An unsupervised strategy for defending against multifarious reputation attacks

Xin WANG
Network and Information Center, Shandong University of Science and Technology, Qingdao, China
Shu-juan JI*, Yong-quan LIANG,
Key Laboratory for wisdom mine information technology of Shandong Province, Shandong University of Science and Technology, Qingdao, 266590, China
Ho-fung Leung
Department of Computer Science and Engineering, The Chinese University of Hong Kong, China
Dickson K.W. Chiu
Faculty of Education, The University of Hong Kong, China

***Corresponding author email:** jane_ji2003@aliyun.com

Abstract: In electronic markets, malicious sellers often employ reviewers to carry out different types of attacks to improve their own reputations or destroy their opponents' reputations. As such attacks may involve deception, collusion, and complex strategies, maintaining the robustness of reputation evaluation systems remains a challenging problem. From a platform manager's view, no trader can be taken as a trustable benchmark for reference, therefore, accurate filtration of dishonest sellers and fraud reviewers and precise presentation of users' reputations remains a challenging problem. Based on impression theory, this paper presents an unsupervised strategy, which first design a nearest neighbor search algorithm to select some typical *lenient* reviewers and *strict* reviewers. Then, based on these selected reviewers and the behavior expectation theory in impression theory, this paper adopts a classification algorithm that pre-classify sellers into *honest* and *dishonest* ones. Thirdly, another classification algorithm is designed to classify reviewers (i.e., buyers) into *honest*, *dishonest*, and *uncertain* ones according to their trading experiences with the pre-classified sellers. Finally, based on the ratings of various reviewers, this paper proposes a formula to estimate seller reputations. We further designed two general sets of experiments over simulated data and real data to evaluate our scheme, which demonstrate that our unsupervised scheme outperforms benchmark strategies in accurately estimating seller reputations. In particular, this strategy can robustly defend against various common attacks and unknown attacks.

Keywords: Reputation attack; Nearest neighbor search; Lenient reviewer; Strict reviewer; Behavior expectation theory

1 Introduction

Trust and reputation of an entity is an opinion of that entity based on what has happened in the past, typically evaluated based on a set of social criteria (Longman Dictionary of Contemporary English; Chiu et al., 2009). In real life, reputation is a ubiquitous and basic measurement of social order facilitating distributed social control behavior. As such, in multi-agent systems that are open, large, and dynamic, reputation evaluation plays a vital role against deceptive and strategic self-interested agents. For example, in electronic markets, dishonest seller agents often commission deceptive reviewers to enter unfairly high ratings to boost their reputations (Dellarocas, 2000), which may result in buyers' perception of unsatisfied quality of the delivered products. As a result, buyers may lose their trust and feel risky in trading with such sellers subsequently after some unsatisfactory transactions. For better estimating sellers' reputations and supporting honest buyers in choosing trustable sellers, reputation systems should model reputations of sellers more accurately and scrutinizing the ratings and reviews shared by buyers, as dishonest reviewers are often hired to give fraud and unfair ratings to mislead buyers into further deceptive transactions (Jøsang, 2012; Zhang et al., 2012). Such attacks typically include pure attacks like *Camouflage*, *AlwaysUnfair*, *Whitewashing*, *Sybil*, as well as combined attacks like *Sybil_Camouflage* and *Sybil_Whitewashing* (Jiang et al., 2013). For example, even if Yelp attempts to filter suspicious reviews with some authentication algorithms, approximately 16% of their restaurant reviews are still unfair (Luca and Zervas, 2013).

Although existing reputation systems (such as Taobao, Dianping, and Yelp) has adopted various strategies to filter out fraud ratings and reviews, these companies are not willing to publish their defensing strategies or even share the desensitized data. That may be caused by two reasons. First, as there are no quality supervision institutions for products in these electronic commerce platforms, no trustable sellers and reviewers (i.e., buyers) can serve as benchmarks for reference. Therefore, it is still a challenge to filter out accurately dishonest sellers and fraud reviewers to ensure the robustness of reputation systems. Second, once the defending strategies are published, the reputation systems and the defending strategies will be exposed to more attacks. However, researchers in the academia, such as Mukherjee et al. (2013) and Rayana & Akoglu (2015), never give up designing methods that are more accurate to exclude fraud reviews and ratings. To estimate sellers' reputation more accurately and improve the robustness of reputation systems, we propose an unsupervised method called *impression-based strategy* (*IBS*).

Comparing with existing defense approaches, the novelty of our approach are as follows. Firstly, this paper introduces two concepts (i.e., *lenient* and *strict*) from the impression theory into the analysis of reviewers' behavior characteristics, and takes it as the key criteria for selecting centroids of nearest neighbor search algorithm. Secondly, the nearest neighbor search algorithm outperforms traditional clustering algorithm (Liu et al., 2014) because of two reasons. One is that the algorithm in this paper only clusters two kinds of reviewers (i.e., *lenient* reviewers and *strict* reviewers) while disregarding other kinds of reviewers not useful in our evaluation strategy, thereby decreasing the overall time complexity. The other is that according to the natural assumption *lenient* reviewers and *strict* reviewers being relatively rare in complex electronic environment, we adopt a parameter *IC* for controlling their numbers, which ensures the convergence of the algorithm. Thirdly, based on the impression theory, this paper takes an assumption that "once a reviewer is classified as a *strict* or *lenient* one, the reviewer is expected to remain in the same category in the near future." Thus, such intuition provides two rules to classifying *honest* and *dishonest* sellers. Experimental results show that these rules can improve the unsupervised filtering approach in accurately estimating sellers' reputations and robustly defending against various common attacks.

To develop this paper, Section 2 reviews the literature, and then we formalize the concepts, assumptions, and rules defined in Section 3. Section 4 details our main idea and algorithms based on impression-based theory. Section 5 illustrates the performances of our strategy through two general sets of experiments. Finally, we conclude this paper with our continuing research plans.

2 Related work

In centralized reputation systems, seller's reputation is usually estimated according to reviewers' ratings. As the central mechanism does not know the exact quality of the sellers' products, it is difficult to discriminate the honesty of reviewers and then accurately estimate a reputation value for the sellers under various attacks (Jøsang et al., 2007). To solve this problem, many filtering approaches have been designed and adopted (even though most of them are trade secrets of electronic commerce companies). There are two general types of existing approaches, namely, data mining and multi-agent methods. The former methods focus on analyzing real data for training an accurate classifier. The latter methods concentrate on generating simulation data and testing the performance of strategies in some extreme environments. These data mining methods first obtain the characteristics of reviewers' linguistic, behaviors, and social relationship. Then they design supervised, unsupervised, or semi-supervised machine learning methods such as Mukherjee's strategy (2013), SpEagle (Rayana, 2015), or SpEagle$^+$ (Rayana, 2015), to classify whether a review is true or false and whether a reviewer is *honest* or *dishonest*. In contrast, our *IBS* strategy focus on filtering out fake reviewers for better evaluating sellers' reputations.

Existing multi-agent approaches can be divided into three categories, namely aggregation methods, filtering methods, and incentive methods. These methods are briefly reviewed as follows.

(1) *Aggregation methods:* This kind of methods are widely used by companies such as eBay, Amazon, and so on since the emergence of electronic markets. Though aggregation methods are playing effective roles in evaluating sellers, they are vulnerable to various reputation attacks. To improve robustness of these kinds of methods, there are many researches. We can trace back to the Sporas model (Zacharia et al., 2000) that considered reviewers'

reputation in the process of reputation accumulation. In addition, to prevent collusion, when two agent review each other many times, only the most recent rating is considered. The model ignores the possibility of repeated trading between two traders. Furthermore, the model does not take into account the timeliness of ratings. This model is only resilient to *Whitewashing* and *Collusive* attacks, but not *Sybil* or *Camouflage* attacks. To improve the Sporas model, Guo et al. (2009) proposed the E-Sporas model, which considers also the influence of the transaction volume and the number of transactions. At the same time, a penalty factor is introduced into E-Sporas to realize the phenomenon of "*slow rise and fast decline*" regarding reputation. Based on traditional reputation accumulation models, Ji et al. (2017) proposed a model called AARE, which further introduced incentive mechanisms to defend all common attacks in monopoly market. However, whether this model is effective in non-monopoly market or not needs further verification.

(2) *Filtering methods:* These kind of methods is popular in research and industry fields, which aims at filtering out suspicious ratings or reviewers to cut off the propagation of fraud information. Table 1 summarizes the characteristics of related filtering methods. One of the most traditional filtering method is the Beta Reputation System (BRS), which discards reviews with scores out of the majority range of q to 1−q quantile (Whitby et al., 2004). With such a "majority-rule," this approach is vulnerable to *Sybil* attacks, as it incorrectly filters out *honest* reviewers' ratings as the minority.

Liu et al. (2014) proposed an algorithm named *iClub*, which divides reviewers into different clubs using the DBSCAN clustering algorithm. In the clustering process, two components (local vs global) are used for filtering unfair reviews. If the reviewer has adequate transactions with the designated seller, the local component clusters only on a reviewer's private information. Otherwise, the global component makes use of the global information instead. So, *iClub* can largely defend against collusion attacks with effective filtering of unfair ratings, but vulnerable to *Sybil* attacks (Jiang, 2013).

In order to improve the model given in a preliminary study (Wang et al., 2017), this paper modifies the calculation method of seller's reputation aggregation as detailed in section 4.4. In this new method, we introduce a parameter *CF* (Confidence) to replace parameter *HD* (ratio be of *honest* buyers to *dishonest* buyers). In the evaluation of sellers' reputations, this modified method completely ignores the *dishonest* reviewers' ratings as soon as they are identified, so that the *CF* parameter enables a gradual reduction of the uncertainty of reviewers' trustworthiness (*honest* or *dishonest*) over time, thereby increasing the overall reputation evaluation reliability. Besides, Wang et al. (2017) have only verified very briefly the robustness of *IBS* strategy with two sets of simulation experiments. In this paper, we perform four sets of detailed experiments to show that such modification improves the performance of our scheme in a variety of settings. Further, there are three limitations in the experiments of Wang et al. (2017) that we improve significantly in this paper.

(a) Wang et al. (2017) fixed a parameter *IC* (ratio of *lenient* and *strict* reviewers to normal reviewers) of the *IBS* strategy at 0.15, and did not test the performance of this strategy under various values of this parameter. Is the performance of this strategy affected by the value of this *IC* parameter? Which value can maximize the performance of this strategy? To answer these questions, we perform another set of experiments to evaluate our strategy under different values of *IC* in order to discover an optimal *IC* value.

(b) Wang et al. (2017) evaluated the accuracy of the *IBS* strategy using the *MARHS* (mean aggregation reputation of *honest* sellers). However, this criterion can only reflect the predicted reputation of *honest* sellers, while it cannot reflect the degree to which the predicted reputation deviates from the true values. Similarly, *MARDS* (mean aggregation reputation of *dishonest* sellers) can only reflect the true reputations of *dishonest* sellers. Therefore, in this paper we adopt the MAE (Mean Absolute Error) as a criterion to reflect the degree to which the predicted reputation of sellers deviates from their true values, and use Matthews Correlation Coefficient (MCC) as an alternative evaluation criterion for measuring the classification accuracy of our strategy. Besides, to reveal the performance of our strategy in a more comprehensive manner, we compare the run-time of our strategy with that of traditional global-viewed strategies under similar configuration.

(c) The experiments of Wang et al. (2017) were performed over simulation data, in which the simulated attacks were simple and lack of adaptability and intelligence in contrast to human attacks. Moreover, there are many noise data in real transaction ratings and reviews. To further demonstrate real-life practicability of our approach, this paper enriches the experiments by adding a set of experiments over the real-life data from Yelp (http://yelp.com), a typical B2C review website. Therefore, these new and enhanced experiments are essential to further illustrate the practicability of our strategy.

Table 1. Comparison of related filtering methods

Method	Classification fashion	features used	Data set	Objectives
Mukherjee (2013)	Supervised	review text+ reviewers behavior	Yelp data	filter out fake reviews
SpEagle (Rayana, 2015)	Unsupervised	review text+ reviewers' behavior +reviewers' social network	Yelp data	filter out fake reviews filter out fake reviewers
SpEagle^{+} (Rayana, 2015)	Semi-supervised	review text+ reviewers' behavior +reviewers' social network	Yelp data	filter out fake reviews filter out fake reviewers
BRS (Whitby, 2004)	Unsupervised	reviewers' behavior	simulation data	calculate sellers' reputation
iClub (Liu,2014)	Unsupervised	reviewers' behavior +reviewers' social network	simulation data	calculate sellers' reputation
IBS	Unsupervised	reviewers' behavior	simulation+ Yelp data	calculate sellers' reputation

(3) *Incentive methods:* Different from the above methods that focus on evaluating historical trustworthiness of reviewers and sellers, incentive methods focus on setting some mechanisms to decrease their motivation to generate fraud ratings and reviews. For example, Kerr and Cohen (2006) proposed the use of numeric *Trunits* to model trust in electronic markets, which 'flow' during the course of transaction in much the same way like monetary value to serve as incentives. If a seller acts honestly, his Trunit balance increases; otherwise his Trunits decreases. With such an approach, a reviewer need not estimate the trustworthiness of a seller according to individual experience or others' opinions, as all the sellers are incented to be honest. Kerr and Cohen argued that their model is invulnerable to many attacks, but have other problems. For example, granting a new trader with some initial Trunits upon startup may lead to vulnerability of re-entry or whitewashing attacks. Moreover, Trunits is vulnerable to surplus trust (extra Trunits).

To address the above problems, this paper presents a new filtering method under the assumptions that once a reputation evaluation mechanism is devised, there are various kinds of possible reputation attacks in electronic markets. Therefore, our filtering method aims at 'purifying' the ratings and defending against the attacks by filtering out false or unfair ratings from the global or platform-level view. As the platform manager does not have direct trading experience with sellers, all the sellers and reviewers may be suspicious, and trustable one cannot be easily identified as evaluation benchmark for filtering. Therefore, it is a challenging unsupervised learning problem to filter out unfair or false ratings. Similar to BRS, our strategy evaluates the reputation of sellers from a global viewpoint. Different from BRS, our scheme classifies *honest/dishonest* sellers and reviewers based on the *lenient* reviewers and *strict* reviewers with rules derived from the impression theory, instead of based on the majority range between the q and 1−q quantile. As reviewers' impression changes dynamically with transactions and ratings, the chosen reviewers as well as the resultant classification of sellers and reviewers change accordingly.

The main difference between data mining algorithms and our strategy is that the former algorithms (Mukherjee, 2013) are supervised ones and they only focus on improving the accuracy (or precision) and F1 (weighted average of precision and recall) of reviews, while neglecting the estimation of sellers' reputations (trustworthiness). In comparison, not only can our *IBS* strategy categorize reviewers (*honest*, *dishonest*, or *uncertain*) and sellers (*honest* or *dishonest*), but also it can estimate sellers' reputations based on the classification results. The second difference is that the former algorithms require manually processed balance data, i.e., administrators have to preprocess the train and test data set as 50% true and 50% false reviews, while our strategy does not have this requirement. Thirdly, former work such as (Mukherjee, 2013; Rayana,2015) reach high F1 by using lots of linguistic, behavior, and relationship features of reviews and reviewers, while our strategy simply use reviewers' ratings. Further, our strategy can accurately estimate sellers' reputations under various attacks in B2C and B2B markets. Therefore, our strategy is applicable to a wider range of applications and has lower computation complexity.

3 Formalization of concepts

In this section, we describe the concepts and framework used in the *impression-based strategy* (*IBS*) (Wang et al., 2017). Figure 1 describes a framework of our centralized reputation system for electronic markets, in which sellers and buyers interact, transact, and review one another. The system records their actions and reviews for calculating their reputations. The following is an overview of our approach. First, the central management agent selects some typical *lenient* reviewers and *strict* reviewers (Algorithm 1) according to their behavior characteristics. Then, the central agent pre-classifies the sellers traded with the selected *lenient*/*strict* reviewers based on their behavior expectation according to impression theory (Algorithm 2). Thirdly, we classify all the reviewers into *honest*, *dishonest*, and *uncertain* ones (Algorithm 3) according to the pre-classified sellers. Finally, sellers' reputations are calculated by aggregating various reviewers' reputations (Algorithm 4).

The nearer the calculated reputations approach their real ones, the more accurate the centralized reputation system is. Moreover, if the calculated reputations approximate to the real ones very well under multifarious attacks, the centralized reputation system can be deemed as robust. This section first illustrates the concepts used in our framework formally, and then defines the concepts of *lenient* reviewers and *strict* reviewers based on the impression theory of social psychology and statistical metrics for identifying them. Finally, based on the expectation about impressed reviewers' future action in impression theory, two rules are given for classifying sellers traded with *lenient* and *strict* reviewers.

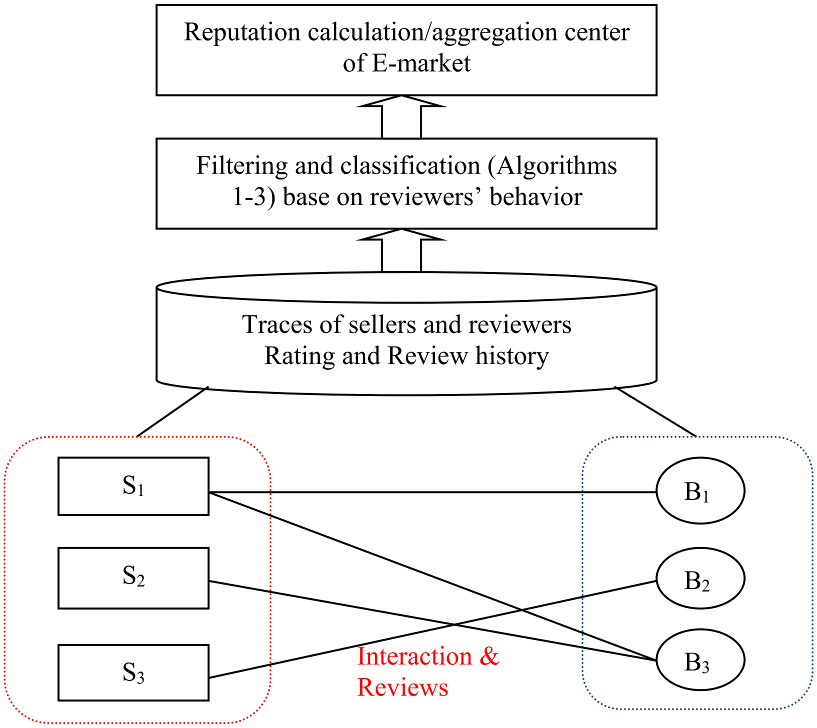

Reputation calculation/aggregation center of E-market
Filtering and classification (Algorithms 1-3) base on reviewers' behavior
Traces of sellers and reviewers Rating and Review history
S1
S2
S3
B1
B2
B3
Interaction & Reviews

Figure 1. A framework of centralized reputation system in electronic market

3.1 Formal concept representation

To model the concepts involved in electronic markets, we denote them by the symbols as shown in Table 2. We assume that: (1) reviewers in the electronic market are willing to give ratings; (2) the quality of services or products that are provided by each seller is stable (not fluctuate frequently). These two assumptions are quite common in major e-commerce markets. Under these assumptions, the number of ratings is proportional to Tr . After transactions provide adequate ratings, the *IBS* strategy can be executed. If there are more ratings in a certain period (or time window), the time window will be smaller, and the *IBS* strategy should be executed more frequently.

In this paper, we use the *combined reputation function* (Jøsang, 2002) defined with a binary rating scheme to calculate the seller's reputations. However, the rating mechanism in Yelp (which is the dataset we used as the simulation environment in this paper) is *K*-nomial. So, $r_t^K(b_i,s_j)(K>2)$ should be converted to $r_t^2(b_i,s_j)$. Definition 1 defines our conversion scheme and Definitions 2-5 define the reputations concept of our approach formally. The symbols used in these definitions and their meanings are summarized in Table 2.

Table 2. Symbols and their meanings used in this paper

Symbols	Meaning of symbols
Tr	The volume of ratings in recent period (i.e., a time window *T*)
T	The length of time window
$B=\{b_i \mid i=1,2,...,n\}$	The *active* reviewers set in recent time window *T*
$S=\{s_j \mid j=1,2,...,m\}$	The set of sellers
$r_t^K(b_i,s_j)(K>2,K\in\square)$	*K*-nomial rating of seller s_j suggested by reviewer b_i at time *t*
$r_t^2(b_i,s_j)$	Binary rating of seller s_j suggested by reviewer b_i ,1 and -1 denote positive and negative, respectively
$N_{pos}(b_i,s_j)$	The number of positive ratings reviewer b_i given to seller s_j
$N_{neg}(b_i,s_j)$	The number of negative ratings reviewer b_i given to seller s_j
$\bar{r}_{b_i}$	The average of the ratings that b_i gave to all the trading sellers
σ_{b_i}	The RMSE (Root Mean Square Error) of b_i 's rating

Definition 1. Equation (1) specifies the formula to convert a multinomial rating $r_t^K(b_i,s_j)$ into a binary value $r_t^2(b_i,s_j)$.

$$\forall \mathrm{t},\ r_t^2(b_i,s_j)=\begin{cases} 1\ ,\ if\ \ r_t^K(b_i,s_j)\geq\overline{K} \\ -1\ ,\ if\ \ r_t^K(b_i,s_j)<\overline{K} \end{cases} \tag{1}$$

where *K* represents the *K*-nomial ratings, $r_t^K(b_i,s_j)(K>2,K\in\square)$ denotes the rating of seller s_j according to reviewer b_i at time *t*, and $\overline{K}(\overline{K}\in\square)$ the overall average ratings within time window T.

Definition 2. (Jøsang, 2002) Equation 2 defines the reputation $Rep(b_i,s_j)$ that b_i rates s_j:

$$Rep(b_i,s_j)=\frac{N_{pos}(b_i,s_j)+1}{N_{neg}(b_i,s_j)+N_{pos}(b_i,s_j)+2} \tag{2}$$

Definition 3. (Jøsang, 2002) Equation 3 defines the reputation $Rep(b_i,S')$ that reviewer b_i rates the seller group S':

$$Rep(b_i,S')=\frac{\sum_{s_j\in S'} N_{pos}(b_i,s_j)\ \ +1}{\sum_{s_j\in S'} N_{neg}(b_i,s_j)+\sum_{s_j\in S'} N_{pos}(b_i,s_j)\ \ +2} \tag{3}$$

Definition 4. (Jøsang, 2002) Equation (4) defines the reputation $Rep(B',s_j)$ of seller *S* as reviewed by group B':

$$Rep(B',s_j)=\frac{\sum_{b_i\in B'}N_{pos}(b_i,s_j)\ \ +1}{\sum_{b_i\in B'}N_{pos}(b_i,s_j)+\sum_{b_i\in B'}N_{neg}(b_i,s_j)\ \ +2} \tag{4}$$

Definition 5. In the electronic market, Equation (5) aggregates the baseline reputation value that group *B* gives to seller *S*:

$$Rep(B,S)=\frac{\sum_{b_i\in B}\sum_{s_j\in S}N_{pos}(b_i,s_j)\ \ +1}{\sum_{b_i\in B}\sum_{s_j\in S}N_{pos}(b_i,s_j)+\sum_{b_i\in B}\sum_{s_j\in S}N_{neg}(b_i,s_j)\ \ +2} \tag{5}$$

Definition 6. Two statistics metrics of reviewer b_i 's ratings are given by Equations (6) and (7).

$$\bar{r}_{b_i}=\frac{\sum_{t\in T}\sum_{s_j\in s}r_t^k(b_i,s_j)}{Tr_{b_i}} \tag{6}$$

$$\sigma_{b_i}=\sqrt{\frac{\sum_{t\in T}\sum_{s_j\in s}(r_t^k(b_i,s_j)-\bar{r}_{b_i})^2}{Tr_{b_i}}} \tag{7}$$

3.2 The behavior characteristics of *lenient* reviewers and *strict* reviewers

In real life, different reviewers may have split opinions to similar products or services, and therefore may rate differently. Reviewers who tend to rate highly give us an *impression* as *lenient* ones, while those who tend to rate lowly leave a *strict impression* to us. These two kinds of reviewers can be defined according to the statistical characteristics of their historical ratings with following definition.

Definition 7. Suppose $\chi_t(b_i,s_j)$ is the rating deviation that buyer b_i rated s_j, whose value can be calculated according to $\chi_t(b_i,s_j)=r_t^k(b_i,s_j)-\bar{R}_j$, where $\bar{R}_j=\frac{\sum_{t\in T}\sum_{b_i\in B}r_t^k(b_i,s_j)}{Tr_{s_j}}$ is the average ratings that all reviewers who rated seller s_j in time windows T, and Tr_{s_j} is the number of transaction that seller s_j traded in time windows T. Let χ_H be a threshold of the rating deviation. *Lenient* and *strict* reviewers can be defined as follows.

- $\exists b_i\ for\ \forall s_j, if\ \ \chi_t(b_i,s_j)>0$ and $\max(\chi_t(b_i,s_j))\le\chi_H$, then $b_i\in B_{lenient}$;
- $\exists b_i\ for\ \forall s_j, if\ \ \chi_t(b_i,s_j)<0$ and $\max(-\chi_t(b_i,s_j))\le\chi_H$, then $b_i\in B_{strict}$.

From Definition 7, we can see that *lenient* reviewers have lenient or optimistic personality, and often give higher ratings than normal reviewers do, as they subjectively feel that the quality of the product is better than other persons do. However, *strict* reviewers have strict or captious personality, and often give lower ratings than normal reviewers do, as they subjectively feel that the quality of the product is poorer than other persons do. In particularly, the setting of χ_H guarantees that *lenient* and *strict* reviewers' rating deviation is not very large, which can exclude *dishonest* reviewers whose rating deviation is very large being chosen as benchmarking *lenient* and *strict* reviewers to a certain extent.

Based on Definition 7 and common knowledge in real life, we can infer that *lenient* and *strict* reviewers simultaneously satisfy following characteristics:

1) The average rating of lenient reviewers is moderately large, while the average rating of strict reviewer is moderately small.

2) Small rating variance σ_{b_i}. Since *lenient* are more optimistic and *strict* reviewers are more captious than normal ones, their ratings are always much higher or lower than normal ones. That is to say, the ratings from *lenient* and *strict* reviewers are much close to the highest/lowest score of 5/1 in a 5-rank rating mechanism. Therefore, their rating variances are relative small than normal ones.

3) Rareness. Due to the information asymmetry characteristics of e-commerce environment, honest people are cautious optimistic and captious when give ratings, therefore, *lenient* and *strict* reviewers are rare in general.

According to social psychology theory, impression depicts the phenomenon under which *one subjectively follows the understanding formed with previous experiences, and categorizes others under new situations based on the concepts formed under old situations*. Such a process reflects the clear *orientation of people's actions*, during which others are categorized (Anderson & Sedikides, 1991; Jin, 2005). Upon the formation of an impression, it dominates people's evaluation and interpretation of subsequent information. As such, *impression thus formed remains unchanged in the near future* (Xiang, 2006). Thus, we naturally assume the following about *lenient* reviewers and *strict* reviewers.

Assumption 1. Once a reviewer is classified as *strict* or *lenient*, the reviewer is expected to remain in the same category in the near future.

According to the impression theory and Assumption 1, if a reviewer has been *lenient* recently, we believe that he/she remains *lenient* in the near future. Therefore, if a *lenient* reviewer suddenly gives a low rating to a seller, we can intuitively believe that the quality of the product or service provided by this seller is indeed low. Similarly, if a *strict* reviewer sudden gives a high rating to a seller, then we can intuitively believe that quality of the product or service from this seller is indeed high.

Base on the above definition and analysis, we formalize the following two rules for classifying sellers who have traded with *strict* or *lenient* into *honest* and *dishonest* ones.

Rule 1: (*honest* sellers) If a seller s_j has traded with b_i Î B_{strict} and b_i gave positive rating to s_j (i.e., $r_t^2(b_i,s_j)$ =1), then s_j Î S_{honest} ;

Rule 2: (*dishonest* sellers) If a seller s_j has traded with b_i Î $B_{lenient}$ and b_i gave negative rating to s_j (i.e., $r_t^2(b_i,s_j)$ =-1), then s_j Î $S_{dishonest}$.

According to Rule 2/Rule 1, we can conclude that, in classifying sellers, we always discard *lenient*/*strict* reviewers' positive/negative ratings and only use their negative/positive ratings. However, it is still possible that *dishonest* reviewers may be hired to disguise their selves as *lenient*/*strict* reviewers and give unfair ratings to misguide the classification of sellers. To decrease such threat of misclassifying sellers, in the next section, we design a conflict elimination mechanism (see steps 10-11 in Algorithm 2).

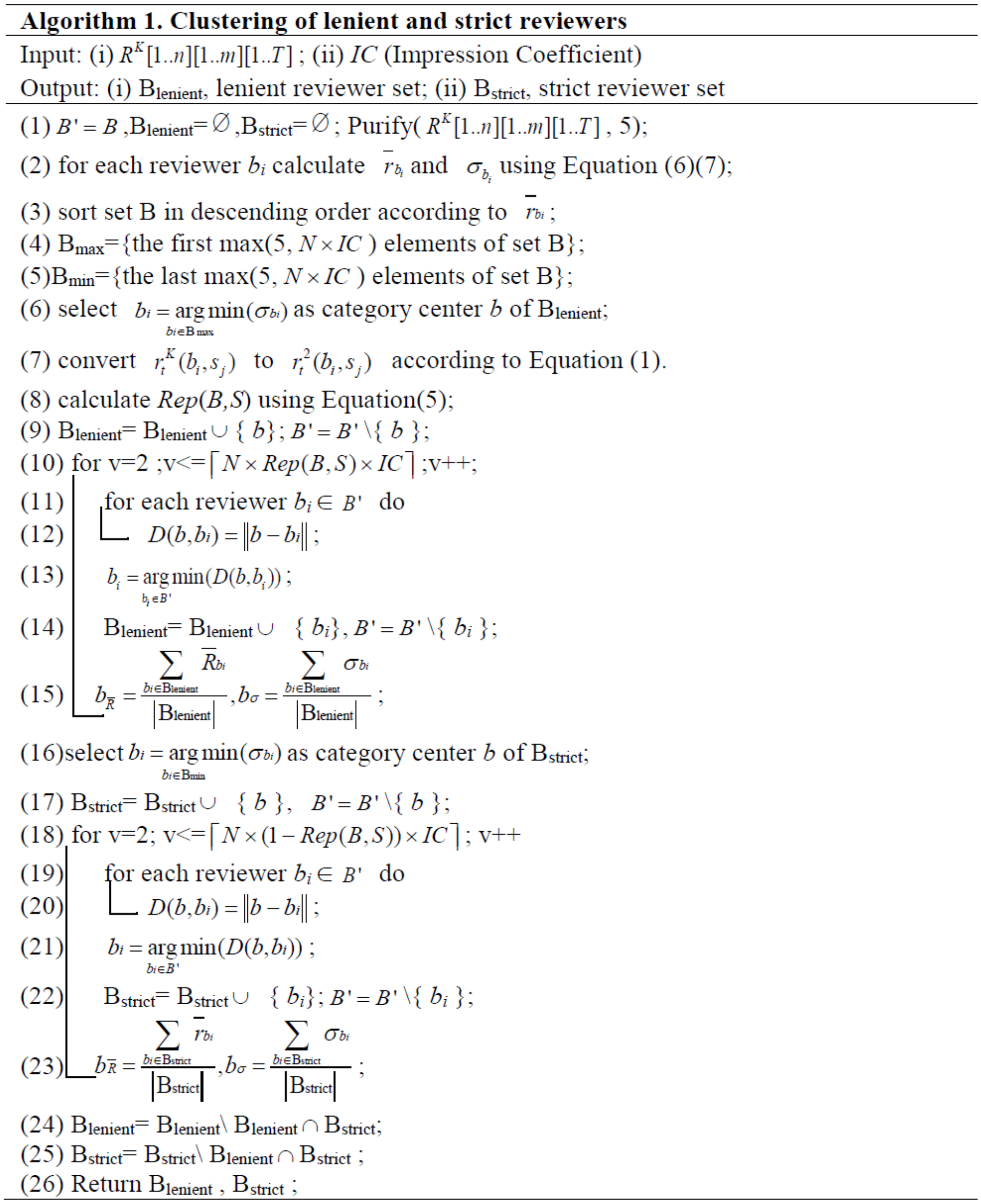
4 An impression-based defending strategy

Based on the framework and concepts defined in Section 3, we present an impression-based reputation attacks defending strategy (*IBS*) (Wang et al., 2017) comprising four steps. (1) Select some typical *lenient* and *strict* reviewers (Algorithm 1). (2) Based on the behavior expectation of impressed *lenient* and *strict* reviewers, pre-classify the sellers traded with these *lenient* and *strict* reviewers into *honest* and *dishonest* ones (Algorithm 2). (3) Classify all the reviewers into *honest*, *dishonest*, and *uncertain* ones based on their ratings to the pre-classified sellers (Algorithm 3). (4) Aggregate all sellers' reputations (Algorithm 4).

4.1 Clustering of lenient reviewers and strict reviewers

This paper proposes an algorithm to classify *lenient* and *strict* reviewers based on nearest neighbor search algorithm (see Algorithm 1). In our algorithm, we first initialize the parameters in this algorithm and purify the reviewers by discarding those with less than 5 reviews (considered as inactive) and their rating records (line 1). Secondly, we choose the reviewers whose rating mean is the largest but with the smallest rating deviation as the center for the *lenient* category. Similarly, the reviewers whose rating mean and rating deviation are combined to be smallest are chosen as the center for the *strict* category (lines 3-6). As we believe that *lenient* and *strict* reviewers are scarce in electronic markets, our algorithm sets a parameter called impression coefficient (i.e., *IC* and $IC << 1$) to manipulate the chosen ratio of *lenient* to *strict* reviewers from normal reviewers. The loop in lines 9-15 selects the closest point to the *lenient* category until reaching its upper limit of $\lceil N \times Rep(B,S) \times IC \rceil$, and updates the category center as soon as a new reviewer is considered. Similarly, *strict* reviewers are selected as shown in lines 18-23.

Finally, as a reviewer cannot be *strict* and *lenient* simultaneously, therefore the intersection set of *strict* reviewers and *lenient* reviewers ($B_{lenient} \cap B_{strict}$) are excluded (lines 24-25) for ruling out ambiguity.

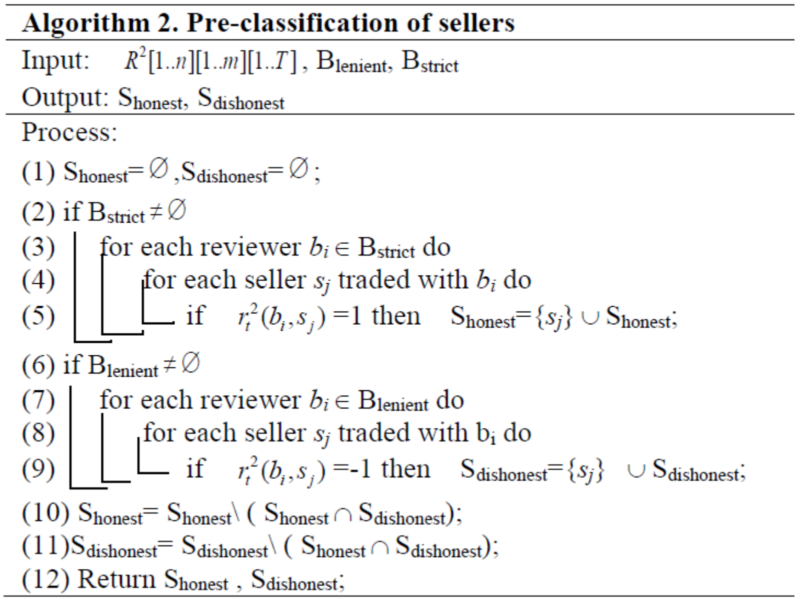


Algorithm 1. Clustering of lenient and strict reviewers

```
Input: (i) $R^K[1..n][1..m][1..T]$; (ii) $IC$ (Impression Coefficient)
Output: (i) $B_{lenient}$, lenient reviewer set; (ii) $B_{strict}$, strict reviewer set
(1) $B' = B$, $B_{lenient} = \varnothing$, $B_{strict} = \varnothing$; Purify( $R^K[1..n][1..m][1..T]$, 5);
(2) for each reviewer $b_i$ calculate $\bar{r}_{b_i}$ and $\sigma_{b_i}$ using Equation (6)(7);
(3) sort set B in descending order according to $\bar{r}_{b_i}$;
(4) $B_{max}$={the first max(5, $N \times IC$) elements of set B};
(5)$B_{min}$={the last max(5, $N \times IC$) elements of set B};
(6) select $b_i = \arg\min_{b_i \in B_{max}}(\sigma_{b_i})$ as category center $b$ of $B_{lenient}$;
(7) convert $r_t^K(b_i, s_j)$ to $r_t^2(b_i, s_j)$ according to Equation (1).
(8) calculate $Rep(B,S)$ using Equation(5);
(9) $B_{lenient} = B_{lenient} \cup \{b\}$; $B' = B' \setminus \{b\}$;
(10) for v=2 ;v<=$\lceil N \times Rep(B,S) \times IC \rceil$;v++;
(11)     for each reviewer $b_i \in B'$ do
(12)         $D(b,b_i) = \|b - b_i\|$;
(13)     $b_i = \arg\min_{b_i \in B'}(D(b,b_i))$;
(14)     $B_{lenient} = B_{lenient} \cup \{b_i\}$, $B' = B' \setminus \{b_i\}$;
(15)     $b_{\bar{R}} = \frac{\sum_{b_i \in B_{lenient}} \bar{R}_{b_i}}{|B_{lenient}|}, b_\sigma = \frac{\sum_{b_i \in B_{lenient}} \sigma_{b_i}}{|B_{lenient}|}$;
(16)select $b_i = \arg\min_{b_i \in B_{min}}(\sigma_{b_i})$ as category center $b$ of $B_{strict}$;
(17) $B_{strict} = B_{strict} \cup \{b\}$, $B' = B' \setminus \{b\}$;
(18) for v=2; v<=$\lceil N \times (1 - Rep(B,S)) \times IC \rceil$; v++
(19)     for each reviewer $b_i \in B'$ do
(20)         $D(b,b_i) = \|b - b_i\|$;
(21)     $b_i = \arg\min_{b_i \in B'}(D(b,b_i))$;
(22)     $B_{strict} = B_{strict} \cup \{b_i\}$; $B' = B' \setminus \{b_i\}$;
(23)     $b_{\bar{R}} = \frac{\sum_{b_i \in B_{strict}} \bar{r}_{b_i}}{|B_{strict}|}, b_\sigma = \frac{\sum_{b_i \in B_{strict}} \sigma_{b_i}}{|B_{strict}|}$;
(24) $B_{lenient} = B_{lenient} \setminus B_{lenient} \cap B_{strict}$;
(25) $B_{strict} = B_{strict} \setminus B_{lenient} \cap B_{strict}$;
(26) Return $B_{lenient}$, $B_{strict}$;
```

4.2 Pre-classification of sellers

Typical *lenient* and *strict* reviewers are selected as benchmarks for pre-classifying sellers who have traded with these reviewers. Algorithm 2 illustrates the steps for pre-classifying the sellers using the selected *lenient* and *strict* reviewers. Firstly, based on the groups $B_{lenient}$ and B_{strict} obtained from Algorithm 1, we classify the sellers recently transacted with *strict* reviewers and were rated positively as *honest* ones (lines 2-5). Then, the sellers recently transacted with *lenient* reviewers and were rated negatively are regarded as *dishonest* ones (lines 6-9). For ruling out the influence of ambiguity, we also remove the controversial sellers who are both *honest* and *dishonest* (i.e., in the set of $S_{honest} \cap S_{dishonest}$) (lines 10-11).



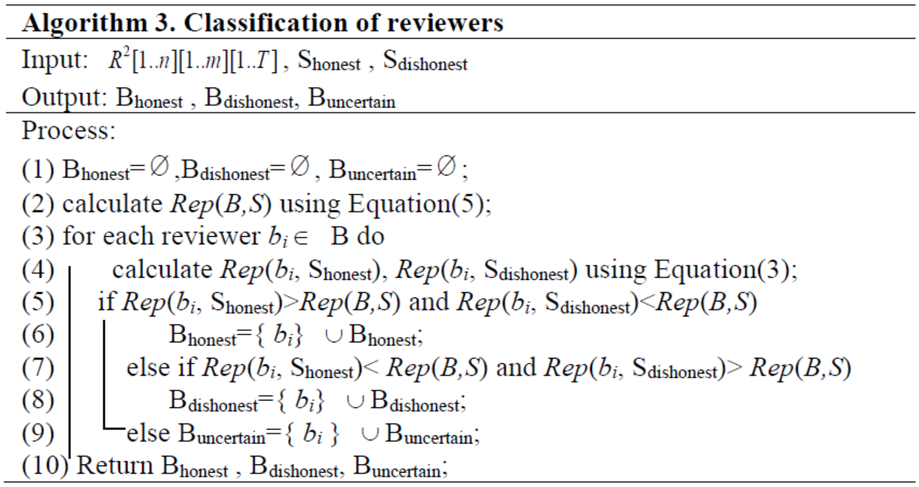
Algorithm 2. Pre-classification of sellers

```
Input:   $R^2[1..n][1..m][1..T]$, $B_{lenient}$, $B_{strict}$
Output: $S_{honest}$, $S_{dishonest}$
Process:
(1) $S_{honest}=\varnothing$ ,$S_{dishonest}=\varnothing$ ;
(2) if $B_{strict} \neq \varnothing$
(3)     for each reviewer $b_i \in B_{strict}$ do
(4)         for each seller $s_j$ traded with $b_i$ do
(5)             if   $r_t^2(b_i, s_j)=1$ then   $S_{honest}=\{s_j\} \cup S_{honest}$;
(6) if $B_{lenient} \neq \varnothing$
(7)     for each reviewer $b_i \in B_{lenient}$ do
(8)         for each seller $s_j$ traded with $b_i$ do
(9)             if   $r_t^2(b_i, s_j)=-1$ then   $S_{dishonest}=\{s_j\} \cup S_{dishonest}$;
(10) $S_{honest}= S_{honest} \setminus ( S_{honest} \cap S_{dishonest})$;
(11)$S_{dishonest}= S_{dishonest} \setminus ( S_{honest} \cap S_{dishonest})$;
(12) Return $S_{honest}$ , $S_{dishonest}$;
```

4.3 A reviewer classification algorithm

The categories of *lenient* and *strict* reviewers are different from the categories of *honest*, *dishonest*, and *uncertain* ones. The former two categories are classified according reviewers' rating behaviors characteristics (i.e., statistic characteristics of Definition 6), while the latter three categories are classified according to the fairness and unfairness property of reviewers' ratings. Algorithm 1 only selects a small number of *lenient* and *strict* reviewers respectively, and does not classify reviewers according the fairness/unfairness property.

Therefore, the aim of Algorithm 3 is to classify all the reviewers into *honest*, *dishonest*, and *uncertain* categories, which is realized by following steps. First, considering each reviewer's ratings to *honest* and *dishonest* sellers, as well as the market's overall reputation score, we first judge accordingly whether the reviewer in question is an *honest* or *dishonest* one. That is to say, according to Definition (5) in Section 3, we calculate the reputation score of the market $Rep(B,S)$ (line 2 in Algorithm 3). Besides, according to Definition (3) in Section 3, the reputation score of the *honest* to *dishonest* sellers $Rep(b_i, S_{honest}) / Rep(b_i, S_{dishonest})$ with respect to each reviewer b_i is calculated in line 4. The reviewers who simultaneously give high ratings to *honest* sellers and low ratings to *dishonest* sellers are regarded as *honest* ones (lines 5-6); while the reviewers who simultaneously give low ratings to *honest* sellers and high ratings to *dishonest* sellers are regarded as *dishonest* ones. Reviewers belonging to neither the *honest* nor the *dishonest* category are considered as *uncertain* ones (lines 3-9).



Algorithm 3. Classification of reviewers

Input: $R^2[1..n][1..m][1..T]$, S_{honest}, $S_{dishonest}$

Output: B_{honest}, $B_{dishonest}$, $B_{uncertain}$

```
Process:
(1) $B_{honest}=\varnothing$, $B_{dishonest}=\varnothing$, $B_{uncertain}=\varnothing$;
(2) calculate $Rep(B,S)$ using Equation(5);
(3) for each reviewer $b_i \in B$ do
(4)     calculate $Rep(b_i, S_{honest})$, $Rep(b_i, S_{dishonest})$ using Equation(3);
(5)     if $Rep(b_i, S_{honest})>Rep(B,S)$ and $Rep(b_i, S_{dishonest})<Rep(B,S)$
(6)         $B_{honest}=\{b_i\} \cup B_{honest}$;
(7)     else if $Rep(b_i, S_{honest})< Rep(B,S)$ and $Rep(b_i, S_{dishonest})> Rep(B,S)$
(8)         $B_{dishonest}=\{b_i\} \cup B_{dishonest}$;
(9)     else $B_{uncertain}=\{b_i\} \cup B_{uncertain}$;
(10) Return $B_{honest}$, $B_{dishonest}$, $B_{uncertain}$;
```

4.4 Aggregation of seller's reputation

After classifying all the reviewers, we calculate the sellers' reputations based on these reviewers' ratings according to Equation (8). Further, Equation (9) computes the weights between the *honest* and *uncertain* reviewers.

$$Ag\_rep(s_j) = (1\text{-}w) \times Rep(\mathrm{B}_{\mathrm{honest}}\ ,\ s_j) + w \times Rep(\mathrm{B}_{\mathrm{uncertain}}\ ,\ s_j) \qquad (8)$$

where $\mathrm{B}_{\mathrm{honest}} \cup \mathrm{B}_{\mathrm{uncertain}} \cup \mathrm{B}_{\mathrm{dishonest}} = B$, and (1-$w$), w are weights assigned to *honest* and *uncertain* reviewers, respectively. $Rep(\mathrm{B}_{\mathrm{honest}}\ ,\ s_j)$, $Rep(\mathrm{B}_{\mathrm{uncertain}}\ ,\ s_j)$ can be evaluated with Equation(4) as defined in Section 3.

$$w = CF \times \frac{|\mathrm{B}_{\mathrm{uncertain}}|}{|B|} \qquad (9)$$

where CF (Confidence, $0 \le CF < 1$) represents the confidence level to *uncertain* reviewers, which can be customized by the platform. The lager the CF value, the more a platform trust *uncertain* reviewers, so that 0 indicates that a platform completely distrust *uncertain* reviewers. $|\mathrm{B}_{\mathrm{uncertain}}|$ and $|\mathrm{B}|$ are the number of *uncertain* reviewers and all reviewers, respectively.

5. Experiment

We evaluate our scheme with two sets of experiments. We design our first set on a multi-agent-based electronic market simulation platform. This set of experiments comprise four subsets of experiments, aiming at analyzing the performance limitations of our strategy when it is confronted with some extreme attacks or in some extreme environments (e.g., the proportion of dishonest reviewers is so large that it may lead to adverse selection and moral hazard, which is a danger for real electronic markets) and evaluates our strategy against the models such as iClub (Liu et al. 2014), Amazon, E-sporas (Guo et al., 2009), and AARE (Ji et al., 2017). The second set of experiments are conducted over the Yelp dataset, which aims at evaluating our scheme under real-life situations when defending unknown attacks. In the real dataset experiment, we do not compare the performance of above strategies because of the limitation and absence of attributes. In the Yelp dataset, the transaction between each pair of seller and reviewer is one shot, i.e., no reviewer rated a seller more than once. So the iClub strategy cannot be implemented because it cannot deal with the situation that each pair of reviewer and seller traded only once. Besides, the E-sporas and AARE strategies consider the reputation of the reviewer when calculating the reputation of the seller, but the Yelp dataset does not provide such kind of attributes, therefore, these strategies cannot be implemented based on the Yelp dataset. The following subsections 5.1 and 5.2 illustrate these two sets of experiments in detail.

5.1 Experiments over simulated dataset

(1) Experiments setting

According to Figure 1 in Section 3, in the simulated electronic market, there are three kinds of agents (i.e., seller, buyer, and platform agents). Both seller and reviewer agents can be divided into honest and dishonest ones. In addition, we simulated three rating behaviors of honest reviewers (normal, *lenient* and *strict*), and six rating behaviors of dishonest reviewers (*AlwaysUnfair*, *Camouflage*, *Whitewashing*, *Sybil*, *Sybil_Camouflage*, and *Sybil_Whitewashing*) (Jiang, 2013). As our simulation assumes that there are no duopoly sellers, all the sellers are equal. Moreover, as we pair the sellers and the buyers randomly in transactions, a buyer may choose a seller as trading partner for multiple times, which is common in electronic markets. In this simulation environment, honest sellers offer superior quality of articles or services, while dishonest sellers offer inferior quality ones. Honest reviewers provide fair ratings, while dishonest reviewers provide unfair ones as attacks. To simplify the modeling of the quality of honest sellers' products or services, we set one-half of the honest sellers' real quality to 1 and the other half of the honest sellers' real quality to 0.8. Similarly, we set one-half of the dishonest seller's real quality to 0 and the other half of the dishonest sellers' real quality to 0.2. The actual quality of honest and dishonest sellers' products or services is secret to the defending strategies.

Moreover, the proportion of *lenient* and *strict* reviewers to the total honest reviewers is 20%. For example, if there are 30 *honest* reviewers in an experiment, then the number of *strict* and *lenient* reviewers is $\lceil 30 \times 20\% \rceil = 6$, respectively. In our simulation, the rating that a *lenient* reviewer gives to the trading seller is assumed one grade higher than the fair quality of the seller's service or product, with the highest grade being 5. This means, if a seller's real quality is 4, the *lenient* reviewer's rating is 5; however, if a seller's real quality is 5, the *lenient* reviewer's rating is also 5 as it cannot be higher. Similarly, a *strict* reviewer's rating to the trading seller is assumed one grade lower than the real value, but the lowest is still 1. Similar to the setting of honest and dishonest sellers, these settings about buyers are also secret.

Based on above settings about sellers and buyers, four sets of experiments are simulated. The first set of experiments aim at analyzing the relationships between the estimation accuracy of seller reputation and the variation of ratings volume, as well as finding the lowest ratings volume that the reputation estimation accuracy can become stable. The second set of experiments is designed to analyze the variation trend of sellers' reputation estimation accuracy under different combinations of parameters. Similarly, assuming that the selected platform ratings volume is large enough to keep estimation accuracy stable. The third set of experiments analyze the robustness (i.e., being able to keep the estimation accuracy stable with an increasing proportion of *dishonest* reviewers) of our strategy. The fourth set of experiments evaluates our strategy against iClub (Liu et al. 2014), Amazon, E-sporas (Guo, 2009), and AARE (Ji et al., 2017). These strategies are selected because they are popular in the industry or recent in the academia, and they all calculate sellers' reputations from a global (or platform) view.

Table 3 lists all the parameters in these four sets of experiments. In the first set of experiments, the volume of ratings in the market varies from 300 to 2500 with increments of 100. The ratings volume that can make the performance of our strategy stable directly determines the appropriate size of the time window for our strategy. In the second set of experiments, to explore the influence of *IC* (i.e., ratio of *lenient* to *strict* reviewers) on the performance of our strategy, *IC* is assigned with 0.1, 0.125, 0.15, 0.175, and 0.2, respectively. Therefore, the second set of experiments aims at determining under which value of parameter *IC* that our strategy can reach optimal performance. To analyze the stability of our strategy, we construct two subsets of experiments in the third set to simulate the variation of the proportion of *dishonest* sellers and reviewers, respectively. In the first subset 3a, the proportion of *dishonest* sellers varies from 20% to 80% in steps of 10%. In the second subset 3b, the proportion of *dishonest* reviewers varies from 20% to 80% also in steps of 10%. If the reputation prediction accuracy is stable with different proportions of dishonest seller/reviewers, we can say that our strategy is stable. In the fourth set of experiments, to compare the performance of our strategy IBS, three other strategies such as Amazon, E-Sporas, and AARE are selected as benchmarks under various electronic market environments, in which the proportion of dishonest reviewers is even larger than 60% (i.e., the majority of reviewers are dishonest ones). Table 4 lists the parameters and the assigned values of E-sporas and AARE in experiments. Besides, in this set of experiments, the behavior of sellers is assumed to be consistent, which means that the products or services provided by honest sellers tend to be good, while those provided by dishonest sellers tends to be fake or inferior.

Table 3. Parameter settings in the experiments

Parameter	**Set 1**	**Set 2**	**Set 3**		**Set 4**
			a	b	
Rating grades	5	5	5	5	5
IC	0.15	0.1,0.125, 0.15,0.175,0.2	0.15	0.15	0.15
Dishonest reviewers	30	40	30	20,30,40,50,60,70,80	20,30,40,50,60,70,80
Honest reviewers	70	60	70	80,70,60,50,40,30,20	80,70,60,50,40,30,20
Dishonest sellers	16	16	8,12,16,20,24,28,32	16	16
Honest sellers	24	24	32,28,24,20,16,12,8	24	24
CF	0.7	0.7	0.7	0.7	0.7
Ratings volume	300-2500	2500	2500	2500	2500

Table 4. The experiment parameters settings for models

models	**Parameters**	**meanings**	**values**
E-Sporas	D	maximal value of reputation	10
	θ	adjusting parameter	10
	σ	acceleration factor	1
AARE	α	scaling factor of damping function	0.05
	θ	damping factor	15
	λ	time discount factor	0.7
	Price	transaction price	500
	l	coefficient of compressibility	25

(2) Evaluation criteria

We evaluate the accuracy of our scheme with the *MAE* (mean absolute error) of the aggregated reputation of sellers (denoted as $Ag\_rep(s_j)$) and real reputation score of sellers (denoted as $Rel\_rep(s_j)$) as the criteria. Equation

(11) defines how *MAE* is calculated (ranging 0 to 1), with a smaller value representing a more accurate defending strategy or a better defense performance. In this paper, the *MAE* of *dishonest* sellers is not adopted because the combination of *honest* sellers' *MAE* and the following *MCC* is adequate to reveal the performance of a strategy.

$$MAE = \frac{\sum_{s_j \in S} \left| Ag\_rep(s_j) - Rel\_rep(s_j) \right|}{|\, S \,|} \tag{11}$$

where $Ag\_rep(s_j)$ is seller s_j's aggregated reputation computed with Equation (8), $|\, S \,|$ the number of sellers in the electronic market, and $Rel\_rep(s_j)$ s_j's real reputation score.

We also use Matthews Correlation Coefficient (*MCC*, Matthews 1975) as an alternative evaluation criterion for measuring classification accuracy of our strategy, which is computed as follows.

$$MCC = \frac{tp \times tn - fp \times fn}{\sqrt{(tp + fp) \times (tp + fn) \times (tn + fp) \times (tn + fn)}} \tag{12}$$

where *fp*, *tp*, *fn*, and *tn* represent the numbers of false positives, true positives, false negatives, and true negatives, respectively.

The *MCC* value ranges from −1 to 1, where 1 reflects perfect filtering, −1 completely wrong filtering, and 0 an arbitrary result. *MCC* reveals the classification accuracy of sellers. The nearer a MCC value is to 1, the more accurate the classification.

(3) Results and analysis on effectiveness and accuracy

As our strategy considers mainly historical information, we consider the volume of ratings starting from 300. Figure 2 shows the performance of our strategy measured by the *MAE* that is calculated according to Equation (11). From Figure 2, we can see that the *MAE* curves of *whitewashing, Sybil_whitewashing* attacks tend to decrease stably with the increase of ratings volume, and converge to 0.1 after the ratings volume exceeds 1100. Specially, our strategy is effective against *Camouflage* attack and its combination with *Sybil* attack, in which attackers frequently change their actions of giving fair and unfair ratings to break the defense of impression-based strategies of the reputation system. From the characteristics of *lenient* and *strict* reviewers and the nearest neighbor search algorithm given in Algorithm 1, our impression-based strategy is theoretically robust against these two kinds of attacks. In contrast, *Camouflage* attackers (hired by dishonest or collusive sellers) change their ratings very frequently to avoid being detected. Such frequent changes exclude the attackers from the *lenient* and *strict* reviewer sets. One may argue that a seller may employ many *Camouflage* attackers acting as *strict* reviewers giving very low ratings to competitors, as well as *lenient* reviewers giving very high ratings to its own for very long time windows. However, such a *Camouflage* approach will bear a high cost, which may even outweigh the gain, and thus is seldom considered by attackers. Therefore, according to Figure 2, we can see that the curves of *camouflage* and *Sybil_camouflage* attacks approach 0.12 the fastest, and they are very stable too. These demonstrate the robustness of our *lenient* and *strict* reviewer selection algorithm empirically.

In Figure 2, the *MAE* curves of *AlwaysUnfair* and *Sybil* attack decrease irregularly and slowly with the increase of ratings volume. Moreover, the final *MAEs* approach 0.1, as good as the other four attacks after the ratings volume exceeds 1600. However, it is still acceptable. The *MAEs* under *AlwaysUnfair* and *Sybil* attack is inferior because the majority of reviewers are dishonest and the dishonest attackers' identities have changed before the defense strategy accumulates enough experiences to correctly judge their honesty.

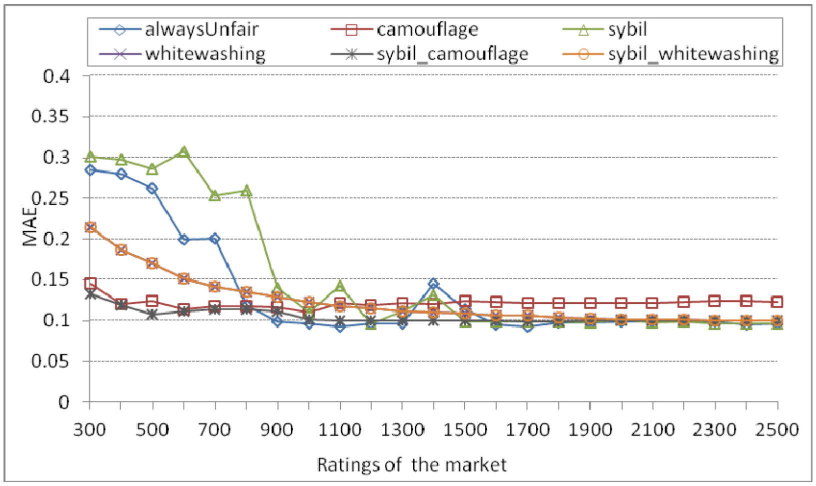


Figure 2. Effectiveness of *IBS* under various attacks

Taking *MAE* as benchmark, the above results reveal our strategy's prediction accuracy of *honest* sellers' reputations under various attacks. We can also illustrate the performance of our strategy by *MCC*. From Figure 3, we can see that the curves of *MCC* stably converge to 1 under all attacks, except for *Camouflage* and *Sybil_camouflage*. However, more transactions (after 1600 and 1200, respectively) is needed under the *AlwaysUnfair* attack and the *Sybil* attack, because these attacks are more difficult to defend compared to the others. The curves of MCC under *Camouflage* attack and *Sybil_camouflage* attack converge to 0.9 stably. It is still acceptable. From above results, it can be concluded that the overall performance of our strategy is desirable.

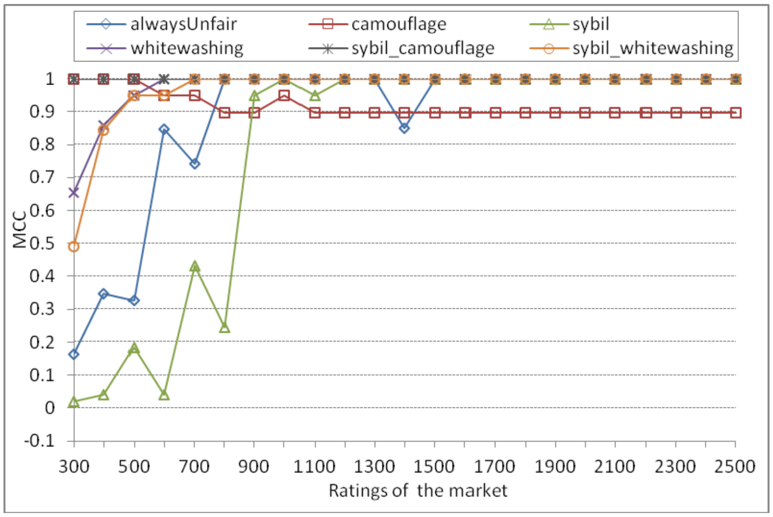


Figure 3. Accuracy of *IBS* defending against different attacks

(4) Results and analysis about various parameters

As the *IC* parameter reflects the rarity of *lenient* and *strict* reviewers (the ratio of these two types of reviewers to all reviewers), the value of *IC* should not be too large. In this paper, we assume that the value of *IC* in the 0.1-0.2 range. A proper value of *IC* parameter can directly improve the classification accuracy of reviewers. However, as the ratio of *lenient* and *strict* reviewers is dynamically changing and unknown to the platform, it is difficult for platform and the designer to choose a proper value for *IC*. To find a proper value of *IC*, we design and implement a set of experiments by assigning different values of *IC* according to interpolation method.

Tables 5 and 6 show the mean and deviation of the *MCC* values and *MAE* values (mean ± deviation) upon defending against various attacks, respectively. To compare the performance of these values, we should note that the deviation after symbol "±" should be compared first (the smaller the deviation, the more stable the performance), and then the mean (the larger the mean *MCC*, the better the performance; the near the *MAE* to 0, the better the performance). That means, in *IBS*, we pay much attention to the stability of an estimation. From these tables, we can see that, under the *Whitewashing* and *Sybil_whitewashing* attacks, the changes of *IC* do not make significant difference to *MCC* and *MAE*. Under the *AlwaysUnfair*, *Camouflage*, *Sybil*, and *Sybil_camouflage* attacks, the outstanding *MCC and MAE* are highlighted in bold type. Comparing the performance highlighted in Tables 5 and 6, we can see that *IC*=0.175 can gain an overall optimal performance.

Table 5. Results of *MCC* with various *IC*

IC	***AlwaysUnfair***	***Camouflage***	***Whitewashing***	***Sybil***	***Sybil_camouflage***	***Sybil_whitewashin***
0.1	0.90±0.16	0.90±0.01	1.00±0.00	0.61±0.43	1.00±0.00	1.00±0.00
0.125	0.90±0.17	0.90±0.01	1.00±0.00	0.77±0.29	1.00±0.00	1.00±0.00
0.15	0.93±0.10	**0.90±0.00**	1.00±0.00	0.81±0.25	1.00±0.00	1.00±0.00
0.175	**0.96±0.06**	**0.90±0.00**	1.00±0.00	0.89±0.20	1.00±0.00	1.00±0.00
0.2	0.93±0.10	**0.90±0.00**	1.00±0.00	**0.90±0.16**	1.00±0.00	1.00±0.00

Table 6. Results of *MAE* with various *IC*

IC	***AlwaysUnfair***	***Ccamouflag***	***Whitewashing***	***Sybil***	***Sybil_camouflage***	***Sybil_whitewashin***
0.1	0.13±0.03	0.12±0.00	0.11±0.00	0.18±0.08	0.10±0.00	0.11±0.00
0.125	0.13±0.04	0.13±0.00	0.11±0.00	0.16±0.06	0.10±0.00	0.11±0.00
0.15	0.12±0.02	**0.12±0.00**	0.11±0.00	0.14±0.05	0.10±0.00	0.11±0.00
0.175	**0.11±0.02**	**0.12±0.00**	0.11±0.00	0.13±0.04	0.10±0.00	0.11±0.00
0.2	0.12±0.04	0.13±0.00	0.11±0.00	0.14±0.04	0.10±0.00	0.11±0.00

According to the experimental results, the value of *IC* changes in positive correlation with the reputation baseline ($Rep(B,S)$ of Equation 5) of the whole market. That is, when the value of baseline is small, *IC* should be assigned with a little smaller value; otherwise, *IC* should be set to a little larger.

(5) Results of IBS under different proportions of dishonest sellers and dishonest reviewers

Figure 4 depicts the change of *MAE* when the proportion of *dishonest* sellers increases from 20% to 80%. The curves of *Camouflage*, *Whitewashing*, *Sybil_whitewashing,* and *Sybil_camouflage* remain at quite a low value, especially when the proportion of dishonest sellers is near to 50%.As the proportion of dishonest sellers is higher than 50%, the *MAE* values increase only slightly. The trend of *AlwaysUnfair* and *Sybil* curves is quite similar and the two curves are almost parallel to each other at most proportions. However, when 70% or 80% sellers are dishonest, the *MAE* values are larger than 0.1 significantly (i.e., the estimated reputation of *honest* sellers deviates from their real reputation greatly). That is because most of the sellers being dishonest lower the overall reputation of the sellers, which is consistent with common sense. As such, our strategy can be regarded as very stable under the *Camouflage*, *Whitewashing*, *Sybil_whitewashing,* as well as *Sybil_camouflage* attacks. However, under the *AlwaysUnfair* and *Sybil* attacks, our strategy is not so stable under hypothetical extreme cases.

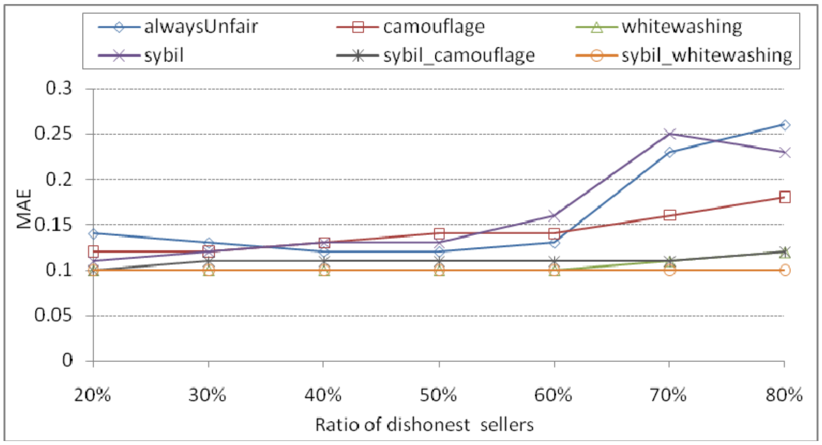


Figure 4. Robustness of our IBS strategy with increasing proportion of dishonest sellers

Figure 5 depicts the *MAE* when the proportion of dishonest reviewers increases from 20% to 80%. In general, with the proportion less than 60%, the *MAEs* remain stable around 0.1 under different attacks. However, as the proportion of dishonest reviewers further increases, the *MAEs* of *Camouflage* and *Sybil_camouflage* attacks increase slightly (from 0.1 to 0.2), but the *MAEs* of other attacks increase abruptly from 0.1 to 0.5. This is likely because determining *lenient* and *strict* reviewers accurately is hard when most reviewers are dishonest. From the above results, it can be concluded that the IBS strategy is stable even in some extreme environments (e.g., the proportion of dishonest reviewers or sellers is larger).

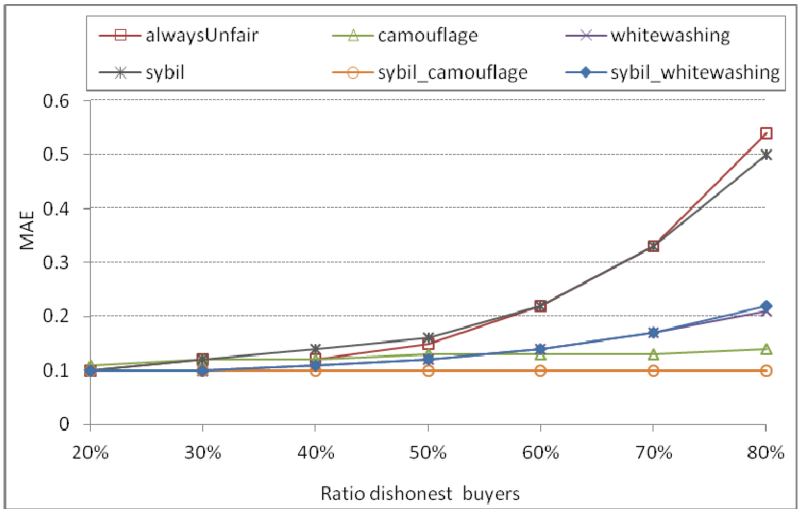
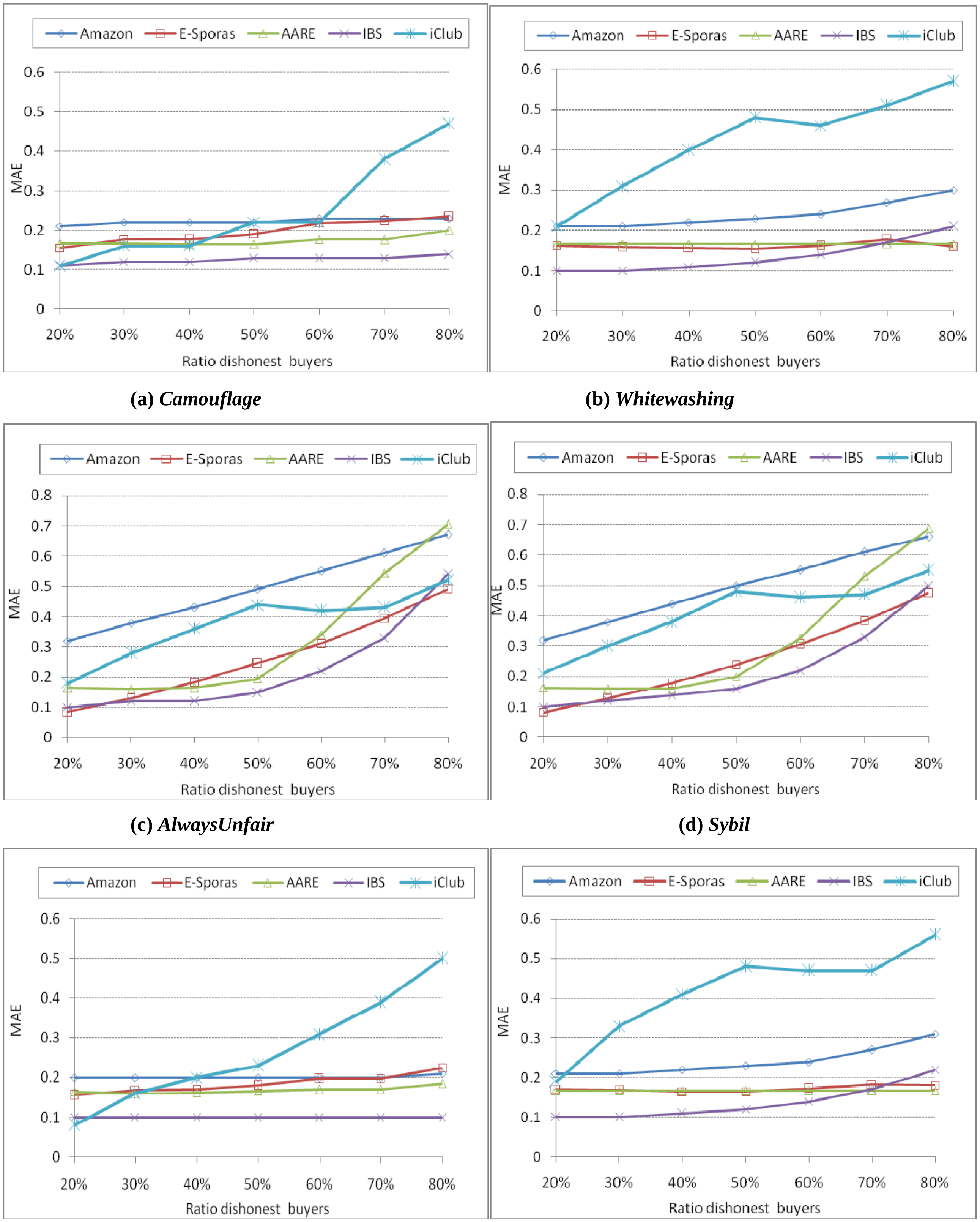


Figure 5. Robustness of our IBS strategy with increasing proportion of dishonest reviewers

(6) Comparisons with other methods

The fourth set of experiments evaluate the *MAE* of our strategy against four other strategies (i.e., iClub, Amazon, E-sporas, and AARE) under different proportions of *dishonest* reviewers, with the proportion of *dishonest* sellers fixed at 40%. Considering the fact that e-commerce platform managers would try their best to maintain market order in reality, the proportion of dishonest sellers should not be so high. That is because, if the proportion of dishonest sellers is very high, there will be moral hazards and adverse activities in the market, and then the market may crash. Therefore, it is reasonable to assume that 40% is the worst proportion of dishonest sellers. If a filtering strategy is stable at this point, it should be stable when the proportion of *dishonest* sellers is smaller than 40%. Figure 6 depicts the *MAE* curves of iClub, Amazon, E-sporas, AARE, and IBS under different attacks. Under *AlwaysUnfair* and *Sybil* attacks (Figure 6(c) and (d)), our strategy performs the best when the proportion of dishonest reviewers is less than 70%. With an increasing proportion of dishonest reviewers, The *MAE* values of IBS strategy increase from 0.1 to above 0.5, those of E-sporas increase from 0.09 to 0.5, and those of AARE increase from 0.16 to about 0.7. In comparison, the performance of the Amazon strategy performs the worst, because it employs neither an accumulation method nor a filtering method. However, with an increasing proportion of dishonest reviewers, iClub, E-Sporas and AARE cannot readily discover trustable reviewers. In contrast, our IBS method can still identify *lenient* and *strict* reviewers quite accurately even if most reviewers are *dishonest*, as our selection method is resilient to their rareness.

Under *Camouflage* and *Sybil_camouflage* attacks (see Figures 6(a) and (e), respectively), our strategy remains stable (with *MAE* around 0.1) and significantly performs better than the other four strategies. The *MAE* values of E-Spore and AARE strategies increase slightly from 0.15/0.16 to 0.23/0.20 with the increase proportion of *dishonest* reviewers, respectively. The Amazon strategy performs inferior than others, with its *MAE* value remaining around 0.20. Under *Whitewashing* and *Sybil_whitewashing* attacks, our method outperforms the other four methods when the proportion of *dishonest* reviewers is less than 70%, its *MAE* value increases slightly from 0.1 to 0.16 (see Figures 6(b) and (f)). The *MAE* values of E-Spore and AARE remains about 0.16 in all cases. Although the E-Spore and AARE strategies perform better than IBS strategy when the proportion of *dishonest* reviewers is 80%, such extreme situation is impossible in reality. Compared to other methods, the iClub method has the worst performance, especially when defending against *Whitewashing* and *Sybil_whitewashing* attacks. This is because iClub needs to accumulate certain trading experience of buyers and sellers when filtering. If *dishonest* reviewers frequently change their identity, it is difficult for the iClub strategy to identify them.



(a) ***Camouflage***

(b) ***Whitewashing***

(c) ***AlwaysUnfair***

(d) ***Sybil***

(e) ***Sybil_camouflage***

(f) ***Sybil_whitewashing***

Figure 6. MAEs of different methods defending against various attacks

In summary, based on the experimental results over simulated dataset, we can draw following conclusions.

Conclusion 1. As long as enough trading experiences (i.e., ratings) are accumulated, the IBS strategy is accurate and stable in predicting sellers' reputation.

Conclusion 2. Even in the extreme environment with a high proportion of dishonest reviewers, the IBS strategy presented in this paper outperforms the other four benchmarks when defending against most simulated attacks.

5.2 Experiments over real dataset

(1) Experiments setting

To further demonstrate the performance of our strategy, we test it over a real-life dataset, the Yelp restaurant data (Mukherjee, 2013). It comprise a total number of 67,019 rating records and the rating time spans from October 2004 to October 2012. All "Y" reviews are obtained from the filtered section and "N" reviews from the regular pages. The proportion of reviews labeled with "N" is 87.6%. The total number of reviewers and sellers are 35,028 and 129, respectively. Besides, in the Yelp dataset, the reputation of each seller is denoted as $Yelp\_rep(s_j)$, where $Yelp\_rep(s_j) \in \mathbb{R}$ and $0 < Yelp\_rep(s_j) \leq 5$. Different from the reputations as recorded in the dataset, the estimated reputation presented in this paper can be calculated over any period. To bridge the gap between these two kinds of reputations, we first arrange the Yelp reviews in reverse chronological order. The more recent a review is given, the nearer is it to the front of the queue (see the bottom rectangle of Figure 7). It should be noted that the labeled reputations given by Yelp is accumulated ones since sellers' account creation until the crawling time. According to the data extracted method in Figure 7, the bigger the time window, the closer the predicted value of IBS's seller reputation is to Yelp's label value.

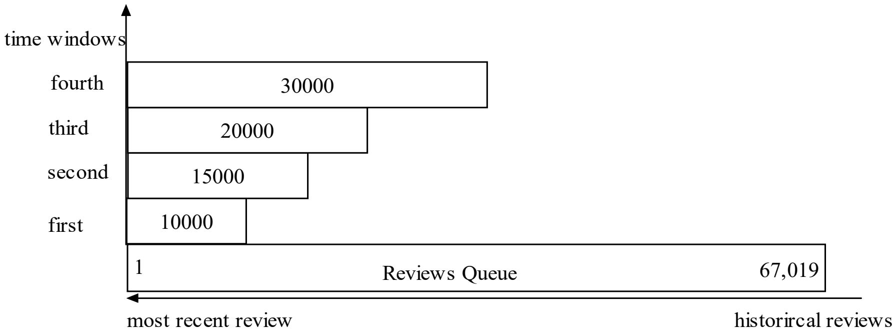


Figure 7 The extraction method of time window in two subsets of experiments

Before implementing our strategy, in order to eliminate data noise, we also pre-process the extracted data. Three data preprocessing methods (ATV-3, ATV-4, and ATV-5) are used and compared to delete reviewers with ratings volume below 3, 4, and 5, respectively. Figure 7 illustrates these preprocessing methods in graphical form. Table 7 shows a sample of data features that are extracted and pre-processed following the ATV-5 pre-processing methods.

Table 7 Characteristics of samples after preprocessing (ATV-5)

Time windows (ratings volume)	reviewers	sellers	reviews	proportion of fair reviews
10000	139	100	903	0.873
20000	386	105	2723	0.874
30000	732	109	5477	0.873
40000	1139	112	8890	0.871
50000	1578	116	13019	0.869
60000	2065	117	17834	0.871

Two sub-sets of experiments are designed and implemented over the pre-processed real dataset. The first subset of experiments is to verify the influence of various values of parameters such as time window (i.e., ratings volume), *IC* and *CF*. The second subset of experiments aims at evaluating the effectiveness and stability of our approach. In the first subset of experiments, to explore the influence of time window, *IC* and *CF* on the performance of our strategy. Since a small window (lack of trading experience) will lead to unstable prediction results of the algorithm, in experiments, we assign time windows with values of 30000, 40000, and 50000, respectively. Moreover, *IC* is assigned with 0.14, 0.16, 0.18, 0.2, and 0.25, *CF* is set with 0.2, 0.4, 0.6, and 0.8, respectively.

To evaluate the effectiveness and stability of our approach, we set the parameter of *IC* and *CF* according to the results of the first subset of experiment. Therefore, in the second subset of experiments, we fix *IC* and *CF* to 0.18 and 0.4, respectively. In the second subset of experiments, we compare four strategies, i.e., the Amazon strategy, three IBS strategies with different pre-processing methods such as ATV-3, ATV-4, and ATV-5. To analyze the stability of these strategies, we used these four strategies to predict the reputations of all sellers and analyze the variation trends of the predicted MAE of all sellers by increasing the size of time window gradually.

(2) Evaluation criteria

In the experiment over real-life data set, the MAE between Yelp labeled reputation (i.e. $Yelp\_rep(s_j)$) and the estimated window-based reputation (i.e., $Ag\_rep(s_j)$) is calculated over above time windows according to the calculation principle given in Section 4.

$$MAE = \frac{\sum_{s_j \in S} \left| Ag\_rep(s_j) - \frac{Yelp\_rep(s_j)}{5} \right|}{|S|} \tag{13}$$

where $Ag\_rep(s_j)$ denotes the aggregated reputation of seller s_j computed with Equation (8), $|S|$ the total number of sellers in Yelp, and $Yelp\_rep(s_j)$ the labeled reputation of seller s_j that has been accumulated since user account creation.

(3) Results and analysis about various parameters

Table 8 lists the results we get in the first subset of experiment, in which our strategy is assigned with various parameters values of ratings volume, *IC*, and *CF*. In this table, “ratings volume” is the number of ratings extracted according to Figure 7. “Seller MAE” is the average reputation error of all sellers predicted by this strategy (the smaller the better). The ATV-5 data preprocessing method is used to delete reviewers with ratings volume below 5.

According to Table 8, when *IC*=0.25, the MAE value is correspondingly larger than those with *IC* smaller than 0.25. Therefore, *IC*=0.25 is not quite appropriate. Excluding the case of *IC*=0.25, for all ratings volumes such as 40000, 50000, and 60000, the change of *CF* value has little effect on MAE value. When the *IC* value is 0.16-0.2, the experimental results of MAE are slightly better, no matter how the *CF* and ratings volume change. Moreover, when *IC* and *CF* are fixed, no matter how ratings volume changes, the value of MAE can always be stabilized at about 0.08.

From above results, we can conclude that the best combination of parameters is *IC*=0.18 and *CF* 0.2-0.4. In addition, we can also conclude that: even though the numbers of market participants (Table 7) vary dynamically, as long as the platform adopting our strategy has accumulated enough trading experiences (i.e., ratings), it can predict sellers’ reputations stably.

Table 8. MAE Result Display of *IC* and *CF* with Different Values

Ratings volume	*IC*	*CF*	Seller MAE	Ratings volume	*IC*	*CF*	Seller MAE	Ratings volume	*IC*	*CF*	Seller MAE
40000	0.14	0.2	0.077	50000	0.14	0.2	0.081	60000	0.14	0.2	0.077
		0.4	0.078			0.4	0.083			0.4	0.077
		0.6	0.078			0.6	0.085			0.6	0.078
		0.8	0.079			0.8	0.088			0.8	0.079
	0.16	0.2	0.078		0.16	0.2	0.081		0.16	0.2	0.077
		0.4	0.078			0.4	0.083			0.4	0.077
		0.6	0.078			0.6	0.085			0.6	0.078
		0.8	0.078			0.8	0.088			0.8	0.079
	0.18	0.2	0.078		0.18	0.2	0.081		**0.18**	**0.2**	**0.076**
		0.4	0.078			0.4	0.083			**0.4**	**0.076**
		0.6	0.078			0.6	0.086			**0.6**	**0.076**
		0.8	0.078			0.8	0.089			**0.8**	**0.076**
	0.2	0.2	0.078		**0.2**	**0.2**	**0.08**		0.2	0.2	0.078
		0.4	0.078			**0.4**	**0.081**			0.4	0.078
		0.6	0.078			0.6	0.082			0.6	0.079
		0.8	0.078			0.8	0.084			0.8	0.08
	0.25	0.2	0.078		0.25	0.2	0.086		0.25	0.2	0.107
		0.4	0.079			0.4	0.097			0.4	0.151
		0.6	0.081			0.6	0.11			0.6	0.196
		0.8	0.085			0.8	0.124			0.8	0.243

(4) Results and analysis about effectiveness and stability of our approach

Figure 8 shows the variation of MAE over 26 time windows (the ratings volume changes from 10000 to 60000 with increments of 2000). The vertical and horizontal axes represent the sellers' MAE values and the 26 time windows extracted from the recent starting point, respectively. The larger the value of horizontal axis, the earlier the window starts and the older the data samples are. ATV-5, ATV-4, and ATV-3 represent the three MAE trend curves after the data preprocessing of deleting reviewers with ratings volume below 5, 4, and 3, respectively. Amazon is the MAE trend curve calculated using Amazon Platform reputation Method.

From Figure 8, we can see that the four MAE curves decrease with the increase of the ratings volume. Amazon performed better than ATV-5 and ATV-4 when the ratings volume is smaller than 16000. However, it becomes the worst when the accumulated ratings is larger than 30000. Besides, the ATV-3 curve is the best one when the calculated ratings reach a large volume (MAE=0.064). However, its stability is worse than other two curves when the ratings volume increases from 18000 to 40000. The ATV-4 and ATV-5 curves are more stable, and ATV-4 performs better than ATV-5, regardless of the number of ratings. The more the ratings volume, the smaller the difference is. As such, these experiments demonstrate that our strategy is very stable and closer to Yelp's filtering strategy in performance. These results are due to the fact that the IBS strategy has not accumulated enough trading experience, which leads to the inaccuracy of predicting the seller's reputation. Once enough experiences are accumulated, the performance of IBS strategy will increase no matter what kind of pre-processing methods (e.g., ATV-3, ATV-4 and ATV-5) are adopted. Therefore, we can conclude that the IBS strategy is more effective and stable than the Amazon one when enough experiences are accumulated.

Based on the results we get from the two subset of experiments, we can draw following conclusion.

Conclusion 3. Over the real-life Yelp dataset, the IBS strategy is also validity and stability when defending unknown attacks. Therefore, it is appropriate to apply in real e-commerce environment.

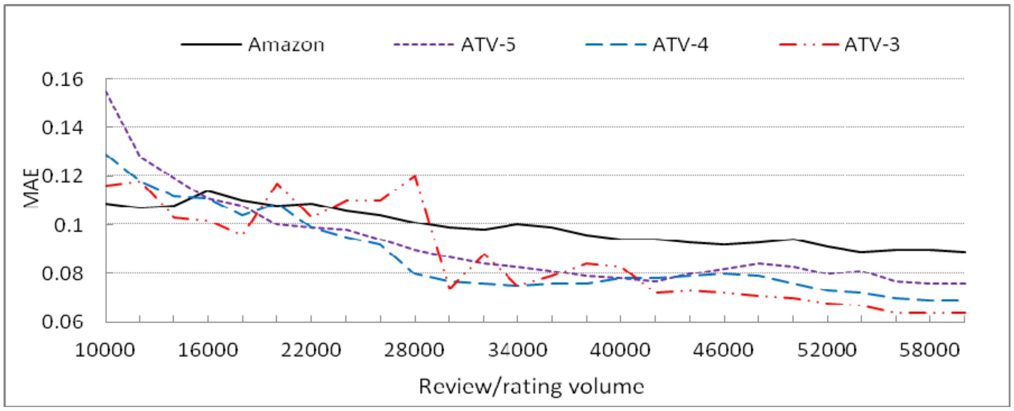


Figure 8 Variation trend of MAE over time windows (*IC*=0.18, *CF*=0.4)

6 CONCLUSIONS AND FUTURE WORK

As electronic markets do not have any prior knowledge about the trustworthiness of the sellers, they can only estimate the reputation of sellers according to reviewers' historical ratings. However, there are no hints on the trustworthiness of the reviewers' ratings either. Though researchers tried to design some filtering mechanisms to make the reputation system more robust against multifarious attacks, there are still great challenges in accurately estimating sellers' reputations and improving the robustness of reputation systems.

In this paper, we present an unsupervised strategy composed of several algorithms. First, a novel nearest neighbor search algorithm is proposed for discovering some rare *lenient* reviewers and *strict* reviewers as benchmark for a cluster based pre-classification of *honest* and *dishonest* sellers (Dellarocas, 2000; Liu, 2014). A key novelty of our nearest neighbor search algorithm is that it only needs to select two special kinds of reviewers (i.e., *lenient* reviewers and *strict* reviewers) based on their statistics characteristics instead of expensive computation of object density or adaptive convergence, which results in tremendously speedup. Besides, to overcome the rareness of *lenient* reviewers and *strict* reviewers, our algorithm imposes a limit to the size of selection of these two kinds of reviewers, which guarantees the convergence of our strategy. Secondly, the valid assumption "once a reviewer is classified as a *strict* or *lenient* one, the reviewer is expected to remain in the same category in the near future" according to the impression theory of social psychology help reduce the need of re-classification. Moreover, based on this assumption, two rules are formalized and used in pre-classifying sellers who have traded with these *lenient* and *strict* reviewers into *honest* and *dishonest* ones. Thirdly, the pre-classified partial sellers serve as benchmarks for evaluating trustworthiness of all the reviewers in the electronic market and dividing them into *honest*, *dishonest*, and *uncertain* ones. These results are finally used in calculating sellers' reputation. We further design two general sets of experiments to evaluate the performance of our approach. Firstly, we simulate a B2B e-commerce market (in which each pair of buyer and seller may have long-term cooperative relationship) under different attacks through four sets of sub-experiments. The second set of experiments are based on real-life Yelp data set (a typical B2C market that most reviewers trade with the sellers only once). Experimental results show that our strategy not only can accurately estimate sellers' reputations, but also can robustly defend against various attacks. Therefore, this strategy opens a new unsupervised research direction in defending reputation attack problems.

This paper takes the behavioral characteristics of the reviewers as the premise of filtering and classification, which implies that the more active reviewers and the more transaction volume, the higher the accuracy of the seller's reputation prediction will be. Moreover, we validate the effectiveness of this strategy against common simulated reputation attacks (such as *AlwaysUnfair*, *Sybil*, *Whitewashing*, *Camouflage*, *Sybil_Camouflage*, and *Sybil_Whitewashing*) over simulated dataset as well as unknown attacks over real dataset. However, for sophisticated and evolutionary attacks, the effectiveness of our model needs further verification. With the running of each e-commerce platform and the setting of some accusation mechanism, a platform can receive more reports and build a black list of poor-reputation sellers. Therefore, we are planning to design a semi-supervised algorithm to accelerate the learning rate from the reviewers' historical experience.

The strategy in this paper is applicable to situations where the quality of products or services provided by sellers is relatively stable. For most of the sellers who sell products (clothing, electrical appliances, etc.), most of the cases satisfy our assumptions because of the less frequent updating of commodity production equipment and the slow and steady progress of production technology. For the part of the sellers to provide services (restaurants, travel, etc), quality of service may change frequently. Under this situation, we can reduce the time windows so that our strategy can adapt to the change of service quality quickly. Further, if the actual situation is completely beyond the scope of application of this strategy, we will take into account the factors of frequent changes in quality in the subsequent research, and design a more widely used strategy. Besides, for electronic market platforms, passively waiting for reports and complaints is inadequate to defend dynamic evolution attacks. It is necessary for platforms to enhance the accuracy of deceptive actions detection on particular sellers and reviewers using their limited resources. Therefore, in the near future, we will study how to allocate detection resource based on the research of Hao et al. (2014, 2015, 2016). We are also interested in protecting the privacy of the reviewers and sellers (Hung et al., 2007) as well as applying this approach under disastrous situations (Chiu et al., 2010). Moreover, we plan to incorporate the impression-based classification method of the strict and lenient persons into the approach of Zhao et al. (2015, 2017) for the application of social media data mining.

ACKNOWLEDGEMENT

This paper is supported in part by the Natural Science Foundation of China (No. 71772107, 71403151, 61502281, 61433012, U1435215), Qingdao social science planning project (No. QDSKL1801138), the training program of the major research plan of the National natural science foundation of China (No. 91746104), the National Key R&D Plan of China (Nos. 2017YFC0804406, 2018YFC0831002), Humanity and Social Science Fund of the Ministry of Education (No. 18YJAZH136), the Key R&D Plan of Shandong Province (No.2018GGX101045), the Natural Science Foundation of Shandong Province (Nos. ZR2018BF013, ZR2013FM023, ZR2014FP011), Shandong Education Quality Improvement Plan for Postgraduate, the Leading talent development program of Shandong University of Science and Technology and Special funding for Taishan scholar construction project.

References

Anderson C.A., & Sedikides C. (1991). Thinking about people: Contribution of Typological Alternative to Associanistic and Dimensional Models of Person Perception. Journal of Personality and Social Psychology, 60: 203-217.

Chiu, D. K. W., Leung, H. F., & Lam, K. M. (2009). On the making of service recommendations: An action theory based on utility, reputation, and risk attitude. *Expert Systems with Applications*, *36*(2), 3293-3301.

Dellarocas. C. (2000). Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior, in: Proceedings of the second ACM conference on Electronic Commerce, 150-157.

Guo, H. (2009). Modeling for reputation computing in c2c communities. *Chinese Journal of Management*.

Hao J., Kang E., Jackson D., & Sun J. (2014). Adaptive Defending Strategy for Smart Grid Attacks. The ACM Workshop on Smart Energy Grid Security, 23-30.

Hao J., Xue Y., Chandramohan M., Liu Y., & Sun J. (2015). An Adaptive Markov Strategy for Effective Network Intrusion Detection. IEEE International Conference on Tools with Artificial Intelligence, 1085-1092.

Hao, J., Kang, E., Sun, J., Wang, Z., Meng, Z., & Li, X., et al. (2016). An adaptive Markov strategy for defending smart grid false data injection from malicious attackers. IEEE Transactions on Smart Grid, DOI: 10.1109/TSG.2016.2610582

Ji, S., Liu, B., Zou, B., & Zhang, C. (2017). An Anti-attack Model for Centralized C2C Reputation Evaluation Agent. *IEEE International Conference on Agents* (pp.63-69). IEEE.

Jiang Siwei, Zhang Jie, & Ong Yew-Soon (2013). An Evolutionary Model for Constructing Robust Trust Networks. In: Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 813-820

Jin, S.H. (2005). Social Psychology. Higher Education Press, Beijing, 102-109.

Jøsang A., & Ismail R. (2002). The Beta Reputation System. In: Proceedings of the 15th Bled Electronic Commerce Conference, 324-337.

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

Jøsang A. (2012). Robustness of Trust and Reputation Systems: Does It Matter? Ifip Advances in Information &Communication Technology, 253-262.

Kerr R., & Cohen R. (2006). Modeling trust using transactional, numerical units. In Proceedings of the 2006 International Conference on Privacy, Security and Trust: Bridge the Gap Between PST Technologies and Business Services (p. 21). ACM.

Liu S., Zhang J., Miao C., Theng Y.-L., & Kot A. C. (2014). An integrated clustering based approach to filtering unfair multi-nominal testimonies. In: Computational Intelligence, 30(2):316-341.

Longman Dictionary of Contemporary English. Reputation. http://www.ldoceonline.com/dictionary/reputation

Luca M., & Zervas G. (2013). "Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud." Harvard Business School Working Papers.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). What yelp fake review filter might be doing? In Seventh International AAAI Conference on Weblogs and Social Media.

Rayana, S., & Akoglu, L. (2015). Collective Opinion Spam Detection: Bridging Review Networks and Metadata. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp.985-994). ACM.

Tran T., & Cohen R. (2004). Improving User Satisfaction in Agent-Based Electronic Marketplaces by Reputation Modelling and Adjustable Product Quality. In Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2. International Conference on Autonomous Agents. IEEE Computer Society, Washington, DC, 828-835.

Wang, X., Ji, S. J., Liang, Y. Q., & Chiu, D. K. W. (2017). An Impression-Based Strategy for Defending Reputation Attacks in Multi-agent Reputation System. International Symposium on Computational Intelligence and Design (pp.383-388). IEEE.

Whitby A., Jøsang A., & Indulska J. (2004). Filtering out unfair ratings in Bayesian reputation systems. In: Proceedings of International Conference on Autonomous Agents and Multiagent Systems Workshop on Trust in Agent Societies (AAMAS).

Xiang S.L. (2006). Social Psychology (2ndед.). China Renmin University Press, Beijing, 108-117.

Zacharia, G., Moukas, A., & Maes, P. (2000). Collaborative reputation mechanisms for electronic marketplaces. *Hicss, 29*(4), 371-388.

Zhang L., Jiang S., Zhang J., & Ng, W. K. (2012). "Robustness of Trust Models and Combinations for Handling UnfairRatings."Ifip Advances in Information & Communication Technology, 36-51.

Zhao, Z., Zhang, Y., Li, C., Ning, L., Fan, J., & Zhao, Z., et al. (2017). A system to manage and mine microblogging data. *Journal of Intelligent & Fuzzy Systems, 33*(1), 1-11.

Zhao, Z., Li, C., Zhang, Y., Huang, J. Z., Luo, J., & Feng, S., et al. (2015). Identifying and analyzing popular phrases multi-dimensionally in social media data. *International Journal of Data Warehousing & Mining, 11*(3), 98-112.