaches

Jøsang and Ismail proposed the Beta Reputation System (BRS) [5]. In BRS, ratings for a trustee are expressed as either positive or negative, which can be considered as two events in the beta probability distribution [4]. The provider's reputation is calculated as the expected value of the positive rating happening in the future by substituting the numbers of the positive and negative ratings into the beta probability density function. The older ratings will carry lower weights by introducing a forgetting factor. This model considers the temporal aspect of the testimonies, and is effective in the cases of the unfair testimonies caused by changing situation, but not deliberate alterations or subjective difference.

In [16], the authors proposed TRAVOS, which is a probabilistic reputation model for agent-based virtual organizations. It is designed following the intuition that only if a truster is not confident about the trust evidence regarding a trustee derived from his past direct transactions should third-party testimonies be required. In this case, the testimonies contain not only the overall assessment of the trustworthiness of a trustee, but also the self-reported frequency of success and failure between the witness and the trustee. A probability value of a given witness providing fair ratings is derived based on his past opinions. This probability is used to determine the weight of the witness's opinion in the reputation evaluation. The model considers the similarity aspect of the testimonies, and is more effective in the cases of unfair testimonies caused by deliberate alterations and subjective differences than those caused by changing situations due to the preference of direct trust evidence to third-party testimonies.

2.2 Information Theory Based Approaches

Yu and Singh [21] proposed a distributed reputation model based on the Dempster-Shafer theory of evidence [13]. The model defines three types of unfair testimonies: positively exaggerated, negatively exaggerated and complementary. If testimonies deviate significantly from a truster's actual personal experience, the corresponding witness's credibility is reduced. When a truster has accumulated enough direct transaction experience with the trustee, it will reduce the elicitation of testimonies from the witnesses. This model considers the similarity aspect of the testimonies and is effective to unfair testimonies caused by deliberate alterations and subjective differences.

In [18], the authors enhanced BRS through filtering the unfair testimonies by assuming the direct experience of a truster as the most reliable source of reputation evidence. The authors proposed an entropy-based approach to measure how much a testimony deviates from the current belief of the truster. If the difference is within a predefined threshold, the new testimony is aggregated into the current belief. This model depends on a truster having sufficient direct transaction experience with a trustee in a given context in order to be effective, which somehow reduces the necessity of third-party testimonies. This improved BRS model considers the temporal and similarity aspects of the testimonies, and can handle unfair testimonies caused by deliberate alterations, subjective differences and changing situations with varying degrees of effectiveness.

2.3 Statistical Approaches

Weng et al. [19] proposed a credibility model based on statistical approaches. It records the testimonies of each witness for each trustee as a witness profile stored locally by every truster. The credibility of each witness, in the context of providing testimonies for each trustee, is evaluated based on its history of success and recorded separately. The model considers the similarity aspect of testimonies and is effective to unfair testimonies caused by deliberate alterations and subjective differences.

2.4 Learning Based Approaches

In [15], the authors proposed a Bayesian trust model (HABIT) to assess trust based on direct experience and third-party information. HABIT learns agent behaviors and infers behaviors back to ratings based on observed similarities between groups of agents. HABIT considers the similarity aspect of testimonies and is effective in handling unfair testimonies caused by deliberate alterations and subjective differences.

In [22], the authors proposed a reinforcement learning based reputation model which adjusts the relative importance given to the testimonies from each witness based on the actual gain or loss derived from the actual transactions following their advice. The model provides a method for evaluating the impact of each witness's testimonies on the truster's wellbeing and rewarding/penalizing them accordingly. The set of selected witnesses as well as the importance of their opinions relative to each other are updated by the truster through the act of trusting relying on their testimonies, observing the transaction outcomes, and assessing the changes in its wellbeing. The weights given to the truster's direct trust evidence and third-party testimonies with regard to each trustee also dynamically change according to the gain/loss in utility attributable to these two sources of trust evidence. The model considers the similarity aspect of testimonies and can handle unfair testimonies caused by deliberate alterations and subjective differences.

Our work belongs to the category of information theory based approaches as we consider fuzzy logic system as a kind of information fusion theory. It can handle unfair testimonies caused by deliberate alterations, subjective differences and changing situations by considering multiple aspects of the testimonies while not requiring an arbitrary threshold setting.

3. THE PROPOSED FUZZY LOGIC BASED REPUTATION MODEL

In this section, we introduce the proposed fuzzy logic based reputation model. Before we come into the details of the proposed model, let's first imagine a similar face-to-face scenario in reality. Suppose a person P is going to a restaurant for dinner. Before P selects a restaurant, he can ask for advice from one of his friends F on how much F will recommend a candidate restaurant. Suppose F highly recommends P to this restaurant, then a new question arises – how much P will rely on F's advice. There are some factors impacting P's decision.

First, if F went to the restaurant just one day before P asking for his advice, then the advice should carry more weight than in the case that F went to the restaurant one year before because the advice provided more recently can better reflect the current quality of the restaurant. Second, from P's point of view, the weight of F's advice is also related to the similarity between P and F's taste. If P prefers Japanese food and F prefers Korean food, and if the restaurant in question primarily serves Japanese food, then F's advice on the restaurant will carry less weight as it can be unfair due to F's subjectiveness. Third, F's confidence will also impact the weight of F advice. If F has provided a lot of advice regarding different restaurants in the past, then P can consider that F's advice for this restaurant may be more convincing.

Mapping this to a reputation evaluation scenario, a truster can evaluate the weights of the testimonies provided by the witnesses using the same way. The time the ratings are provided, the similarity between the witnesses and the truster for the commonly rated trustees, and the confidence of the witnesses in providing ratings, are the three factors we consider to calculate the weights the ratings carry. In the following parts, we will first introduce the proposed fuzzy logic based reputation model, including how to measure the three factors and how to use a fuzzy logic system to combine the three factors together. Then, we use an example to show how the proposed reputation model, especially the fuzzy logic system, works.

3.1 Factor Measurement

Suppose a truster U of an online transaction system is evaluating the reputation of a trustee I which has received N ratings expressed as a rating set $\{r_1, ..., r_i, ..., r_N\}$, where r_i is a numerical integer value. If all the ratings are fair, then I's reputation R can be calculated as:

$$R = \frac{\sum_{i=1}^{N} r_i}{N}. \tag{1}$$

However, as we have mentioned in Section 1, the ratings can be unfair due to various factors. Therefore, before we aggregate them together, we need to calculate the weights put on them. For a particular rating r_i $(1 \leq i \leq N)$ provided by a witness $\hat{U}$, its weight is calculated by taking the following three factors into consideration:

- x_1: the gap between the time r_i is provided, and the time r_i's weight is evaluated
- x_2: the similarity between $\hat{U}$ and U's subjectiveness in providing ratings.
- x_3: the confidence of $\hat{U}$ in providing ratings.

The first factor x_1 is measured as:

$$x_1 = \lambda^{T_c - T_i}, \tag{2}$$

where λ is the forgetting factor [5] in the range of [0,1], T_c is current time window, and T_i is the time window in which r_i was provided. This equation implies that the impact of the older ratings will be smaller as the old ratings will be forgotten gradually with time passing. λ controls the rate of the old ratings being forgotten. If $\lambda = 1$, the old ratings will be never forgotten. If $\lambda = 0$, only the ratings in the current time window will have impacts on the trustee reputation evaluation.

The second factor x_2 is measured through comparing U and $\hat{U}$'s ratings for their commonly rated trustees. Suppose U and $\hat{U}$ have L commonly rated trustees, denoted as $\{I_1, ...I_j, ..., I_L\}$. For a particular trustee I_j $(1 \leq j \leq L)$, the ratings provided by U and $\hat{U}$ are r_{I_j} and $\hat{r}_{I_j}$, respectively. Then, x_2 is measured as the cosine similarity[2] between U and $\hat{U}$'s ratings for the L trustees, as follows:

$$x_2 = \frac{\Sigma_{j=1}^{L} r_{I_j} \times \hat{r}_{I_j}}{\sqrt{\Sigma_{j=1}^{L} r_{I_j}^2} \times \sqrt{\Sigma_{j=1}^{L} \hat{r}_{I_j}^2}}. \quad (3)$$

The third factor x_3 represents the confidence of $\hat{U}$ providing ratings. Currently we simplify the measurement of x_3 by associating it with the number of the trustees $\hat{U}$ has rated. The greater the number of the trustees $\hat{U}$ has rated, the larger his confidence is. Suppose $\hat{U}$ provided ratings for $N_{\hat{U}}$ trustees, then x_3 is measured as:

$$x_3 = \min(\frac{N_{\hat{U}} - N_{min}}{N_{max} - N_{min}}, 1), \quad (4)$$

where N_{min} and N_{max} are constants to normalize x_3 into the range of [0,1].

After we get x_1, x_2 and x_3 values for a particular rating r_i, we will pass them to the fuzzy logic system to calculate the weight w_i that r_i carries. Then I's reputation R is evaluated as:

$$R = \frac{\Sigma_{i=1}^{N} w_i \times r_i}{\Sigma_{i=1}^{N} w_i}. \quad (5)$$

3.2 The Fuzzy Logic System

Generally speaking, a fuzzy logic system is a rule-based system [17]. It translates the human expert's knowledge to some rules following which the system operates. For example, the sentence below describes a rule in the reputation evaluation scenario:

IF the rating is provided recently, and the similarity between the witness and the truster is high, and the witness is confident in providing ratings, THEN the rating's weight is very high.

A typical fuzzy logic system works as shown in Figure 1. A real-valued point x is first transformed into a fuzzy set in X through a fuzzifier. Then according to the predefined rules, the fuzzy inference engine transforms the fuzzy set in X into another fuzzy set in Y. Finally, an output y can be derived through the defuzzifier.

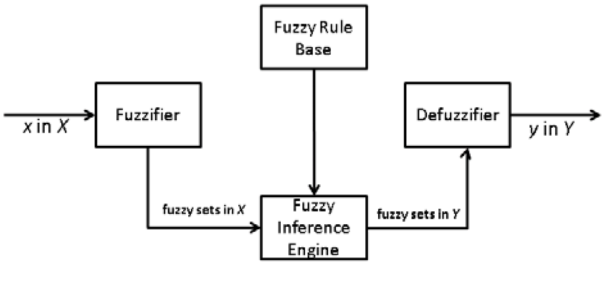

Figure 1: A typical fuzzy logic system [17]

In our proposed reputation model, the input x is a 3-tuple – $< x_1, x_2, x_3 >$. Each x_i $(i = 1, 2, 3)$ is associated with the fuzzy sets – Low (L), Medium (M) and High (H). The output variable y is associated with the fuzzy sets – Very Low (VL), Low (L), Medium (M), High (H) and Very High (VH). For a fuzzy logic system to work, a Membership Function (MF) [11] is first defined for each fuzzy set. MF is used to characterize a fuzzy set to associate a variable with it. In the proposed reputation model, Gaussian MFs are used to express the input fuzzy sets as follows:

$$\mu_{X_i^d}(x_i) = e^{-(\frac{x_i - \overline{x}_i^d}{\sigma_i^d})}, \quad (6)$$

where X_i^d ($i = 1, 2, 3$ and d=L, M, H) represents the fuzzy set that can be interpreted as x_i is a a member of the fuzzy set d (e.g., X_i^d is the fuzzy set of *low similarity* when $i = 2$ and d = L). $\overline{x}_i^d$ and σ_i^d are the constant mean and standard deviation values for the corresponding Gaussian MF, respectively.

The Gaussian MF defined for the output fuzzy set is as follows:

$$\mu_{Y^d}(y) = e^{-(\frac{y - \overline{y}^d}{\sigma_y^d})}, \quad (7)$$

where Y^d (d=VL, L, M, H, VH) represents the fuzzy set interpreted as y is a member of the fuzzy set d. $\overline{y}_d$ and σ_y^d are the constant mean and standard deviation values for the Gaussian MF, respectively.

The Gaussian fuzzifier is equipped to map a real-valued point $x^* = [x_1^*, x_2^*, x_3^*]$ to a fuzzy set X' as the Gaussian fuzzifier has the advantage of suppressing noise. The mapped fuzzy set X' has the following MF:

$$\mu_{X'}(x) = e^{-(\frac{x_1 - x_1^*}{a_1})^2} \times e^{-(\frac{x_2 - x_2^*}{a_2})^2} \times e^{-(\frac{x_3 - x_3^*}{a_3})^2}, \quad (8)$$

where a_1, a_2, and a_3 are constants. To suppress noise, they need to satisfy the following condition:

$$a_i >> \sigma_i^d, \quad (9)$$

for all d =L, M, H, where $i = 1, 2, 3$[3].

After using the fuzzifier to map the point x^* to the fuzzy set X', the output fuzzy set Y' can be inferred according to X' and the predefined rules. In the proposed reputation model, we define the following rules as shown in Table 1.

Table 1: Rules

Rule	x_1	x_2	x_3	y	Rule	x_1	x_2	x_3	y
1	L	L	L	VL	2	L	L	M	L
3	L	L	H	L	4	L	M	L	L
5	L	M	M	L	6	L	M	H	M
7	L	H	L	L	8	L	H	M	M
9	L	H	H	H	10	M	L	L	VL
11	M	L	M	L	12	M	L	H	M
13	M	M	L	L	14	M	M	M	M
15	M	M	H	H	16	M	H	L	M
17	M	H	M	H	18	M	H	H	H
19	H	L	L	VL	20	H	L	M	L
21	H	L	H	M	22	H	M	L	L
23	H	M	M	H	24	H	M	H	M
25	H	H	L	M	26	H	H	M	M
27	H	H	H	VH					

[2] The advantage of using cosine similarity is that it can achieve a similarity value in the range of [0,1] regardless of the rating scales the reputation system adopts.

[3] The original format of the Gaussian fuzzifier is $\mu_{X'}(x) = e^{-(\frac{x_1 - x_1^*}{a_1})^2} \star e^{-(\frac{x_2 - x_3^*}{a_2})^2} \star e^{-(\frac{x_3 - x_3^*}{a_3})^2}$, where $\star$ denotes t-norm [17]. And we use the commonly used algebraic product for t-norm. Then Eq.(8) is achieved.

Each combination of x_1, x_2, x_3, y in Table 1 can be interpreted as a rule. For example, the last combination (i.e., rule 27) can be interpreted as the following rule:

IF x_1 *is* H, *and* x_2 *is* H, *and* x_3 *is* H, *THEN* y *is* VH.

It is actually the same as the example rule previously mentioned. Based on the input MF, fuzzifier output and defined rules, the output fuzzy set is obtained through the fuzzy inference engine as:

$$\mu_{Y'}(y) = \max_{l=1}^{M}[\prod_{i=1}^{3} e^{(-\frac{x_{iP}^l - \overline{x}_i^l}{\sigma_i^l})^2} e^{(-\frac{x_{iP}^l - x_i^*}{a_i})^2} \mu_{Y^l}(y)], \quad (10)$$

where

$$x_{iP}^l = \frac{a_i^2 \overline{x}_i^l + (\sigma_i^l)^2 x_i^*}{a_i^2 + (\sigma_i^l)^2}, \quad (11)$$

for $i = 1, 2, 3$. M is the number of rules. l denotes the index of the rules. $\overline{x}_i^l$ and σ_i^l denote the mean and standard deviation values of the Gaussian MF corresponding to the fuzzy set of x_i in the lth rule, respectively. Y^l denotes the fuzzy set of y in the lth rule. For example, if $l = 1$ (i.e., rule 1 in Table 1), then $\overline{x}_2^l$ and σ_2^l are the mean and standard deviation values of the Gaussian MF corresponding to the fuzzy set of low similarity, and Y^l represents the fuzzy set VL.

After we get the output fuzzy set through the fuzzy inference engine, a defuzzifier is designed to map the output fuzzy set Y' to a crisp value y^*. We currently use the widely adopted center average defuzzifier to map Y' to y^* as follows:

$$y^* = \frac{\Sigma_{l=1}^{M} \omega_l \overline{y}^l}{\Sigma_{l=1}^{M} \omega_l}, \quad (12)$$

where $\overline{y}^l$ denotes the center of the output fuzzy set corresponding to the lth rule, and ω_l be its height. Then the output y^* is considered as the weight of the rating under evaluation.

3.3 Example

In this part, we use an example to show how the fuzzy logic system works. The example data we use in this part are from our experiments. Suppose there are 300 trusters and 300 trustees. For one particular trustee, it received 134 ratings which were provided in different time windows and may include unfair ratings. Suppose one truster is evaluating the reputation of this trustee. The truster first calculates the weight for each rating. Suppose one of these ratings is 5 provided in the time window which is 34 time windows older than the current time window. Then for this rating, x_1 is calculated as λ^{34} according to Eq.(2). When $\lambda = 0.95$, $x_1 = 0.1748$. Suppose the truster and the witness who provided the rating have 7 commonly rated trustees, the ratings for the 7 trustees from the truster are $\{5, 4, 4, 4, 4, 4, 5\}$, and the ratings from the witness are $\{4, 4, 4, 4, 4, 4, 5\}$. The similarity x_2 is calculated as 0.9967 according to Eq.(3). Suppose the witness has rated 64 items in total. Then the witness's confidence x_3 is calculated as 1 according to Eq.(4) when $n_{min} = 1$ and $n_{max} = 20$. Now, the input real-valued point x^* is equal to $[0.1748, 0.9967, 1]$. The parameters for the MFs are set using the values shown in Table 2.

Table 2: Parameter settings

d	VL	L	M	H	VH
$\overline{x}_i^d$ $(i = 1, 2, 3)$	-	0	0.5	1	-
σ_i^d $(i = 1, 2, 3)$	-	0.1	0.1	0.1	-
$\overline{y}^d$	0	0.25	0.5	0.75	1
σ_y^d	0.1	0.1	0.1	0.1	0.1

$a_i = 2\max\{\sigma_i^L, \sigma_i^M, \sigma_i^H\}$ for $i = 1, 2, 3$. After going through fuzzifier and fuzzy inference engine (Eq.(10)), we have:

$$\begin{aligned} \mu_{Y'}(y) = \max[& 0 \times \mu_{VL}(y), 0 \times \mu_L(y), 0 \times \mu_L(y), 0 \times \mu_L(y), \\ & 0 \times \mu_L(y), 0.0039 \times \mu_M(y), 0 \times \mu_L(y), \\ & 0.0037 \times \mu_M(y), 0.5426 \times \mu_H(y), 0 \times \mu_{VL}(y), \\ & 0 \times \mu_L(y), 0 \times \mu_M(y), 0 \times \mu_L(y), 0 \times \mu_M(y), \\ & 0.0009 \times \mu_H(y), 0 \times \mu_M(y), 0.0008 \times \mu_H(y), \\ & 0.1206 \times \mu_H(y), 0 \times \mu_{VL}(y), 0 \times \mu_L(y), \\ & 0 \times \mu_M(y), 0 \times \mu_L(y), 0 \times \mu_M(y), 0 \times \mu_M(y), \\ & 0 \times \mu_M(y), 0 \times \mu_M(y), 0 \times \mu_{VH}(y)]. \end{aligned} \quad (13)$$

By using the defuzzifier (Eq.(12)), the weight of the rating is 0.7472.

4. EXPERIMENTAL STUDIES

In this section, we present the experiments conducted to examine the effectiveness of the proposed model in reputation evaluation with the existence of unfair ratings. We used a set of real user ratings collected from a cyber competition [8] [14] to conduct the experiments. We also compared the effectiveness of the proposed model with a witness credibility model [19] from the aspect of reputation evaluation accuracy. In the following sections, we will first introduce the experimental setup, then present the results.

4.1 Experimental Setup

In current research on reputation systems coping with the problem of unfair ratings, it has always been difficult to obtain real-world data to conduct experiments due to the following two reasons. First, it is difficult to get the data for an online transaction system due to the privacy considerations by the business owners. Second, even if we can get the data from an online transaction system (e.g., eBay), there is usually no ground truth based on which ratings can be labeled as unfair. Therefore, most of the current research work uses simulations to conduct experiments. However, simulations cannot realistically reflect the way unfair ratings happen in the real world.

To resolve this dilemma, we obtained a set of competition data [8] [14] to evaluate the effectiveness of the proposed model in reputation evaluation with the existence of unfair ratings. The set of data includes two parts – normal data and attack data. The normal data include real online ratings (numerical integer values from 1 to 5) collected from a famous e-commerce website [1]. The ratings are for 300 products (i.e., trustees), denoted as $I_1, ..., I_i, ..., I_{300}$, provided by 300 users (i.e., trusters and witnesses), denoted as $U_1, ..., U_i, ..., U_{300}$, during 150 days. These 300 users are referred to as normal users. The normal data are considered as the ground truth in our experimental studies. Then a competition was launched to encourage players to provide unfair ratings to attack the target product (i.e., downgrade the reputation of the first product in the competition). Each player can control 20 user IDs to attack the target product (i.e., Sybil attack). Each ID can only rate one product at most once. The ratings from the manipulated IDs are submitted as an attack profile. The attack data consist of 13,084 attack profiles in total. The effectiveness of an attack profile is measured by the difference between the reputation score of the target product calculated using the normal data and that calculated using the normal data together with the attack profile data. The larger the difference, the greater the impact of the attack profile produces. The player who produced the largest difference wins the prize.

In our experiments, we compared the proposed model with Weng et al.'s credibility model [19] in terms of the accuracy of reputation evaluation regarding the target product. The effectiveness of the proposed approach and Weng's approach in reputation evaluation is measured through the reputation difference metrics as follows:

$$D_{max} = \max_{k=1}^{K}\{|R_{U_k} - \hat{R}_{U_k}|\}, \tag{14}$$

$$D_{mean} = \frac{\sum_{k=1}^{K} |R_{U_k} - \hat{R}_{U_k}|}{K}, \tag{15}$$

where K is the number of normal users who have no ratings for the target product (i.e., I_1). If a normal user has rated I_1, then the reputation score of I_1 is his own rating from his point of view. If a normal user U_k has no rating for I_1, then R_{U_k} is the reputation score for I_1 evaluated by U_k using only the ratings from the normal data. $\hat{R}_{U_k}$ is the reputation score evaluated by U_k using the ratings from the normal data and the attack data. For a particular user U_k, the evaluated reputation for I_1 may be different from the reputation evaluated by other users as both the proposed approach and Weng's approach calculate the weights of the ratings specifically for different users. Therefore, we examined the effectiveness of an approach in worse case by exploring the largest reputation difference (i.e., D_{max}) from a particular normal user's point of view. We also examined the average case effectiveness by exploring the average reputation difference (i.e., D_{mean}) to investigate how an approach can evaluate reputation accurately in general.

4.2 Single Attack Profile

Table 3 shows the results when I_1 is attacked by a single attack profile, where "Attack", "Weng" and "Proposed" mean the scenarios when no mechanisms adopted to counter the unfair ratings, using Weng's approach and using the proposed approach, respectively. The value in each cell is the number of attack profiles which have produced the D_{max} or D_{mean} values in the range corresponding to the cell row. For example, the third row for the D_{max} table represents that there are 56, 0, and 10203 single attack profiles making D_{max} value fall into the range of $[0, 0.1)$ when no mechanisms adopted to counter the unfair ratings, using Weng's approach and using the proposed approach, respectively. The third row for the D_{mean} table represents that there are 56, 0, and 13065 single attack profiles making D_{mean} value fall into the range of $[0, 0.1)$ when no mechanisms adopted to counter the unfair ratings, using Weng's approach and using the proposed approach, respectively.

Table 3: Results for a single attack profile

	D_{max}		
	Attack	Weng	Proposed
$[0, 0.1)$	56	0	10203
$[0.1, 0.2)$	1040	0	2639
$[0.2, 0.3)$	1930	0	217
$[0.3, 0.4)$	2773	0	25
$[0.4, 0.5)$	7168	0	0
$[0.5, 0.6)$	53	0	0
$[0.6, 0.7)$	45	0	0
$[0.7, 0.8)$	19	0	0
$[0.8, 1.0)$	0	0	0
$[1.0, 1.1)$	0	13019	0
$[1.1, 1.2)$	0	30	0
$[1.2, 1.3)$	0	23	0
$[1.3, 1.4)$	0	3	0
$[1.4, 2.9)$	0	0	0
$[2.9, 3.0]$	0	9	0

	D_{mean}		
	Attack	Weng	Proposed
$[0, 0.1)$	56	0	13065
$[0.1, 0.2)$	1040	13084	17
$[0.2, 0.3)$	1930	0	2
$[0.3, 0.4)$	2773	0	0
$[0.4, 0.5)$	7168	0	0
$[0.5, 0.6)$	53	0	0
$[0.6, 0.7)$	45	0	0
$[0.7, 0.8)$	19	0	0
$[0.8, 3.0]$	0	0	0

We can see that D_{max} produced by using the proposed approach is smaller than 0.4 for all single attack profiles, and that produced by using Weng's approach is about 1 for most single attack profiles. D_{max} is even greater than 3 for some profiles when using Weng's approach. D_{mean} produced by using Weng's approach all falls into the range of $[0.1, 0.2)$. D_{mean} produced by using the proposed approach is smaller than 0.1 for most (99.85%) attack profiles. In some cases, D_{mean} produced by using the proposed approach is greater than 0.2 (2 attack profiles). There are 5 attack profiles (including the 2 attack profiles just mentioned) making D_{mean} produced by Weng's approach smaller than that produced by the proposed approach. We examined the 5 attack files and found them using a similar strategy. The players manipulate the attack IDs to provide fair ratings for other products, and then provide unfair ratings for I_1. Therefore, the attack IDs may present a high similarity to the normal users. This is a classical case of strategic attack which is very difficult to counteract. In addition, the influence of the attacks can be further strengthened if the attack IDs provide a large number of fair ratings for other products, which will make the confidence of the attack IDs higher. Under this particularly vicious type of attacks, the effectiveness of the proposed approach is not as good as in other cases.

4.3 Multiple Attack Profiles

Figure 2 and 3 show the effectiveness of the proposed approach and Weng's approach when I_1 is attacked by multiple attack profiles. The X-axis is the number of attack profiles. Its value is from 0 to 10. We repeat the experiments for 1000 times for each X-axis value. At each time, the corresponding number (equal to X-axis value) of randomly selected profiles are merged together to attack I_1. For example, if X-axis value is 2, then 2 attack profiles will be randomly selected from 13,084 profiles and the unfair ratings from the 2 attack files will be merged together to attack I_1. Each attack profile contains the ratings from 20 attack IDs. When there are 10 attack profiles, there will be 200 attack IDs. As the number of the normal users providing ratings for I_1 is 134, the scenario of 10 attack profiles can actually represent the scenario that there are more unfair ratings than fair ratings. A statistical hypothesis test is performed to analyze the results as suggested in [9]. The test shows that the mean values for D_{max} and D_{mean} produced using the proposed approach are smaller than the corresponding mean values for D_{max} and D_{mean} when unfair ratings exist and when Weng et al.'s approach is adopted with 95% confidence. This implies that the outperformance of the proposed approach is statically significant.

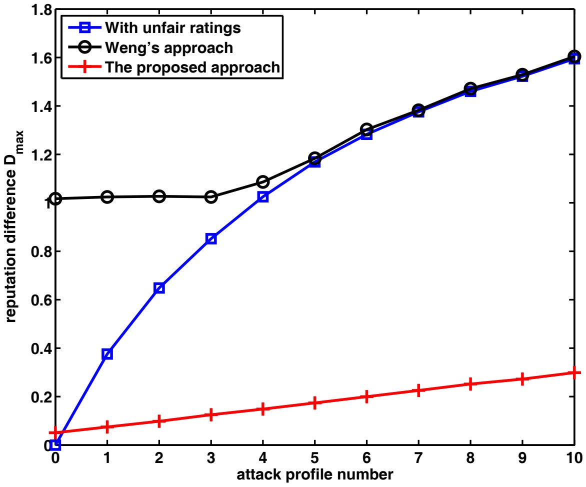

Figure 2: The largest difference D_{max} for multiple attack profiles

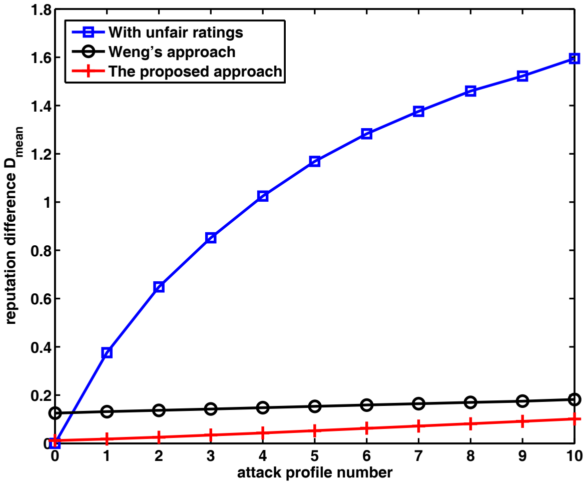

Figure 3: The mean difference D_{mean} for multiple profiles

Figure 2 and 3 show the mean values of D_{max} and D_{mean} achieved from the 1000 experiments, respectively. It can be seen that the D_{max} and D_{mean} values after using the proposed approach are smaller than those using Weng's approach. A slight increase can be noticed for D_{max} and D_{mean} after using the proposed approach. As the number of attack IDs increases with the number of attack profiles, the proposed approach cannot erase the effects of the unfair ratings completely though it can produce smaller weights for unfair ratings. Therefore, the reputation difference value (D_{max} and D_{mean}) increase slightly with the number of attack profiles increasing.

Generally speaking, Weng's approach performs worse than the proposed approach. It produces an almost 3 rating level difference in the worst case as shown in Table 3. Its performance is even worse than that without using it when there are no more than 4 attack files in the worst case as shown in Figure 2. By examining the attack profiles causing the worst performance of Weng's approach, some similarity can be found. From some particular normal users' point of view, Weng's approach produces very low credibility values for other normal users as the number of the past ratings for the particular normal user to calculate the credibility for other normal users is quite small. It outputs comparatively high credibility values for the attack IDs as the attack IDs insert ratings strategically to make the number of past ratings for the particular normal users calculating the credibility of the attack IDs comparatively large. Therefore, a large reputation difference will be resulted in as the attack IDs' ratings carry larger weights (i.e., credibility) and have a greater impact among all the ratings. For the scenarios with multiple attack profiles, the attack IDs' ratings will have a smaller impact compared with the normal ratings when there are less attack profiles when Weng's approach is not used. But when Weng's approach is used, for some particular normal users, as it gives the attack IDs' ratings higher weights, its performance will be worse than when it is not being used.

The performance of the proposed approach is better than Weng's approach because of the following reasons. First, the proposed approach can measure the similarity between users (no matter normal users or malicious users) more accurately. Second, the proposed approach takes the confidence of the witnesses in providing ratings into consideration. The combination of the similarity and confidence makes the output weights for the testimonies more reasonable.

5. DISCUSSIONS

In this section, we will further discuss some potential issues that require more considerations.

First, the similarity and confidence of a witness providing ratings are currently calculated based on the ratings regarding the commonly rated trustees without the consideration of context information and trustee behavior change. In the experiment data, each truster or witness only has at most one rating for one trustee (i.e., a product in the context of TV). But there may exist scenarios that a truster can provide multiple ratings for one trustee in different contexts. For example, if we consider a seller in an e-commerce environment as a trustee, then a buyer (i.e., truster) may have multiple transactions with the seller, and hence, provide more than one ratings regarding the seller for different items at different time windows. A way to address this scenario is to study and calculate the similarity and confidence in different contexts and time windows individually [7] [10] [12].

Second, as the experimental results show, the proposed approach performs worse when some kind of strategic attack exists. Currently, there are some ways proposed to counter this attack. For example, we can follow the majority rule to first filter the ratings which are different from majority ratings [20]. Then we calculate the weights for the remaining ratings. But using this way will face difficulty when majority ratings are unfair, or when multi-nominal ratings are adopted (e.g., it is difficult to tell rating level 4 and rating level 5 should be considered as same or not). Another way is to use the truster's direct experience as benchmark to filter the ratings different from his personal ratings as unfair [16] [18]. But using this way will face difficulty when the truster's direct experience is lacking. A stereotype model was proposed in [3] to address the scenario where direct experience is totally missing (i.e., a newcomer to online transaction system). We will continue investigating how to address this strategic attack in the future.

6. CONCLUSIONS

In this paper, we proposed a fuzzy logic based reputation model against unfair ratings. The proposed model considers three aspects – temporal, similarity and quantity – of a rating to calculate the weight of it based on fuzzy logic. It can resist the unfair ratings caused by deliberate alterations, situation changes, and subjective difference. The experimental results based on a set of real user ratings show that the proposed model can effectively mitigate the adverse effects of unfair ratings and significantly outperform the related work. In the future, we will continue studying other factors that can impact the rating weights. A more general similarity and confidence calculation which considers the context information and trustee behavior change is under exploration. Addressing the strategic attacks is also one future direction.

7. ACKNOWLEDGMENTS

The authors would like to thank the research group leaded by Dr. Yan Sun from the University of Rhode Island for generous data sharing.

8. REFERENCES

[1] www.douban.com.

[2] *Buy iTunes Ratings and Commnets – Increase iTunes Sales and Downloads.* http://www.youtube.com/watch?v=-TWV4XaxCo.

[3] C. Burnett, T. J. Norman, and K. Sycara. Bootstrapping trust evaluations through stereotypes. In *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems*, pages 241–248, 2010.

[4] A. Gelman. *Bayesian Data Analysis.* Chapman & Hall/CRC, Florida, USA, 2004.

[5] A. Jøsang and R. Ismail. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, pages 324–337, 2002.

[6] A. Jøsang, R. Ismail, and C. Boyd. A survey of trust and reputation systems for online service provision. *Decision Support System*, 43(2):618–644, 2007.

[7] S. Liu, A. C. Kot, C. Miao, and Y.-L. 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, 2012.

[8] Y. Liu and Y. Sun. Anomaly detection in feedback-based reputation systems through temporal and correlation analyssis. In *Proceedings of the 2nd IEEE International Conference on Social Computing*, 2010.

[9] D. Montgomery. *Design and analysis of experiments.* John Wiley and Sons, Chichester, UK, 2004.

[10] Z. Noorian, S. Marsh, and M. Fleming. Multi-layer cognitive filtering by behavioral modeling. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems*, pages 871–878, 2011.

[11] T. J. Ross. *Fuzzy Logic with Engineering Applications.* Wiley, 2010.

[12] M. Sensoy, J. Zhang, P. Yolum, and R. Cohen. POYRAZ: context-aware service selection under deception. *Computational Intelligence*, 25(4):335–366, 2009.

[13] G. Shafter. *A Mathematical Theory of Evidence.* Princeton University Press, 1976.

[14] Y. Sun and Y. Liu. Security of online reputation systems: evolution of attacks and defenses. *Signal Processing Magazine, Special Issue On Signal and Information Processing for Social Learning and Networking*, 29(2):87–97, 2012.

[15] W. T. L. Teacy, N. R. Jennings, A. Rogers, and M. Luck. A hierarchical bayesian trust model based on reputation and group behaviour. In *the 6th European Workshop on Multi-Agent Systems*, pages 206–212, 2008.

[16] W. T. L. Teacy, J. Patel, N. R. Jennings, and M. Luck. TRAVOS: trust and reputation in the context of inaccurate information soucres. *Autonomous Agents and Multi-Agent Systems*, 12(2):183–198, 2006.

[17] L.-X. Wang. *A course in fuzzy systems and control.* Prentice Hall, 1997.

[18] J. Weng, C. Miao, and A. Goh. An entropy-based approach to protecting rating systems from unfarir testimonies. *IEICE Trans. INF. and SYST.*, E89-D(9):2502–2511, September 2006.

[19] J. Weng, Z. Shen, C. Miao, A. Goh, and C. Leung. Credibility: How agents can handle unfair third-party testimonies in computational trust models. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 22:1286–1298, 2010.

[20] A. Whitby, A. Jøsang, and J. Indulska. Filtering out unfair ratings in bayesian reputation systems. *ICFAIN J. Management Res.*, 4(2):48–64, 2005.

[21] B. Yu and M. Singh. Detecting deception in reputation management. In *Proceedings of the 2nd International Conference on Autonomous Agents and Multiagent Systems*, pages 73–80, 2003.

[22] H. Yu, Z. Shen, and B. An. An adaptive witness selection method for reputation-based trust models. In *Proceedings of the 15th International Conference on Principles and Practice of Multi-Agent Systems (PRIMA)*, pages 184–198, 2012.

[23] H. Yu, Z. Shen, C. Miao, C. Leung, and D. Niyato. A survey of trust and reputation management systems in wireless communications. *Proceedings of the IEEE*, 98:1755–1772, 2010.