Computational Intelligence, Volume 30, Number 2, 2014

AN INTEGRATED CLUSTERING-BASED APPROACH TO FILTERING UNFAIR MULTI-NOMINAL TESTIMONIES

SIYUAN LIU,[1] JIE ZHANG,[2] CHUNYAN MIAO,[2] YIN-LENG THENG,[3] AND ALEX C. KOT[1]

[1]*School of Electrical and Electronic Engineering*
[2]*School of Computer Engineering*
[3]*Wee Kim Wee School of Communication and Information*
Nanyang Technological University, Singapore

Reputation systems have contributed much to the success of electronic marketplaces. However, the problem of unfair testimonies has to be addressed effectively to improve the robustness of reputation systems. Until now, most of the existing approaches focus only on reputation systems using binary testimonies, and thus have limited applicability and effectiveness. In this paper, We propose an **i**ntegrated **CLU**stering-**B**ased approach called **iCLUB** to filter unfair testimonies for reputation systems using multinominal testimonies, in an example application of multiagent-based e-commerce. It adopts clustering techniques and considers buyer agents' local as well as global knowledge about seller agents. Experimental evaluation demonstrates the promising results of our approach in filtering various types of unfair testimonies, its robustness against collusion attacks, and better performance compared to competing models.

Received 31 October 2011; Revised 12 March 2012; Accepted 23 June 2012; Published online 4 September 2012

Key words: clustering, multiagent-based electronic commerce, reputation system, robustness, unfair testimony.

1. INTRODUCTION

With the development of Internet technology, electronic commerce systems have been made widely accessible in our daily life, e.g., eBay, through which transactions are made conveniently. However, there are a number of challenging issues arising. With respect to this, one challenging issue is to accurately evaluate the trustworthiness of the potential sellers. Due to the nature of e-commerce, buyers and sellers usually do not meet face-to-face during an online trading process or inspect the quality of the item before a transaction is completed. Hence, accurately evaluating the trustworthiness of the potential sellers is important in electronic commerce systems. Moreover, as e-commerce is becoming more popular, it is common that there might exist a lot of sellers providing the same items at almost the same price. In such a scenario, buyers are more willing to have transactions with the sellers who are more likely to be trusted. However, buyers are also hesitant to decide which sellers to have transactions with if the buyers cannot accurately evaluate the trustworthiness of the sellers. Therefore, despite e-commerce's convenience, people are usually more concerned about its reliability when using it.

To cope with this dilemma, reputation systems have been developed for multiagent-based e-commerce (Jøsang, Ismail, and Boyd 2007). Reputation systems represent *soft security* mechanisms as a complement to the traditional information security mechanisms (Rasmusson and Janssen 1996). In a reputation system, a buyer agent can give a rating regarding his transaction parter—a seller agent—after completing the transaction. Then one buyer agent can aggregate ratings provided by other buyers regarding one seller agent to derive a reputation score, which can further be used to assist the buyer to evaluate the trustworthiness of the seller and decide whether to carry out a transaction with the seller.

Address correspondence to Siyuan Liu at School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798; e-mail: lius0036@e.ntu.edu.sg

Although reputation systems have contributed much to the success of e-commerce system, their robustness remains to be a big concern. With respect to this, the problem of *unfair testimonies* is one important issue. For instance, suppose in a reputation system, a buyer *B* is evaluating a seller *S*'s reputation to decide whether to carry out transactions with *S*. To assist *B*'s evaluation, *B* requests ratings (called testimonies[1]) from other buyers (called witnesses) who had transactions with *S* before. However, to illude *B* to buy from *S*, *S* might collude with some witnesses, who only report positive testimonies to *B* regarding *S* no matter what *S*'s real behavior is. It is highly possible that those unfairly positive testimonies will lead to *B*'s inaccurate evaluation on *S*'s reputation. As a consequence, *B* might make a wrong decision to conduct transactions with *S*. Actually, the problem of unfair testimonies also exists in the reputation systems in other application domains such as recommender systems and voting mechanisms.

Various approaches, such as the beta reputation system (Whitby, Jøsang, and Indulska 2005), TRAVOS model (Teacy et al. 2006), the personalized approach (Zhang and Cohen 2008), and cognitive filtering by behavioral modeling (Noorian, Marsh, and Fleming 2011) have been proposed to cope with the problem of unfair testimonies in reputation systems. However, most of these approaches focus only on the reputation systems using binary testimonies, and thus are not applicable to the ones supporting multi-nominal testimonies (e.g., the Dirichlet reputation system proposed by Jøsang and Haller (2007) and Fung et al. (2011)) where more than two levels of ratings are also accepted.

In this paper, we propose an integrated CLUstering-Based approach (iCLUB)[2] to effectively filter unfair testimonies for this kind of generic reputation systems.[3] More specifically, our approach adopts clustering techniques and integrates two components, Local and Global. The Local component makes use of only buyers' knowledge about the sellers being currently evaluated (called target sellers). The Global component makes use of buyers' knowledge about other sellers that the buyers have previously encountered. This is particularly useful when the buyers do not have much experience with the target sellers.

We carry out experiments in a simulated e-commerce environment where witnesses may provide different types of unfair testimonies and may collude with each other. We first use experiments to explore the impact of parameter values on the accuracy of iCLUB for filtering unfair testimonies. Second, we present the accuracy of iCLUB against different types of unfair testimonies, especially in the collusion attack scenario. Third, we integrate iCLUB with a multinominal reputation system to demonstrate that it can improve the robustness of the reputation system. Finally, we conduct comparison experiments to compare iCLUB with two representative filtering approaches. Experimental results demonstrate that our approach is effective in filtering unfair testimonies, it is robust against collusion attacks to a good extent, and our approach outperforms other competing models in the scenario where only binary testimonies are allowed. Thus, our approach is proven to improve the robustness of reputation systems and contribute to the goal of developing reliable e-commerce for users.

The remainder of this paper is structured as follows: A review of related work is given in Section 2. Section 3 provides a brief description of the notations we use. Section 4 presents the proposed iCLUB approach. Section 5 provides an example to illustrate how the proposed iCLUB approach works. The experimental studies and results are presented in Section 6. Finally, Section 7 concludes the paper with an overview of the future work.

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

[2] It is extended from our previous work (Liu et al. 2011) by including more detailed descriptions of the approach, some examples and more extensive experimentation.

[3] We will show that our approach is also applicable to the reputation systems using only binary testimonies.

2. RELATED WORK

To find an effective approach to handle the problem of unfair testimonies has been studied for a long time. Various approaches have been proposed. Here, we briefly summarize some representative approaches.

Jøsang and Ismail (2002) proposed the beta reputation system (BRS). In BRS, ratings for a seller are expressed as either positive or negative, which can be considered as two events in the beta probability distribution (Gelman 2004). A seller's reputation is estimated as the expected value of the positive event happening in the future by using the aggregated numbers of the positive and negative ratings regarding the seller from all buyers. To address the problem of unfair testimonies, Whitby et al. (2005) further proposed an iterated filtering approach by testing whether a witness's testimonies are outside of the q quantile or $1 - q$ quantitle of the majority testimonies. If the testing witness's testimonies are beyond the range, then the testimonies are considered as unfair and discarded. However, this approach has the disadvantage that its filtering accuracy decreases rapidly with the increase of the percentage of the dishonest witnesses.

Weng, Miao, and Goh (2006) proposed an entropy-based approach to filter unfair testimonies for BRS. The approach first calculates the quality of the buyer's personal ratings and the quality of a particular witness's ratings by using an entropy based metric. Then it measures the difference between the two quality values. If the difference exceeds the threshold, the witness's ratings are considered as unfair and discarded. However, because of using entropy, the approach cannot distinguish the quality between the symmetry positive and negative testimonies pair (i.e., the number of a witness's positive ratings is the same as the number of another witness's negative ratings, and the number of the former's negative ratings is the same as the number of the latter's positive ratings), which will lead to that unfair testimonies cannot be accurately identified.

Teacy et al. (2006) proposed the TRAVOS model which is also based on beta probability distribution to evaluate the reputation of agents in agent-based virtual organizations. This approach first estimates the accuracy of a witness's testimonies by comparing the witness's previous testimonies with the buyer's personal ratings regarding the commonly rated sellers. Then the approach adjusts the witness's testimonies according to the obtained accuracy. However, the computation of the witnesses' accuracy is quite time-consuming if the number of the witnesses is large or the amount of a witness' testimonies is large as TRAVOS repeatedly goes through a witness's testimonies each time. Sharing some similarities with TRAVOS, Regan, Poupart, and Cohen (2006) proposed the BLADE model. The BLADE model uses Bayesian learning to reinterpret a witness's ratings instead of filtering the unfair testimonies, which is similar to the step of estimating accuracy of a witness's testimonies in TRAVOS. But the reinterpretation is dependent on the assumption that the witness's behavior keeps consistent. Otherwise, the reinterpretation of the witness's ratings may be incorrect. As an extension of BLADE, Teacy et al. (2008) proposed the HABIT model, which can be used for discrete and continuous ratings. But HABIT shares the same disadvantage as BLADE that the witness's behavior is assumed to be consistent.

The personalized approach proposed by Zhang and Cohen (2008) has some similar spirit as our approach. The personalized approach uses private reputation and public reputation to measure the reliability of a witness. Private reputation is calculated by comparing the witness's ratings with the buyer's personal ratings regarding the commonly rated sellers. Public reputation is estimated by comparing the witness's ratings with other witnesses' ratings regarding all sellers. This approach has the advantage that it considers two aspects of the reliability of a witness. However, this approach calculates a witness's reputation as a common value for all sellers. Therefore, it may not work when the witness changes his behavior from one seller to another seller.

A multilayer cognitive filtering approach by behavior modeling for binary ratings was proposed by Noorian et al. (2011). In this approach, a witness's testimonies go through two filtering layers. In the first layer, the approach calculates the average difference between the witness's testimonies and the buyer's personal ratings. If the difference value exceeds a threshold value, the witness's testimonies are filtered. In the second layer, the approach models the behaviors of the witnesses who have passed the first layer by measuring the similarity between the witnesses' testimonies and the buyer's personal ratings for all sellers. A tendency value is achieved by using the similarity value. Finally, the witness's behavior is identified as optimistic or pessimistic by considering the similarity and the tendency value. The approach has the advantage that it proposed the idea to differentiate the witness's behavior pattern. But it assumes that the witness's behavior is consistent for all sellers.

An approach also using clustering but designed for the binary testimony case was proposed by Dellarocas (2000). A divisive clustering algorithm is used to separate the testimonies into two clusters—the cluster including lower ratings and the cluster including higher ratings. The testimonies in the higher testimonies cluster are considered as unfairly high testimonies, and are discarded. However, this approach cannot effectively handle unfairly low testimonies. Another approach that also applies clustering was proposed in our previous work (Liu et al. 2010). However, it makes use of only buyers' own knowledge about the target sellers. Ratings for other sellers are not considered.

As described above, most of the approaches for handling the problem of unfair testimonies are designed for reputation systems using binary ratings. In contrast, our iCLUB approach is applicable to reputation systems using multinominal ratings. Our approach has the advantage that it considers the buyer's personal ratings with a different importance from witnesses' ratings. It also considers ratings for the sellers other than the one that is currently under evaluation, to cope with the situation where a buyer has very limited experience with that seller. Several approaches mentioned above assume that the witness's behavior is consistent for all sellers. Our approach does not entirely rely on this assumption. The witness that is honest for some sellers may not be considered as honest in our approach. To be considered as honest, a witness has to be in the group that has the largest number of witnesses who are honest regarding those sellers. Therefore, if the witness is dishonest regarding the target seller, it may end up with falling in the group that mainly involves dishonest witnesses.

3. NOTATIONS

Before getting into the details of the iCLUB approach, we first introduce some notations in this section. Suppose that in a reputation system, there are M seller agents $\{S_1, S_2, \dots, S_M\}$, and N buyer agents $\{B_1, B_2, \dots, B_N\}$. After each transaction between a buyer agent B_n $(1 \leq n \leq N)$ and a seller agent S_m $(1 \leq m \leq M)$ is completed, B_n can rate S_m's behavior by a rating level from a set of predefined discrete rating levels. Suppose that there are K different rating levels and each rating level is indexed by i. If B_n rates S_m's behavior as rating level i, B_n's rating $r_{S_m}^{B_n}$ for S_m is represented as a row vector:

$$r_{S_m}^{B_n} = [0, \dots, 0, 1, 0 \dots, 0],$$

where the ith rating level is 1 $(1 \leq i \leq K)$. For example, suppose that $K = 5$ and B_n rates S_m's behavior as 4 after one transaction, then $r_{S_m}^{B_n} = [0, 0, 0, 1, 0]$. The aggregated ratings

$R_{S_m}^{B_n}$ from B_n for S_m can be represented as a cumulative vector, expressed as:

$$R_{S_m}^{B_n} = \left[R_{S_m}^{B_n}(1), \dots, R_{S_m}^{B_n}(i), \dots, R_{S_m}^{B_n}(K)\right],$$

where $R_{S_m}^{B_n}(i)$ is the aggregated result of $r_{S_m}^{B_n}(i)$ $(1 \leq i \leq K)$. The updating of $R_{S_m}^{B_n}$ can be achieved by adding the new rating vector $r_{S_m}^{B_n}$ to the previous rating vector $R_{S_m}^{B_n}$.[4]

When B_n is evaluating S_m's reputation, it can collect rating vectors from other buyer agents to facilitate his evaluation. Then the set of these buyer agents $W_{S_m}^{B_n}$ who provide rating vectors to B_n regarding S_m are expressed as:

$$W_{S_m}^{B_n} = \left\{B_j \mid j \neq n \wedge \left\|R_{S_m}^{B_j}\right\| \neq 0\right\}.$$

From B_n's point of view, $W_{S_m}^{B_n}$ is called the set of witness agents regarding S_m (each buyer agent in $W_{S_m}^{B_n}$ is a witness agent), and the rating vector provided by each witness is called testimonies from this witness. Then the local information $L_{S_m}^{B_n}$ regarding S_m can be expressed as:

$$L_{S_m}^{B_n} = \begin{cases} \left\{R_{S_m}^{B_j} \middle| B_j \in W_{S_m}^{B_n}\right\}, & \text{if } \left\|R_{S_m}^{B_n}\right\| = 0 \\ \left\{R_{S_m}^{B_j} \middle| B_j \in W_{S_m}^{B_n} \cup \{B_n\}\right\}, & \text{if } \left\|R_{S_m}^{B_n}\right\| \neq 0. \end{cases}$$

And the global information G^{B_n} can be expressed as:

$$G^{B_n} = \bigcup_{m=1}^{M} L_{S_m}^{B_n}.$$

It in fact also contains the local information of B_n about the seller agent whose reputation is currently under evaluation. As mentioned in Section 1, though B_n can use the testimonies to facilitate the evaluation regarding S_m's reputation, the testimonies may mislead B_n's evaluation if the witnesses do not provide testimonies in an honest way. This may even result in an opposite evaluation situation, e.g., where a very low reputation is estimated regarding a reputable seller.

4. THE iCLUB APPROACH

In this section, we present our integrated CLUstering-Based (iCLUB) approach for effectively filtering unfair testimonies. Before elaborating the iCLUB approach, we need to clarify what we mean by "unfair testimonies." According to the definition of trust—the opinion (more technically, an evaluation) of an entity toward a person, a group of people, or an organization on a certain criterion (Jøsang et al. 2007), the trustworthiness of the target seller agent S_t (the seller agent whose reputation is under evaluation) is the opinion held by a buyer agent toward the seller. Therefore, we consider that whether a witness's testimonies are unfair (or whether the witness is trustworthy in reporting testimonies) should also be the opinion held by the buyer agent toward the witness's testimonies.[5] An intuition is that if the witnesses' testimonies are more similar to the buyer agent's past ratings for sellers,

[4] The adding mentioned is referred to as a matrix addition.

[5] This is what the "i" in the name "iCLUB" reflects, and this name also implies that other agents providing similar testimonies as the buyer agent will be allowed to join his club (cluster) and be considered as honest by the buyer.

then the more likely the testimonies are fair. On the contrary, the more different from the buyer agent's past ratings the testimonies are, then the more likely the testimonies are unfair. Therefore, if we can group the witnesses who provide similar testimonies as the buyer agent's past experience with sellers, we can actually find the honest witnesses and filter the unfair testimonies provided by other witnesses. Here we emphasize that the "unfairness" does not exactly mean that the witness report ratings intentionally unfairly. The "unfairness" also possibly comes from the subjective difference between the buyer and the witness. In our current approach, we do not differentiate the two kinds of unfairness.

To group the similar testimonies together, the technique of clustering is a good choice. Clustering is originally used to assign a set of observations into subsets (called clusters) so that observations in the same cluster are similar to each other according to some criteria (Duba, Hart, and Stork 2001). There are many clustering methods designed, such as *k*-means clustering and hierarchical clustering. In our proposed iCLUB approach, we use a density-based clustering approach (Ester et al. 1996) as it can discover clusters of arbitrary shape without specifying the number of clusters.

Our iCLUB approach integrates two components, Local and Global. The Local component applies clustering only on a buyer agent's local information, in the scenario where the buyer agent has a sufficient number of transactions with the target seller S_t. Otherwise, the Global component applies the clustering on the buyer agent's global information. More details are given in the subsequent sections.

4.1. Making Use of Only Local Information

Considering the rating vector from the buyer agent or a particular witness as a feature vector in a K-dimension space, the iCLUB approach groups the feature vectors with the same similarity into one cluster. After clustering, the rating vectors will be considered as unfair testimonies if they are not in the cluster which includes the buyer agent's personal rating vector. A pseudo code summary of this process is given in Algorithm 1.

Algorithm 1: Making Use of Local Information

```
Procedure: Local($S_t$, $B$)
Input    : $S_t$, seller whose reputation is evaluated;
           $B$, buyer evaluating $S_t$'s reputation;
Output   : A set of honest witnesses regarding $S_t$;
1 Collect local information regarding $S_t$ as $L^B_{S_t}$;
2 $C_1, C_2, ..., C_Z$ = DBSCAN($L^B_{S_t}$);
3 $\exists b, R^B_{S_t} \in C_b \ (1 \leqslant b \leqslant Z)$;
4 Return $W_T = \{B_i | R^{B_i}_{s_t} \in C_b \wedge B_i \neq B\}$;
```

The Local component of our approach first collects the local information regarding S_t (see Line 1).[6] DBSCAN (Ester et al. 1996), a density-based clustering routine, is then applied on the collected testimonies $L^B_{S_t}$ to generate a set of clusters (Line 2). Before the clustering process, we need to normalize the rating vectors in $L^B_{S_t}$. The normalization for each rating vector is achieved by having the value of each dimension divided by the sum of the value of each dimension. For example, suppose that a rating vector is [0, 1, 4, 2, 1], then the

[6] How to discover the distributed testimonies is also an important issue. But it is not the focus of our current work.

normalized rating vector is [0, 0.125, 0.5, 0.25, 0.125]. The DBSCAN clustering approach works just like using a circle with the radius r to scan the whole feature space from an arbitrarily selected point. The points will be grouped with the starting point together if they are within the circle area of the starting point (we currently use 2-norm distance to calculate whether a point is within the circle area). Then the scanning process will continue by starting from the starting point and the points which are just included. The scanning process for this starting point will stop when no points are circle-area reachable from all the points included in the starting point's cluster. All the points in the cluster are labeled as "touched." Then another arbitrary point which is not touched in last scan process is selected, and the same scanning process starts again. The whole clustering process will stop when no points are untouched. The most important parameter for the DBSCAN clustering approach to work correctly is the radius of the circle used to scan the feature space. Some work has been done to investigate the optimum radius value setting, such as the work of Ester et al. (1996) and Ankerst et al. (1999). In Section 6.1, we will carry out experiments to investigate the impact of the radius value on the filtering accuracy of our approach, and establish a feasible optimum radius value for our approach to work accurately. After the clusters are generated, the Local component returns as honest witnesses the set of witnesses whose rating vectors are included in the same cluster as the buyer agent's rating vector (Lines 3−4).

4.2. Making Use of Global Information

As pointed out, the Local component is able to work effectively only when the buyer agent has some transactions with the target seller agent S_t. However, it is possible that the buyer agent does not have much experience with S_t in some scenarios, for example, where the buyer encounters the seller agent for the first time. In this case, we have to depend on the buyer's global information to filter unfair testimonies. That is the Global component. A pseudo code summary of how it works is given in Algorithm 2.

Algorithm 2: Making Use of Global Information

```
Procedure: Global($S_t$, $B$)
Input    : $S_t$, seller whose reputation is evaluated;
           $B$, buyer evaluating $S_t$'s reputation;
Output   : A set of honest witnesses regarding $S_t$;
foreach seller agent $S_i$ $(1 \leqslant i \leqslant M, i \neq t)$ do
    if $B$ has transactions with $S_i$ (i.e., $||R^B_{S_i}|| \neq 0$) then
        $W_i = \text{Local}(S_i, B)$;
$W_F = \bigcap_{i=1}^{M} W_i$, where $||R^B_{S_i}|| \neq 0$ and $i \neq t$;
$C_1, C_2, ..., C_L = \text{DBSCAN}(L^B_{S_t})$;
foreach cluster $C_j$ $(1 \leqslant j \leqslant L)$ do
    $W_{C_j} = \{B_i | R^{B_i}_{S_t} \in C_j\}$;
    if $W_F \neq \emptyset$ then
        $W_{F_j} = W_F \bigcap W_{C_j}$;
    else
        $W_{F_j} = W_{C_j}$;
$q = \arg\{\max_j(|W_{F_j}|)\}, j = 1, 2, \cdots, L$;
Return $W_T = \{B_i | R^{B_i}_{S_t} \in C_q\}$ as honest witnesses;
```

The Global component first finds the honest witnesses for each seller agent with whom the buyer agent has transactions, using the Local() procedure (Lines 1–3). Then, a set of common honest witnesses W_F are formed as the intersection of the set of the honest witnesses for each seller agent except S_t (Line 4). The Global component continues by applying the DBSCAN routine to obtain the clustering result for S_t (Line 5). It then calculates the intersection of W_F with the witnesses whose rating vectors are in each cluster achieved in Line 5 if W_F is not an empty set (Lines 6–11). Finally, it returns as honest witnesses the ones whose rating vectors are in the cluster which has the largest intersection result with W_F (Lines 12–13). Note that if two or more clusters are identified to contain the same largest size of intersection with W_F, the one that contains the buyer agent B's rating vector (if any) will be considered as the honest witnesses cluster.

In brief, the Global component of our iCLUB approach makes use of the buyer's experience with sellers in the reputation system to find a set of witnesses who are honest regarding those sellers, and then uses this information to find honest witnesses regarding the seller who is currently under evaluation. As can be noticed from Lines 4 and 12 in Algorithm 2, a witness may be considered as honest only if he has been honest regarding the sellers encountered by the buyer. This restriction is based on the assumption that if a witness is honest for all the common sellers encountered by the buyer, then it is more likely that the witness will be honest for the target seller. Another stronger restriction is that the witness has to belong to the largest intersection cluster regarding the target seller. This restriction is set to cope with collusion attacks to a great extent where a group of witnesses collude in providing unfair testimonies to the target seller, by intentionally being honest regarding other sellers to build up trust from buyers. We demonstrate this through experiments in Section 6.2.

4.3. The Integrated Approach

From the Global() procedure in Algorithm 2, we can see that the Global component has already integrated the Local() procedure (Lines 3–5). Our iCLUB approach further integrates these two components using a threshold ε, as summarized in Algorithm 3.

Algorithm 3: Integrate Local and Global Information

```
Procedure: iCLUB($S_t$, $B$)
Input    : $S_t$, seller whose reputation is evaluated;
           $B$, buyer evaluating $S_t$'s reputation;
Output   : A set of honest witnesses regarding $S_t$;
if $\sum_{i=1}^{K} R_{S_t}^{B}(i) \geqslant \varepsilon$ then
    Return Local($S_t$,$B$);
else
    Return Global($S_t$,$B$);
```

There are some points worth of mentioning. First, the triggering of the use of global information is controlled by the threshold ε. If the number of transactions between the buyer agent and S_t is smaller than ε, the Global component will be triggered. We will investigate how to properly set ε in Section 6.1. Second, it can be noticed that the Global component of our approach can work effectively when the buyer agent has no sufficient transactions with S_t but has transactions with other seller agents. However, for a buyer agent who is a newcomer

TABLE 1. Rating Vectors for the Local Component Working Example Scenario.

	Reported rating vectors					Normalized rating vectors				
Rating level	1	2	3	4	5	1	2	3	4	5
W_1	0	0	0	0	31	0	0	0	0	1.0000
W_2	0	0	0	0	14	0	0	0	0	1.0000
W_3	0	0	0	0	21	0	0	0	0	1.0000
W_4	0	0	4	6	33	0	0	0.0930	0.1395	0.7674
W_5	0	0	2	4	6	0	0	0.1667	0.3333	0.5000
W_6	0	0	2	14	32	0	0	0.0417	0.2917	0.6667
W_7	0	0	4	10	26	0	0	0.1000	0.2500	0.6500
W_8	0	0	0	0	30	0	0	0	0	1.0000
W_9	18	12	4	0	0	0.5294	0.3529	0.1176	0	0
W_{10}	12	6	2	0	0	0.6000	0.3000	0.1000	0	0
B	17	8	4	0	0	0.5862	0.2759	0.1379	0	0

to the system, this agent does not have any transactions with any of the sellers. In this case, the Global component has to follow the majority rule (see Lines 11–12 in Algorithm 2) which is similar to the approaches reviewed in Section 2, such as BRS (Whitby et al. 2005) and the approach proposed by Dellarocas (2000).

5. EXAMPLES

In this section, we use some examples to demonstrate how the iCLUB approach works. These examples involve ten witnesses, indexed by $W_1, W_2, \ldots, W_{10}$, one buyer B and five sellers, indexed by $S_1, S_2, \ldots, S_5$. Five rating levels are used as an illustration. The radius value used for DBSCAN clustering is 0.3, and the threshold value to trigger the global component is set as $\varepsilon = 10$.

5.1. Local Component Example

When there are more than ε transactions between the buyer and the target seller, the Local component is called. Suppose that S_5's reputation is under evaluation. W_1 to W_8 are dishonest witnesses. The numbers of transactions rated as rating level 1 to 5 by each witness and B are shown in Table 1.

After DBSCAN clustering, there are two clusters achieved. W_1, W_2, W_3, W_4, W_5, W_6, W_7, and W_8 are in one cluster. W_9, W_{10}, and B are in the other cluster. As indicated in Algorithm 1, the cluster including the buyer's rating vector is kept. W_9 and W_{10} are considered as honest witnesses. This result is consistent with the initial setting.

5.2. Global Component Example

When the buyer does not have enough transactions with the seller (in this example, it is when the buyer has less than 10 transactions with the seller), we need the global information to filter the unfair testimonies. Suppose that whether the witnesses are honest for the five sellers are shown in Table 2.

Table 2. The Witnesses' Honesty Regarding Respective Sellers.

	Dishonest	Honest
S_1	W_1	$W_2, W_3, W_4, W_5, W_6, W_7, W_8, W_9, W_{10}$
S_2	W_1, W_2, W_3	$W_4, W_5, W_6, W_7, W_8, W_9, W_{10}$
S_3	W_2, W_4	$W_1, W_3, W_5, W_6, W_7, W_8, W_9, W_{10}$
S_4	$W_1, W_2, W_3, W_4, W_5, W_6, W_7$	W_8, W_9, W_{10}
S_5	W_5, W_8	$W_1, W_2, W_3, W_4, W_6, W_7, W_9, W_{10}$

Also suppose that the buyer does not have enough transactions with S_5 whose reputation is under evaluation, but he has enough transactions with other four sellers (S_1, S_2, S_3, and S_4). We first use DBSCAN clustering for the four sellers. By assuming the clustering result is consistent with the setting, we then get W_F:

$$\begin{aligned} W_F &= \{W_2, W_3, W_4, W_5, W_6, W_7, W_8, W_9, W_{10}\} \\ &\quad \bigcap \{W_4, W_5, W_6, W_7, W_8, W_9, W_{10}\} \\ &\quad \bigcap \{W_1, W_3, W_5, W_6, W_7, W_8, W_9, W_{10}\} \\ &\quad \bigcap \{W_8, W_9, W_{10}\} \\ &= \{W_8, W_9, W_{10}\}. \end{aligned}$$

Second, we use DBSCAN to cluster S_5's testimonies, and assume that we get two clusters—C_1 including W_5 and W_8's testimonies, and C_2 including W_1, W_2, W_3, W_4, W_6, W_7, W_9, and W_{10}'s testimonies:

$$W_{C_1} = \{W_5, W_8\},$$

$$W_{C_2} = \{W_1, W_2, W_3, W_4, W_6, W_7, W_9, W_{10}\}.$$

Then we calculate the intersection of W_F with W_{C_1} and W_{C_2} and get W_{F_1} and W_{F_2}, respectively:

$$W_{F_1} = W_F \bigcap W_{C_1} = \{W_8\},$$

$$W_{F_2} = W_F \bigcap W_{C_2} = \{W_9, W_{10}\}.$$

As the cluster C_2 has a larger intersection set size with W_F, C_2 is kept and W_1, W_2, W_3, W_4, W_6, W_7, W_9, and W_{10} are considered as honest witnesses for S_5.

6. EXPERIMENTAL STUDIES

We carry out four sets of experiments to evaluate our iCLUB approach. The first set investigates the relationship between the radius value of DBSCAN and the accuracy of iCLUB in filtering unfair testimonies, and explores the proper ε value used to trigger the Global component. The aim of the second set of experiments is to examine the accuracy of our approach in various scenarios, and in particular, its robustness against collusion attacks. The third experiment is to integrate the iCLUB approach with the Dirichlet reputation system (Jøsang and Haller 2007) to examine whether the iCLUB approach will improve the reputation system's robustness. The fourth set is to compare iCLUB with other representative approaches (BRS (Whitby et al. 2005) and TRAVOS (Teacy et al. 2006)) in terms of the accuracy of filtering unfair testimonies and estimating seller reputation.

In our experiments, we use Matthew's correlation coefficient (MCC) (Matthews 1975) to measure the accuracy of filtering unfair testimonies. MCC is a convenient measure for the accuracy of binary classifications. Its computation is as follows:

$$\mathrm{MCC} = \frac{t_p \times t_n - f_p \times f_n}{(t_p + f_p) \times (t_p + f_n) \times (t_n + f_p) \times (t_n + f_n)},$$

where f_p, t_p, f_n, and t_n represent the numbers of false positives, true positives, false negatives, and true negatives, respectively. In our experiments, a true positive means that an honest witness is correctly detected as honest; a false positive means that a dishonest one is incorrectly detected as honest; a true negative means that a dishonest one is correctly filtered out as dishonest; a false negative means that an honest witness is incorrectly filtered out as dishonest. MCC value is between -1 and 1 where 1 represents a perfect filtering result, -1 represents an inverse filtering result, and 0 represents a random filtering result. As an addition, we use false positive rate (FPR) and false negative rate (FNR) to measure the accuracy of filtering testimonies in more details. Their computations are as follows:

$$\mathrm{FPR} = \frac{f_p}{f_p + t_n}, \mathrm{FNR} = \frac{f_n}{t_p + f_n}.$$

In our experiments investigating the accuracy of filtering unfair testimonies when multinominal rating levels are applied, four types of dishonest witnesses are involved:

(1) *ballot-stuffing* witnesses (Dellarocas 2000) give testimonies that seller agents behave well regardless of the true behaviors of the seller agents;
(2) *badmouthing* witnesses (Dellarocas 2000) give testimonies that seller agents behave badly regardless of the true behaviors of the seller agents.
(3) γ*-low-shifting* witnesses give testimonies that are γ levels lower than the real ratings.
(4) γ*-high-shifting* witnesses give testimonies that are γ levels higher than the real ratings.

6.1. Choosing Proper Parameters

This set of experiments is to investigate how the radius value of DBSCAN clustering will impact the accuracy of our approach and how to set ε value (Algorithm 3) to trigger the Global component. To investigate the influence of DBSCAN radius value, we focus on the Local component as it is sufficient to explore the impact. We simulate a trading community that involves 1 seller agent S, ω witnesses, and 1 buyer agent B. In each round of a simulation, an initial willingness (iw) value ($0 \leq iw \leq 1$) is randomly generated for the seller to represent how much the seller is willing to cooperate. According to the generated iw value, different types of dishonest witnesses are generated for different types of sellers. The relationship between iw value and the types of dishonest witnesses is shown in Table 3.

Each witness or B has I transactions with the seller. For each transaction, one willingness value will be generated from a normal distribution whose mean is iw, and standard deviation is δ. The mapping between the willingness value for each transaction and the rating level for S is shown in Table 4. In this way, a seller's behavior is represented by the normal distribution corresponding to his initial willingness value. For example, if seller S's initial willingness value is 0.2, most of the ratings for him should be 1. When honest witnesses or buyers have more transactions with S, his behavior will be represented by their rating vectors more precisely.

TABLE 3. Relationship between iw Value and Types of Dishonest Witnesses.

iw value	Types of dishonest witnesses
$0 \le iw \le 0.3$	Ballot-stuffing, γ-high-shifting
$0.3 < iw < 0.7$	Badmouthing, ballot-stuffing, γ-low-shifting, γ-high-shifting
$0.7 \le iw \le 1$	Badmouthing, γ-low-shifting

TABLE 4. From Willingness to Rating Level.

Willingness	$(-\infty,0.2]$	$(0.2,0.4]$	$(0.4,0.6]$	$(0.6,0.8]$	$(0.8,\infty)$
Rating	1	2	3	4	5

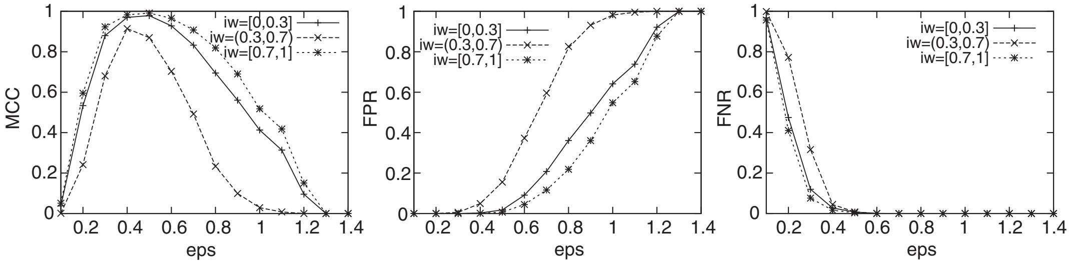
TABLE 5. Simulation Parameters, Meanings, and Values.

Parameter	Meaning	Value
ω	The number of witnesses	$\{10,100\}$
I	The number of transactions between each witness or B and S	$\{10,100\}$
δ	The standard deviation of the normal distribution to simulate S's behavior	$\{0.2,0.3\}$
P_{unfair}	The percentage of the dishonest witnesses	{40%, 80%}
γ	The shifting level of the reported ratings from the γ-low-shifting witnesses or the γ-high-shifting witnesses	$\{1, 2, 3, 4\}$

We explore how the radius value will impact the accuracy of the filtering approach for different levels of scalability and stability. Here scalability represents the number of witnesses who will report ratings to the buyer, and stability represents the number of transactions between each witness or B and S. A high scalability means that there are a lot of witnesses. On the contrary, a low scalability means that there are only a few witnesses. A high stability means that there are a large number of transactions between S and the witnesses (or B), and a low stability means that there are only a few transactions between S and the witnesses (or B). We set two levels of scalability and stability—high (i.e., 100 witnesses and 100 transactions) and low (i.e., 10 witnesses and 10 transactions). Table 5 lists the parameter meanings and values we use in our simulation.

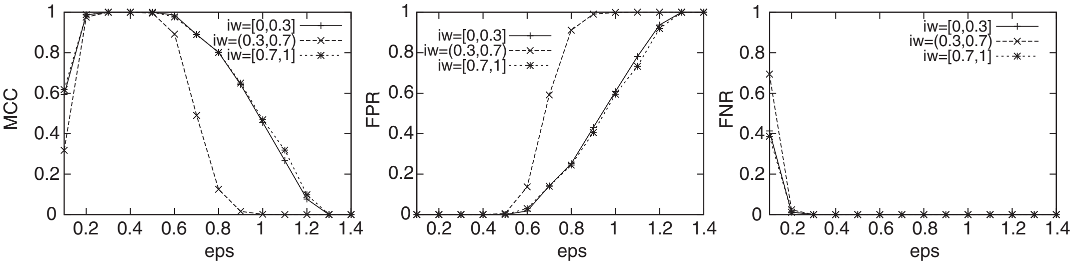
In this table, P_{unfair} is the total percentage of the dishonest witnesses. In each round of a simulation, the percentage of each type of dishonest witnesses is equally distributed. For example, if $p_{\text{unfair}} = 40\%$, and there are badmouthing witnesses and γ-low-shifting witnesses, the percentage of each type of dishonest witnesses is 20%. In general, we have 64 parameter value combinations. Therefore, there are 64 scenarios simulated in total. We run 10,000 rounds for each simulation to achieve a statistical accuracy. As an illustration, we show the results of the MCC, FPR, and FNR value changes with DBSCAN radius (i.e., eps) increasing from 0.1 to 1.4 when $\gamma = 2$ and $\delta = 0.2$.

Figure 1 shows the MCC, FPR, and FRR results when the scalability is low (i.e., $\omega = 10$) and the stability is also low (i.e., $I = 10$) . According to the results, a higher MCC value can



MCC FPR FNR

FIGURE 1. Varying radius when $I = 10$, $\omega = 10$, $P_{\text{unfair}} = 40\%$.



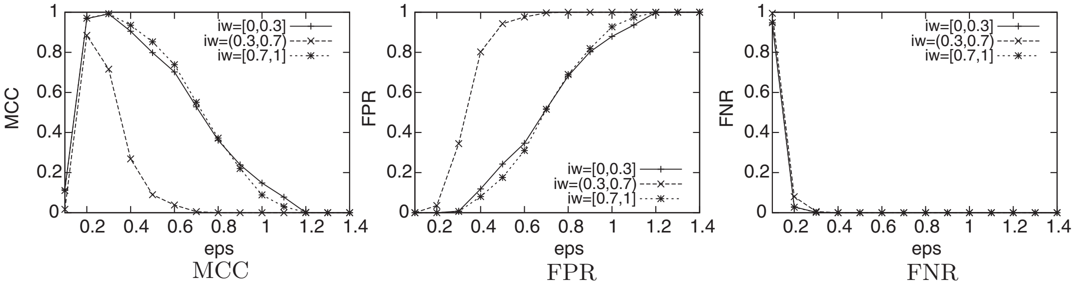
MCC FPR FNR

FIGURE 2. Varying radius when $I = 100$, $\omega = 10$, $P_{\text{unfair}} = 40\%$.

be achieved when the DBSCAN radius value is in the range of $[0.3, 0.5]$. When the radius value is too small (i.e., 0.1 to 0.3), there are more false negatives, meaning that there are more honest witnesses misclassified as dishonest witnesses. When the radius value increases from 0.5 to 1.4, there are more false positives, meaning that there are more dishonest witnesses misclassified as honest witnesses. The reason is as follows. When the stability is low, some honest witnesses or the buyer's rating vector may not represent the seller's behavior. When the radius is small, these honest witnesses cannot be grouped into the same cluster as the buyer's cluster, resulting in that they are incorrectly filtered as dishonest witnesses. When the radius is large, such as 1.4, the dishonest witnesses are included in the honest witnesses' cluster as the DBSCAN scanning radius is too large to differentiate the honest witnesses' rating vectors from the dishonest witnesses' rating vectors. Therefore, the FPR value increases with the radius value increasing.

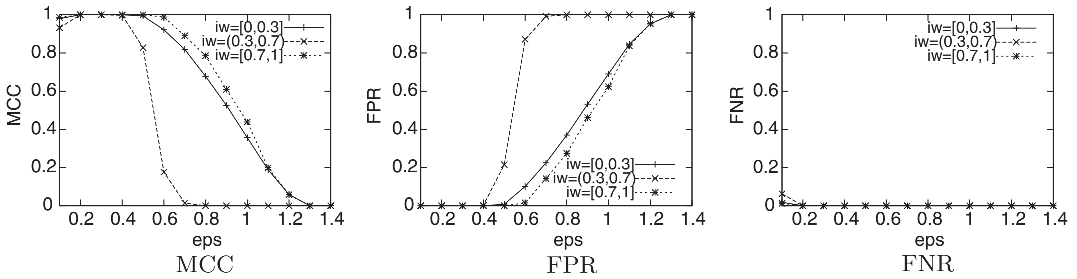
Figure 2 shows the MCC, PPR, and FNR results when the scalability is low (i.e., $\omega = 10$) and the stability is high (i.e., $I = 100$). As there are more transactions between each witness or the buyer and the seller, the honest witnesses or the buyer's rating vector can represent the seller's behavior more accurately. Compared to the results when the stability is low, there is a larger workable radius range—about $[0.2, 0.6]$. Similar to the results when the stability is low, a smaller radius value will lead to a larger FNR value, and a larger radius value will lead to a larger FPR value.

Figure 3 shows the MCC, FPR, and FNR results when the scalability is high (i.e., $\omega = 100$) and the stability is low (i.e., $I = 10$). It can be noticed that the workable radius value range is quite small. As the stability is low and there are a lot of witnesses, there is a larger diversity among the witnesses' rating vectors, which makes the correct clustering



MCC FPR FNR

Figure 3. Varying radius when $I = 10$, $\omega = 100$, $P_{\text{unfair}} = 40\%$.



MCC FPR FNR

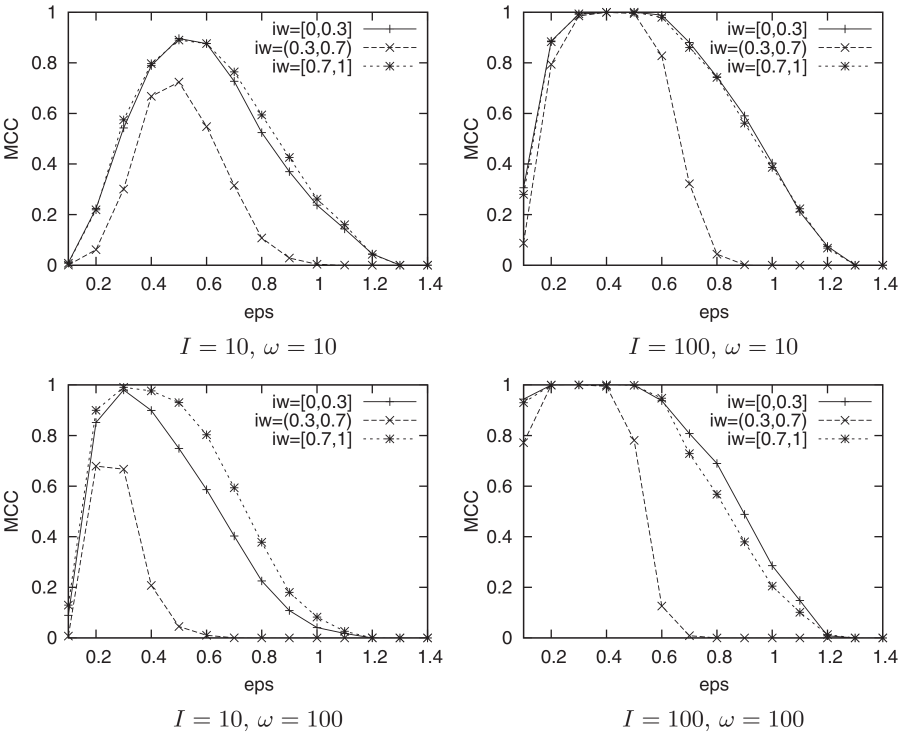
Figure 4. Varying radius when $I = 100$, $\omega = 100$, $P_{\text{unfair}} = 40\%$.

result difficult to be achieved. Similar to the results in the previous two scenarios, more errors happen as false positives when the radius value increases.

Figure 4 shows the MCC, FPR, and FNR results when the scalability is high (i.e., $\omega = 100$) and the stability is also high (i.e., $I = 100$). Similar to the results shown in Figure 2, there is a larger workable radius range. It also can be noticed that there are almost no false negatives. As the stability is high, the honest witnesses or the buyer's rating vector can reflect the seller's behavior more precisely. A small radius value (i.e., 0.1$\sim$0.3) also can correctly differentiate the honest witnesses from dishonest witnesses.

As an illustration, Figure 5 shows the MCC results when $P_{\text{unfair}} = 80\%$ for the four scenarios (the scalability is high or low, and the stability is high or low). We can see that the MCC presents the similar trend as the the results of the corresponding scenario when $P_{\text{unfair}} = 40\%$.

According to the simulation results (including the results we do not present here), it is difficult to find a working radius value when $\gamma = 1$ as there is only a small difference between γ-shifting witnesses' rating vectors and the honest witnesses' rating vectors. But from another point of view, as the difference is quite small, we can consider it as a subtle subjective difference and treat these witnesses as honest witnesses. When $\gamma = 2, 3$, or 4, a workable radius value range can be found. Generally speaking, the range is larger when the stability is higher, and it is smaller when the stability is lower. A very small radius value (i.e., 0.1) will lead to a larger FNR value. And a larger radius value will lead to a larger FPR value. Therefore, if FNR is more concerned, a too small radius value should not be adopted. If FPR is more concerned, then a small or medium radius value can be applied. As we have mentioned in Section 4, we currently do not focus on the problem of collecting testimonies. But here we suggest that when collecting testimonies and estimating the seller's reputation, the witnesses who have a low stability should not be considered as their testimonies may not



$I = 10,\ \omega = 10$

$I = 100,\ \omega = 10$

$I = 10,\ \omega = 100$

$I = 100,\ \omega = 100$

FIGURE 5. Varying radius when $P_{\text{unfair}} = 80\%$.

reflect their behaviors or the seller's behavior, which will lead to an incorrect filtering result especially when the scalability is high. By summarizing the simulation results, the workable radius value more falls into the range of $[0.3, 0.5]$. As we are more concerned with the FPR value, we use 0.3 as the DBSCAN radius value in our following simulations. The δ value is randomly selected (either 0.2 or 0.3) in each round of our other simulations.

To investigate how to properly set ε value which is used to trigger the Global component, we simulate three scenarios. In the first scenario, the frequency of the witnesses having transactions with S is lower than that of B. In the second scenario, the frequency of the witnesses having transactions with S is the same as that of B. In the third scenario, the frequency of the witnesses having transactions with S is higher than that of B. Here the frequency represents how often a witness or B will have one transaction with S. If we use B's frequency as the baseline, the equal frequency scenario means that when B has one transaction with S, each witness also has one transaction with S. The lower frequency scenario means that when each witness has one transaction with S, B already has multiple transactions with S. The higher frequency means that when B has one transaction with S, each witness already has multiple transactions with S.

Figure 6 shows the changes of MCC values with transaction number increasing in 0.5 frequency, equal frequency, and double frequency scenarios, respectively, where 0.5 frequency means that when B has two transactions with S, a witness has only one transaction with S from the view of probability. According to the results, in the 0.5 frequency scenario, the Local component achieves a stable filtering result after B having about 25 transactions. In the equal frequency scenario, the Local component achieves a stable filtering result after B having about 12 transactions. In the double frequency scenario, the Local component achieves

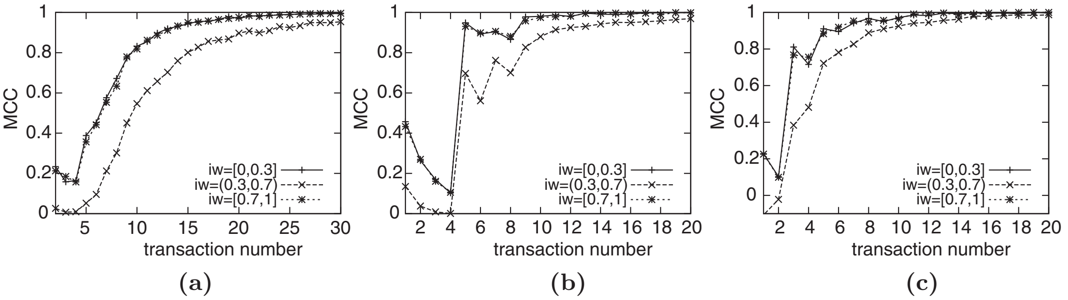


FIGURE 6. (a) Frequency of buyer interaction is double as that of witnesses; (b) Equal frequency of interacting with seller; (c) Frequency of witness interaction is double as that of buyer.

a stable filtering result after B having about eight transactions. These results suggest that the setting of ε value should be related to the relative frequency of the witnesses and B having transactions with the seller. It is a good practice for B to evaluate the frequency of the transactions between the witnesses and the seller before he sets the ε value. If the witnesses' frequency is higher than or equal to that of the buyer, then a small ε value can be adopted. When the witnesses' frequency is lower than that of the buyer, a larger ε value is more preferred. In our following experiments, we use $\varepsilon = 10$ because the equal frequency scenario is simulated.

6.2. Robustness against Collusion Attacks

The goal of this set of experiments is to investigate the accuracy of the iCLUB approach in filtering unfair testimonies. In particular, we investigate the robustness of our approach against collusion attacks (i.e., sellers collude with some buyers who give unfair testimonies for the colluding sellers). As demonstrated in the first set of experiments, the Local component of the iCLUB approach can work well when the buyer agent has some transactions with the target seller agent. But it is obvious that when the buyer has no sufficient number of transactions with the seller, we need the Global component to facilitate the filtering of unfair testimonies.

In this experiment, we simulate a more complicated trading community that involves 10 seller agents, 100 witnesses and 1 buyer agent B. Each seller agent is attached with a profile, describing his initial willingness (iw) value range and the percentage of dishonest witnesses. The first 200 transactions of each witness or B are for the presetting stage. In this stage, the witnesses will randomly select one seller agent among the 10 seller agents as the partner for each transaction, and B will randomly select one seller agent among the first 9 as the partner for each transaction, leaving the last seller agent alone to investigate the accuracy of iCLUB (the Global component). After the presetting stage, B will randomly select one seller agent among the 10 seller agents as partner for each transaction. We simulate two kinds of scenarios: the witnesses' behaviors keep consistent for all the sellers, and the witnesses' behaviors change from one seller to another seller.

When the witnesses' behaviors keep consistent, the dishonest witnesses will report ratings unfairly for all the sellers and the honest witness will report ratings for all the sellers in an honest way. In each round of this simulation, the percentage of dishonest witnesses increases from 10% to 90% for each seller. The sellers' iw values are randomly generated and the percentage of each type of dishonest witnesses is also randomly generated according to the seller's iw value (the sum of the percentage of each type of dishonest

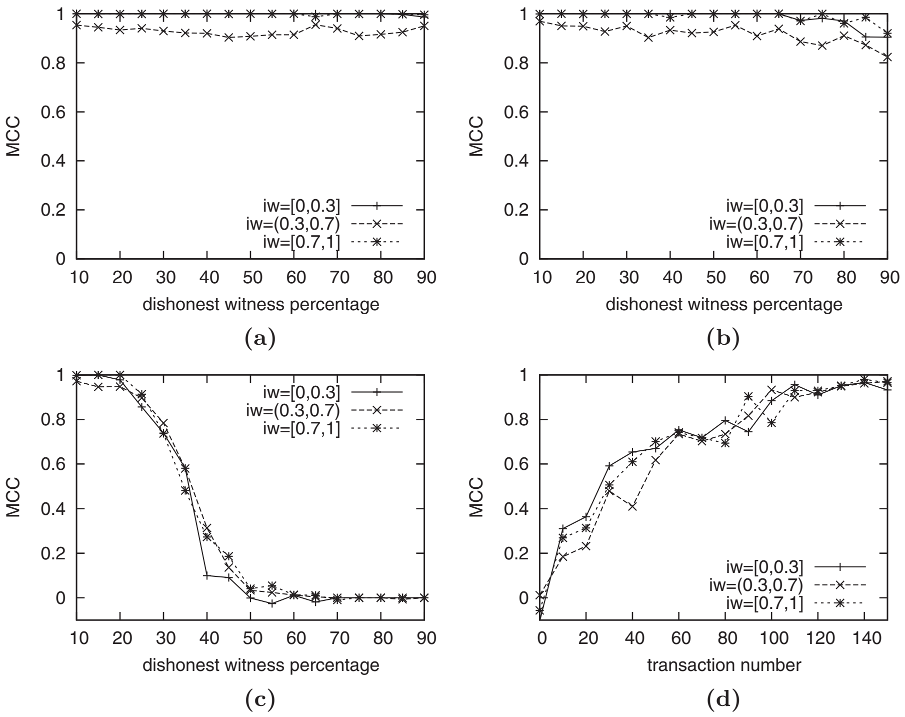


FIGURE 7. (a) The witnesses' behaviors keep consistent for all sellers; (b) The honest witnesses' behaviors keep consistent for all sellers; (c) The witnesses' behaviors change over sellers; (d) MCC varying with transaction number when $P_{\text{unfair}} = 60\%$.

TABLE 6. Profiles of Seller Agents.

Seller index	S_1	S_2	S_3	S_4	S_5
iw	[0, 0.3]	(0.3, 0.7)	[0.7, 1]	[0, 0.3]	[0, 0.3]
P_{unfair}	0	0	0	20%	60%
Seller index	S_6	S_7	S_8	S_9	S_{10}
iw	[0.7, 1]	[0.7, 1]	(0.3, 0.7)	(0.3, 0.7)	Varying
P_{unfair}	20%	60%	30%	90%	Varying

witnesses should be equal to P_{unfair}). Figure 7(a) shows the tenth seller's MCC value changes with the percentage of dishonest witnesses increasing from 10% to 90%. It can be seen that the Global component can work well no matter what the percentage of the dishonest witnesses is.

We then investigate the accuracy of the Global component working when the witnesses' behaviors change when collusion attack exists. First, we simulate a specific scenario—some witnesses keep honest for all the sellers. To simulate this scenario, we set the profiles of the 10 sellers as shown in Table 6.

S_1, S_2, and S_3 are used to simulate the seller agents who have no dishonest witnesses. S_4 and S_5 are used to simulate the seller agents who have ballot-stuffing witnesses and

γ-high-shifting witnesses. They represent two scenarios—one is that the percentage of dishonest witnesses is smaller than that of honest witnesses, and the other one is that the percentage of dishonest witnesses is larger than that of honest witnesses. S_6 and S_7 are used to simulate the seller agents who have badmouthing witnesses and γ-low-shifting witnesses. They also represent the two scenarios as what S_4 and S_5 do. S_8 and S_9 are used to simulate the sellers who have four types of dishonest witnesses at the same time. They also represent the two scenarios as what S_4 and S_5 do. S_{10}'s initial willingness and the percentage of dishonest witnesses will change in the simulation for the purpose of investigating the accuracy of the iCLUB approach in varying scenarios where, for example, the seller agent may change his behavior over time. For each seller, the first P_{unfair} percentage of witnesses are generated as dishonest witnesses, and other witnesses are generated as honest witnesses. Through this setting, only the last 10% witnesses are honest for all sellers. Though 61% to 90% witnesses are honest for the first eight sellers, they are detected as dishonest for seller S_9. If they collude again for seller S_{10}, they still can be detected as dishonest.

Figure 7(b) shows the tenth seller's MCC changes with his dishonest witness percentage increasing from 10% to 90%. According to the results, when a certain number of witnesses keep honest for all the sellers, the Global component can work well. The reason behind is as follows: As described in Algorithm 2, the Global component takes the intersection of all the honest witnesses for all the sellers encountered by B as W_F. If a witness keeps honest for all the sellers, then he must be in W_F (assuming that the clustering result is correct). Then for the target seller S_t, suppose there are five clusters after clustering—C_1 including the badmouthing witnesses' testimonies, C_2 including the ballot-stuffing witnesses' testimonies, C_3 including the γ-low-shifting witnesses' testimonies, C_4 including the γ-high-shifting witnesses' testimonies, and C_5 including the honest witnesses' testimonies. As the witnesses in W_F are honest for all the sellers, their testimonies should not be in C_1, C_2, C_3, or C_4. Then C_5 will have the largest intersection result with W_F. Therefore, the Global component can get the correct filtering result.

Keeping the second scenario of the witnesses changing behaviors, we remove the assumption that some witnesses keep honest for all the sellers by simulating the scenario where the witnesses' behaviors change randomly. In this simulation, the 10 sellers' iw value ranges are kept the same as in last simulation. The percentage of the unfair witnesses is the same for the 10 sellers. For each seller, the dishonest witnesses are randomly assigned. Figure 7(c) shows the tenth seller's MCC value changes with P_{unfair} increasing from 10% to 90%. According to the results, the Global component can work well when P_{unfair} is small (i.e., $< 30\%$). When P_{unfair} is larger than 50%, the filtering result is near to random guessing.

Though the Global component cannot work when P_{unfair} is larger than 30% and the witnesses' behaviors change randomly from one seller to another, the collusion attack problem can still be solved if buyer B can sacrifice some transactions with the target seller agent. Figure 7(d) shows the changes of MCC value for S_{10} with the transaction number increasing after the presetting stage when we set ε=10. The transaction number starts from 0 and ends at 150. Though for a particular transaction, the buyer may not select S_{10} as his partner, we still calculate the MCC value according to the iCLUB result for S_{10}. The percentage of dishonest witnesses is set as 60%. It can be noticed that after about 120 transactions, the MCC value will approximate to 1. Since a seller agent is randomly selected among 10 sellers as the partner for each transaction, the buyer may only sacrifice 12 transactions to accumulate the experiences for the iCLUB approach to cope with the situation where 60% of witnesses randomly collude. Actually, this number is quite affirmative with the ε value to trigger the global component we get in the experiment of exploring the ε value for the equal frequency scenario in Section 6.1.

6.3. Integration with Dirichlet Reputation System

The iCLUB approach is an unfair testimonies filtering approach instead of a reputation system. But the iCLUB approach can be integrated with reputation systems where buyer agents' rating vectors are shared. In this part, we use the Dirichlet reputation system (Jøsang and Haller 2007) as an example to show that the iCLUB approach can contribute to improving the reputation system's robustness.

The Dirichlet reputation system (DRS) works as follows: Let the rating vector from buyer B_n regarding seller S_m be $R_{S_m}^{B_n,t}$ in time period t. Suppose that the buyer cares more about the seller's recent behavior and forgets his old behavior, which can be achieved by introducing a forgetting factor λ. Then after time period T, the aggregated rating vector $A_{S_m}^{T}$ regarding S_m from time period 1 to T is:

$$A_{S_m}^{T} = \sum_{t=1}^{T} \sum_{n=1}^{N} \lambda^{T-t} R_{S_m}^{B_n,t}.$$

DRS assumes that a seller's behavior follows a Dirichlet probability distribution (Gelman 2004). When mapping to K rating levels, the expected probability p_k that S_m will behave as rating level k $(1 \leq k \leq K)$ in the future is:

$$E(p_k) = \frac{A_{S_m}^{T}(k) + \dfrac{C}{K}}{C + \sum_{k=1}^{K} A_{S_m}^{T}(k)},$$

where C is *a priori* constant (Gelman 2004) which will be always equal to the cardinality of the state space over which a uniform distribution is assumed (e.g., the constant $C = 2$ emerges when a uniform distribution over a binary state space is assumed). As a further step, we calculate S_m's reputation E_{S_m} as the following:

$$E_{S_m} = \sum_{k=1}^{K} E(p_k) \times K.$$

We simulate a similar trading community as in Section 6.2, which includes 1 buyer agent, 10 seller agent, and 100 witnesses. Ten time windows are simulated. In each time window, the sellers' behaviors do not change. After each time window, the sellers' behaviors will change. The sellers' behavior changing is simulated by generating different iw value ranges. For example, if in the first time window, seller S_1's iw value falls into the range of [0,0.3], then in the second time window, his iw value will fall into the range of (0.3,0.7) or [0.7,1]. In each time window, the number of the transactions from each witness or the buyer is a randomly generated number which falls into the range of $[100, 200]$. The unfair percentage for each seller is 0 for S_1, 10% for $S_2, \ldots,$ and 90% for S_{10}. The percentage of each type of dishonest witnesses for each seller is randomly generated. In each time window, iCLUB is applied to get the honest witnesses, and only the achieved honest witnesses' ratings are passed to DRS to calculate the sellers' reputation. $C = 5$ is used as DRS *a priori* constant. We measure the accuracy of estimating the sellers' reputation using the mean absolute error (MAE) as follows:

$$\mathrm{MAE} = \frac{\sum_{m=1}^{M} |E_{S_m} - \hat{E}_{S_m}|}{M},$$

where M is the number of sellers, E_{S_m} is the seller S_m's expected reputation which is calculated using the honest witnesses' ratings, and $\hat{E}_{S_m}$ is S_m's reputation using the achieved

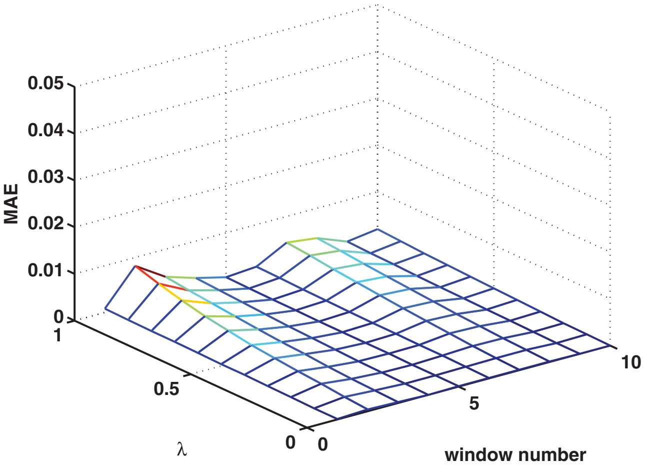


FIGURE 8. MAE changes with forgetting factor and window number.

honest witnesses' ratings after using iCLUB filtering. Figure 8 shows the MAE changes with forgetting factor λ and window number. The result shows that MAE is quite small after applying the iCLUB filtering approach. It implies that the estimated sellers' reputation after using the iCLUB approach is very close to the expected seller reputation.

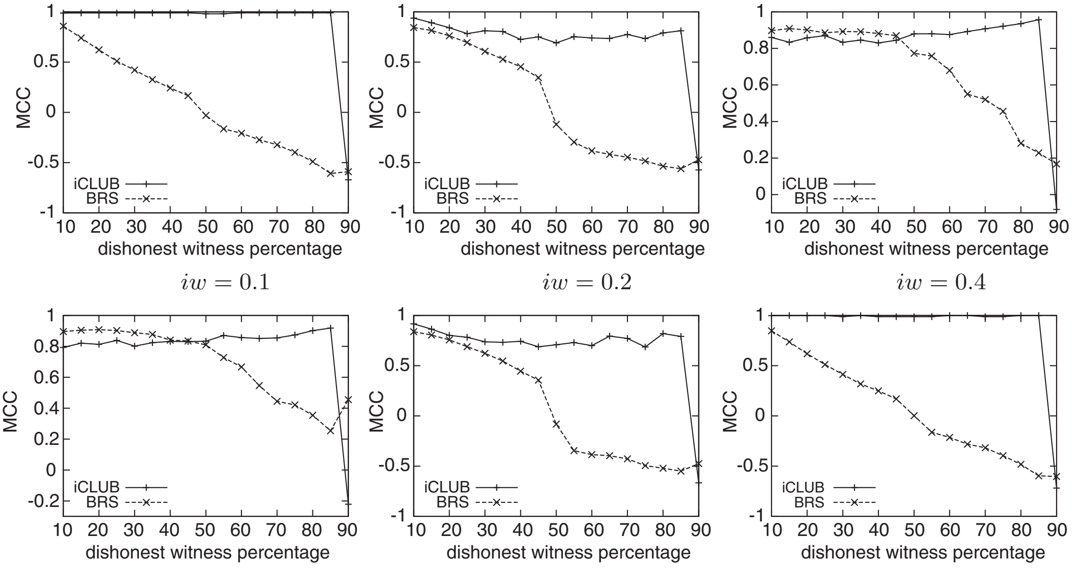
6.4. Comparative Experiments

We compare the iCLUB approach with other approaches from two aspects—filtering accuracy and seller reputation estimation. As pointed out in Section 2, most of the existing approaches for filtering unfair testimonies are designed for reputation systems accepting only binary rating levels. These binary filtering approaches cannot be directly used for multinominal rating levels, but our iCLUB approach can actually be easily adopted for the binary rating level case. In this experiment, we compare our approach with the BRS approach in terms of filtering accuracy. We also compare the accuracy of estimating seller reputation with both BRS and TRAVOS, the two representative probabilistic approaches that are different from clustering-based approaches. Note that in the comparative experiments, after filtering unfair testimonies by different approaches, fair testimonies will be aggregated to estimate seller reputation in a simple way as used in BRS (see Section 2 for more details).

A similar trading community as that used for the second set of experiments is simulated. The difference is the way to generate a rating for each transaction as there are only two rating levels in this experiment. The initial willingness (iw) value assigned to each seller agent is taken from the value set {0.1, 0.2, 0.4, 0.6, 0.8, 0.9}. To simulate seller behavior changes, a seller's first transaction uses the initial willingness value. In the following transactions, a willingness value is generated through one of the three strategies—the willingness value of last transaction subtracting 0.02, equal to the willingness value of last transaction, or the willingness value of last transaction adding 0.02. The three strategies are uniformly selected during the simulation. The willingness value for each transaction is also limited in the range of $[iw-0.1, iw+0.1]$. Compared to the experiments for multinominal ratings, there are no γ-low-shifting or γ-high-shifting witnesses. For the binary ratings case, we study three types of dishonest witness: (1) ballot-stuffing witnesses; (2) badmouthing witnesses; (3) opposite witnesses who report ratings as the opposition of the real ratings. We assume that a seller

TABLE 7. Profiles of Seller Agents.

Seller index	S_1	S_2	S_3	S_4	S_5
iw	0.1	0.4	0.8	0.1	0.2
P_{unfair}	0	0	0	40%	80%
Seller index	S_6	S_7	S_8	S_9	S_{10}
iw	0.8	0.9	0.4	0.6	Varying
P_{unfair}	40%	80%	40%	80%	Varying



$iw = 0.1$ $iw = 0.2$ $iw = 0.4$

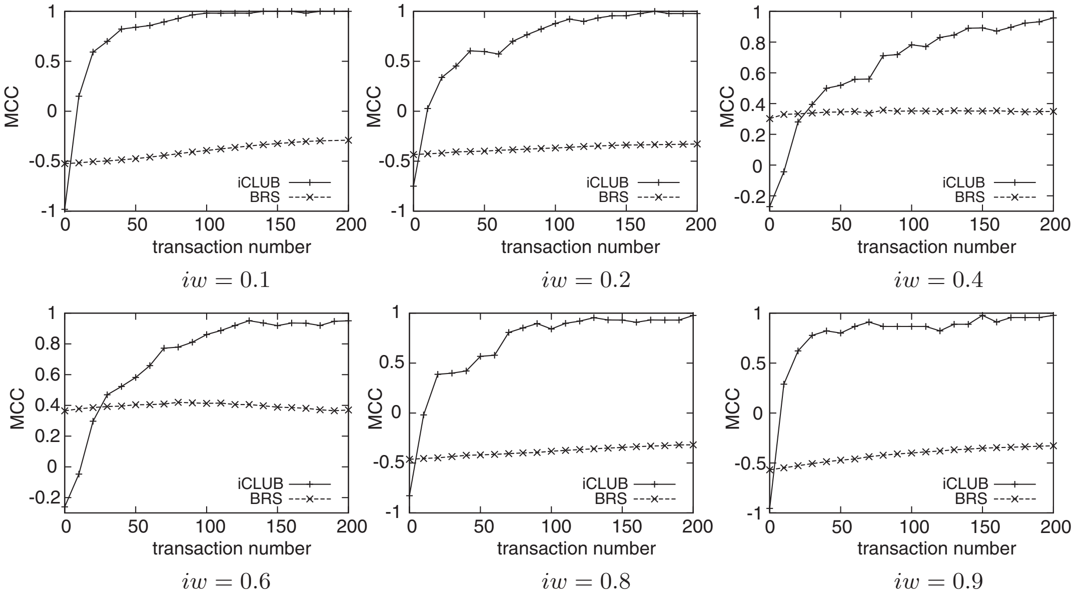
$iw = 0.6$ $iw = 0.8$ $iw = 0.9$

FIGURE 9. MCC changes with dishonest witnesses percentage increasing.

agent with a larger iw value (i.e., 0.8 and 0.9) will not have ballot-stuffing witnesses, a seller agent with a smaller iw value (i.e., 0.1 and 0.2) will not have badmouthing witnesses and a seller agent with a medium iw value (i.e., 0.4 and 0.6) will not have opposite witnesses. Table 7 shows the profiles of the 10 sellers.

We keep the first 200 transactions as the presetting stage. Each witness will randomly select one seller agent among the 10 sellers as his partner for each transaction, and the buyer agent B will randomly select one seller among the first 9 as his partner. Figure 9 shows the changes of MCC value for S_{10} for iCLUB and BRS with the percentage of dishonest witnesses increasing form 10% to 90% after the presetting stage. It can be noticed that the performance of BRS filtering approach decreases with the increase of the percentage of dishonest witnesses. The iCLUB approach performs stably until a significant percentage (more than 85%) of witnesses are dishonest.

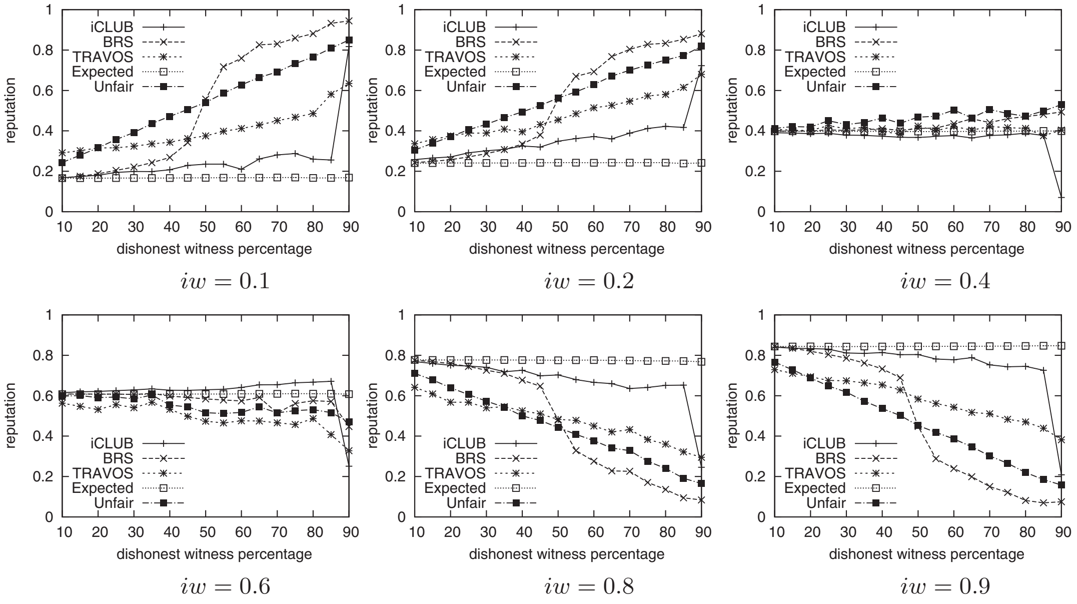
Figure 10 shows the changes of MCC value for S_{10} with the increase of transaction number after the presetting stage when the percentage of dishonest witness is 90%. We set $\varepsilon = 10$. It can be noticed that after about 100−150 transactions, which is about 10−15 transactions between B and S_{10}, the MCC value using iCLUB is approximately 1.



$iw = 0.1$ $iw = 0.2$ $iw = 0.4$

$iw = 0.6$ $iw = 0.8$ $iw = 0.9$

FIGURE 10. MCC changes with transaction number increasing.

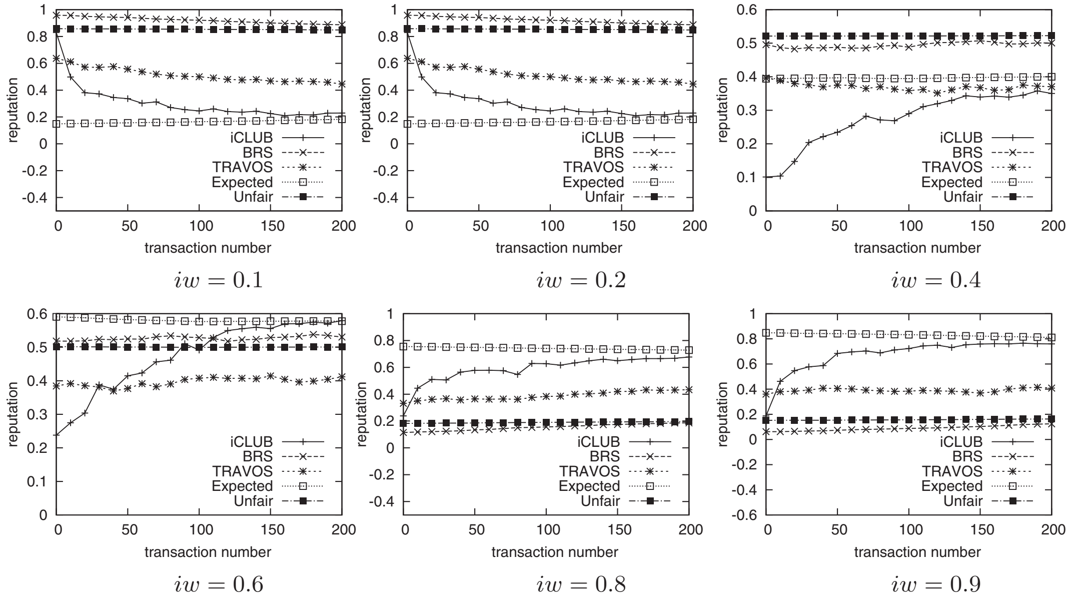


$iw = 0.1$ $iw = 0.2$ $iw = 0.4$

$iw = 0.6$ $iw = 0.8$ $iw = 0.9$

FIGURE 11. Reputation estimation value changes with dishonest witnesses percentage increasing.

Figure 11 shows the comparison result of reputation estimation for S_{10} when the percentage of dishonest witnesses increases by using BRS, iCLUB, and TRVOS. The expected reputation is calculated using only honest witnesses' testimonies. And the unfair reputation is calculated using all witnesses' testimonies. When $iw = 0.1$, $iw = 0.2$, $iw = 0.8$, and $iw = 0.9$, the reputation value after using iCLUB is initially close to the expected reputation. With the increase of the dishonest witness percentage, the reputation value after



$iw = 0.1$ $iw = 0.2$ $iw = 0.4$

$iw = 0.6$ $iw = 0.8$ $iw = 0.9$

FIGURE 12. Reputation estimation value changes with transaction number increasing.

using iCLUB has a small deviation from the expected reputation until a significant percentage of dishonest witnesses ($> 85\%$) arrives. The reputation value after using BRS filtering approach is initially close to the expected reputation, and then continuously deviates from the expected reputation. After the percentage of dishonest witnesses is greater than 50%, the reputation value after using BRS filtering approach is even worse than the unfair reputation. The reputation value after using TRAVOS is initially worse than that using BRS or iCLUB, and when the percentage of dishonest witnesses is greater than 40%, it is better than BRS, but still worse than iCLUB. When $iw = 0.4$ and $iw = 0.6$, the expected reputation value, the unfair reputation value and the reputation value after using BRS, TRAVOS, or iCLUB are quite close since there exist badmouthing witnesses and ballot-stuffing witnesses at the same time. The impacts of the two types of dishonest witnesses counteract each other.

Figure 12 shows the comparison result for S_{10}'s reputation estimation with the transaction number increasing after the presetting stage when the percentage of dishonest witnesses is 90%. We set ε=10. We can see that after about 10–15 transactions (100–150 on the x-axis in the figure), the reputation value estimated using iCLUB is very close to the expected reputation, indicating that iCLUB achieves more accurate filtering result than BRS and TRAVOS.

6.5. Discussions on Experimental Results

According to the experimental results, our iCLUB approach shows promising filtering accuracy when being applied to both reputation systems with multinominal rating levels and those with binary rating levels. To make iCLUB work more effectively, the DBSCAN radius value is very important. Though we have established a feasible DBSCAN radius value range for iCLUB through experimentation, it is worth pointing out that this value range is dependent on the simulation data. We argue that a smaller radius value may be better though it may cause more false negatives but can decrease the number of false positives.

As our aim is to filter unfair testimonies, false positives should carry more weight. Another suggestion is that it is better not to consider the testimonies from the witnesses who have few transactions with the seller when collecting testimonies. As such testimonies may not reflect the witnesses' behaviors or the seller's behavior, these unstable testimonies may cause a wrong clustering result.

The experimental results in Section 6.1 suggests that there is a relationship between the ε value setting and the frequency of buyer agents having transactions with seller agents. But in the scenario where a large number of witnesses collude randomly, ε should also be set smaller to disable the Global component at an early stage as the Global component cannot address this extreme situation. According to the experimental results, setting $\varepsilon = 10$ is a good choice in many cases though iCLUB has not achieved its best performance in the low frequency scenario (as shown in Figure 6(a)).

The accuracy of the Global component is dependent on two factors—the Local component working accuracy and the witnesses' behaviors. The Local component working accuracy can be controlled by adopting an appropriate clustering parameter value. But the witnesses' behaviors are not controlled by our approach. Therefore, in an environment where the witnesses' behaviors keep consistent for the sellers or at least the honest witnesses' behaviors keep consistent for the sellers, the Global component can work accurately. But in an environment where the witnesses' behaviors keep changing over the sellers, the buyer has to sacrifice some transactions with the target seller to get an accurate filtering result. The accuracy of the Global component is also dependent on the sellers from whom the global information is achieved. If it happens that the Global component only gets the testimonies for the sellers to whom the witnesses are honest and does not get the testimonies for the sellers to whom the witnesses are dishonest, then a wrong filtering result may also be led to.

In the comparative experiments, that the iCLUB approach is better than the BRS filtering approach is due to two reasons. First, iCLUB uses global information to filter the unfair testimonies. Second, iCLUB assigns the buyer's personal ratings with a higher weight (the Local component will keep as fair the testimonies from the witnesses whose rating vectors are in the same cluster as the buyer's rating vector). The TRAVOS approach also uses global information and assigns the buyer's personal ratings with a higher weight, but TRAVOS needs more transactions to achieve a high accuracy in reputation estimation compared to the iCLUB approach.

7. CONCLUSIONS AND FUTURE WORK

Reputation systems have contributed much to the success of online trading communities. But its robustness easily deteriorates due to the existence of unfair testimonies. To address the problem of unfair testimonies for reputation systems with multinominal rating levels, we proposed the iCLUB approach for filtering unfair testimonies to improve the robustness of reputation systems. iCLUB supports reputation systems with multinominal rating levels, which is a major limitation of the existing filtering approaches. iCLUB uses local and global information to filter unfair testimonies by adopting the clustering technique. Experimental results confirm that iCLUB is effective in filtering unfair testimonies and is able to cope with collusion attacks to a great extent. iCLUB also outperforms the competing approaches (BRS and TRAVOS) in the scenario where only binary ratings are supported.

For future work, we will first investigate an automatic and dynamic way to decide the DBSCAN parameter value. Currently the DBSCAN parameter value is experimentally decided. Though we suggest some heuristics to choose the parameter value, it may still cause some filtering errors. More experiments are needed to explore the accuracy of the iCLUB filtering when more sophisticated types of dishonest witnesses exist. As what we have pointed out, the Global component working accuracy is dependent on the witnesses' behaviors. Therefore, how to collect testimonies is a direction for future work. Currently, we assume that the buyer can get the testimonies. How to get useful and plentiful testimonies is still a problem, especially in a decentralized reputation system. Another direction of our future work is to model witnesses' behaviors as what Noorian et al. (2011) did. Currently, after iCLUB filter the unfair testimonies, we simply pass these honest witnesses' testimonies to a reputation system. But if we can further model the honest witnesses' behaviors and adjust their testimonies accordingly, the reputation system's performance may be improved.

Besides the future directions mentioned above, how to apply the proposed iCLUB approach into real reputation systems is also a future direction. Though we have presented how to integrate iCLUB with the Dirichlet reputation system in Section 6.3, some factors need to be considered when applying iCLUB into real reputation systems. First, context information can be considered when applying iCLUB to a realistic reputation system. For example, when a new buyer enters the system, he can only depend on the Global component of iCLUB to identify possible honest witnesses. But as what we have pointed out in Section 6.2, the Global component may fail in some scenarios. Therefore, if the buyer can consider other information to filter unfair testimonies, the accuracy of iCLUB will be improved. Such possible information can be the relationship between the buyer and the witnesses, the relationship between the witnesses and the seller, the amount of money involved in the transactions and the frequency of the seller receiving ratings (e.g., that good ratings suddenly come in a short period may imply the existence of dishonest witnesses). If we can consider these kinds of context information when applying iCLUB to real reputation systems, its accuracy in filtering unfair testimonies will be improved. Second, we currently do not differentiate ratings of subjective difference from unfair testimonies intentionally reported by malicious witnesses. However, it is possible that a buyer can tolerate some subtle subjective difference to tune his reputation evaluation in a real reputation system. Therefore, if we can differentiate the witnesses with subjective difference from the malicious witnesses, iCLUB will provide more flexibility when being applied to real reputation systems. This flexibility can be achieved by adjusting the DBSCAN clustering parameter. We can use a larger DBSCAN clustering parameter to include the ratings of subjective difference in the achieved cluster. But how to adjust the parameter needs a careful consideration to avoid the inclusion of malicious ratings.

REFERENCES

ANKERST, M., M. M. BREUNIG, H.-P. KRIEGEAL, and J. SANDER. 1999. Optics: Ordering points to indentify the clustering structure. ACM SIGMOD Record, **28**(2):49–60.

DELLAROCAS, C. 2000. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. *In* Proceedings of 2nd ACM Conference on Electronic Commerce, pp. 150–157.

DUBA, R. O., P. E. HART, and D. G. STORK. 2001. Pattern Classification. Wiley Interscience: New York.

ESTER, M., H.-P. KRIEGEL, J. SANDER, and X. XU. 1996. A density-based algorithm for discoverting clusters in large spatial databases with noise. *In* Proceedings of 2nd Interactional Conference on Knowledge Discovery and Data Mining.

FUNG, C. J., J. ZHANG, I. AIB, and R. BOUTABA. 2011. Dirichlet-based trust management for effective collaborative intrusion detection networks. IEEE Transactions on Network and Service Management, **8**(2):79–91.

GELMAN, A. 2004. Bayesian Data Analysis. Chapman & Hall/CRC: Boca Raton, FL.

JØSANG, A., and J. HALLER. 2007. Dirichlet reputation systems. *In* Second International Conference on Availability, Reliability and Security (ARES'07), pp. 112–119.

JØSANG, A., and R. ISMAIL. 2002. The beta reputation system. *In* Proceedings of the Fifteenth Bled Electronic Commerce Conference, pp. 324–337.

JØSANG, A., R. ISMAIL, and C. BOYD. 2007. A survey of trust and reputation systems for online service provision. Decision Support System, **43**(2):618–644

LIU, S., C. MIAO, Y. L. THENG, and A. C. KOT. 2010. A clustering approach to filtering unfair testimonies for reputation systems (extended abstract). *In* Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems, pp. 1577–1578.

LIU, S., J. ZHANG, C. MIAO, Y.-L. THENG, and A. C. KOT. 2011. iCLUB: An integrated clustering-based approach to improve the robustness of reputation systems (extended abstract). *In* Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems, pp. 1151–1152.

MATTHEWS, B. 1975. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. Biochimica et Biophysica Acta, **405**:442–451.

NOORIAN, Z., S. MARSH, and M. FLEMING. 2011. Multi-layer cognitive filtering by behavioral modeling. *In* Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems, pp. 871–878.

RASMUSSON, L., and S. JANSSEN. 1996. Simulated social control for secure internet commerce. *In* Proceedings of the 1996 New Security Paradigms Workshop.

REGAN, K., P. POUPART, and R. COHEN. 2006. Bayesian reputation modeling in e-marketplaces sensitive to subjectivity, deception and change. *In* Proceedings of the 21st National Conference on Artificial Intelligence, pp. 206–212.

TEACY, W., J. PATEL, N. R. JENNINGS, and M. LUCK. 2006. TRAVOS: Trust and reputation in the context of inaccurate information sources. Autonomous Agents and Multi-Agent Systems, **12**(2):183–198.

TEACY, W. T. L., N. R. JENNINGS, N. R. ROGERS, and M. LUCK. 2008. A hierarchical bayesian trust model based on reputation and group behaviour. *In* 6th European Workshop on Multi-Agent Systems, pp. 206–212.

WENG, J., C. MIAO, and A. GOH. 2006. An entropy-based approach to protecting rating systems from unfarir testimonies. IEICE Transaction on Information and System, **E89-D**(9):2502–2511.

WHITBY, A., A. JØSANG, and J. INDULSKA. 2005. Filtering out unfair ratings in Bayesian reputation systems. ICFAIN Journal of Management Research, **4**(2):48–64.

ZHANG, J., and R. COHEN. 2008. Evaluating the trustworthiness of advice about selling agents in e-marketplaces: A personalized approach. Electronic Commerce Research and Applications, **7**(3):330–340.