

---

## **Mitigating malicious feedback attacks in trust management systems**

---

Soon Keow Chong\* and Jemal H. Abawajy

Parallel and Distributed Computing Lab,  
School of Information Technology,  
Deakin University,  
Victoria 3217, Australia  
Email: s.chong@deakin.edu.au  
Email: jemal@deakin.edu.au  
\*Corresponding author

**Abstract:** In electronic commerce (e-commerce) environment, trust management has been identified as vital component for establishing and maintaining successful relational exchanges between the trading partners. As trust management systems depend on the feedbacks provided by the trading partners, they are fallible to strategic manipulation of the rating attacks. Therefore, in order to improve the reliability of the trust management systems, an approach that addresses feedback-related vulnerabilities is paramount. This paper proposes an approach for identifying and actioning of falsified feedbacks to make trust management systems robust against rating manipulation attacks. The viability of the proposed approach is studied experimentally and the results of various simulation experiments show that the proposed approach can be highly effective in identifying falsified feedbacks.

**Keywords:** e-commerce; trust; feedback; reputation; unfair rating; trust management; online shopping.

**Reference** to this paper should be made as follows: Chong, S.K. and Abawajy, J.H. (2015) 'Mitigating malicious feedback attacks in trust management systems', *Int. J. Trust Management in Computing and Communications*, Vol. 3, No. 1, pp.1–18.

**Biographical notes:** Soon Keow Chong received her Doctor of Philosophy in Computer Science from Deakin University, Geelong, Australia in 2012. She is currently working as a Researcher and Academic at School of Information Technology, Faculty of Science, Engineering and Built Environment, Deakin University, Australia. Her research interests include trust management, e-commerce, cloud computing, business intelligence and security. She is a member of the IEEE.

Jemal H. Abawajy is a Full Professor at School of Information Technology, Faculty of Science, Engineering and Built Environment, Deakin University, Australia. He is currently the Director of the Parallel and Distributing Computing Laboratory. He is a senior member of IEEE Computer Society; IEEE Technical Committee on Scalable Computing (TCSC). His leadership is extensive spanning industrial, academic and professional areas. He has served on the Academic Board, Faculty Board, IEEE Technical Committee on Scalable Computing Performance Track Coordinator, Research Integrity Advisory Group, Research Committee, Teaching and Learning Committee and Expert of International Standing Grant and external PhD thesis assessor.

## 1 Introduction

Trust management has been receiving attention in various domains such as grid computing (Vijayakumar et al., 2012; Slavomír, 2013), cloud computing (Abawajy, 2011; Habib et al., 2012) and e-commerce (Mäntymäki, 2008; Kim et al., 2009). The goal of a trust management system is to minimise risks and develop mutually beneficial cooperation between trading partners. In e-commerce, the biggest threat is not the security of the personal information but the perceptions of consumers (Schlosser et al., 2006). In order to mitigate the consumers perception problem, reputation systems that involve formal feedback mechanisms and online recommendation agents have been proposed in the literature (Wang and Benbasat, 2005, 2008).

While trust management systems are increasingly being used in e-commerce environments, they are susceptible to tampering with feedbacks (Wang and Benbasat, 2005). Although there have been techniques to encourage trustworthy behaviour (Jøsang and Golbeck, 2009; Li, 2012), the general trend in trust management system is to assume that all feedbacks are accurate and not tampered with. Unfortunately, since the trust management systems rely on the feedback provided by the trading partners, they are fallible to strategic manipulation of the feedback attacks. Therefore, identifying and actioning falsified feedbacks remain an important and challenging issue in trust management field (Mäntymäki, 2008). For example, a small percentage of falsified feedbacks could degrade the accuracy of the trust level, compromise the overall trustworthiness of the participating parties and render the trust management system unreliable. While it is impossible to expect all feedback providers to provide actual feedbacks in an open environment such as e-commerce, it is necessary to have an approach that is able to detect falsified feedbacks to protect the integrity of the trust management system.

In online e-commerce environments, the reliability of the trust management system depends on numerous problems such as falsified and biased ratings (Jøsang and Golbeck, 2009; Pittayachawan et al., 2008). The intention of falsifying rating is to inflate or deflate a seller/buyer's reputation. Falsified feedbacks can compromise the reliability of the trust management systems which seriously affects the trust level of both the buyers and the sellers. Various types of rating attacks against the trust management systems such as ballot stuffing, bad-mouthing, negative discrimination and positive discrimination have been discussed in Duma and Shahmehri (2005) and Jøsang et al. (2007). It has been identified that buyers who falsify feedbacks have similar characteristics to online auction shilling bidders such as a higher bidding frequency to outbid legitimate buyers (Trevathan and Read, 2006, 2007). Similarly, the raters who inflate or deflate feedback will attempt to submit feedbacks frequently. Another common characteristic is that raters who falsify ratings usually have low trust value (Raza and Hussai, 2008). They also tend to usually engage in minimum value transactions to meet the requirements of submitting a rating (Trevathan and Read, 2007). Also, falsified ratings tend to be either significantly lower or higher than the average rates. A rater with a higher trust value is more willing to provide a good rating in order to maintain their reputation (Raza and Hussai, 2008). Thus, a trust management system should have the ability to weigh the ratings of highly credible raters more than those with a low credibility rating (Huynh et al., 2006). Hence, there is a need for a way to identify falsified ratings and to improve the credibility of trust management systems.

To address the problem of strategic manipulation of the ratings, we propose an approach that predicates suspicious feedbacks such that the impacts of such feedbacks on the computation of trust level could be minimised. The key contribution of this paper is the design of an approach that verifies suspicious feedbacks with the aims of identifying and actioning feedback-related vulnerabilities such as those identified in Kerr and Cohen (2007, 2009). The proposed approach combines majority ratings and others parameters such as the amount of transaction and the number of ratings submitted by the same rater in order to mitigate the re-entry and value imbalance issues. Our approach avoids such shortcomings as the normal ratings are separated from suspicious ratings. Also, instead of discarding suspicious ratings, a trust metric scheme is proposed to eliminate the issue of ratings sparse and discourage and reduce the impact of suspicious ratings.

The rest of the paper is organised as follows. Related work and the system model are discussed in Section 2 and Section 3 respectively. The proposed feedback verification mechanism is discussed in Section 4. Strategy analysis and performance analysis is discussed in Section 5 and Section 6. The conclusions and future directions are presented in Section 7.

## **2 Related work**

There are several approaches that evaluate trustworthiness of users based on majority opinion, such as beta filtering feedback (Mäntymäki, 2008). An approach that filters feedbacks that are further away from the majority of ratings is discussed in Whitby et al. (2004). This approach works as long as the majority of ratings are not from a group of raters that tend to falsify their ratings. Another approach that uses beta probability density function to estimate the reputation of a seller as either bad or good is discussed in Jøsang and Golbeck (2009). This approach was later extended such that a feedback is considered to be fair if it falls in the range of lower and upper boundaries among all the ratings Whitby et al. (2004). The limitation of this strategy is that the raters could collude as a group to manipulate the majority ratings. However, majority ratings scheme alone is not sufficient to accurately measure the trustworthiness of a user. We combine a majority rating scheme with three other sources to drive the main factors that influence the credibility of the ratings. The basic idea is that if the received ratings agree with the majority opinion, the past history of the rater is taken into account. This is to eliminate the re-entry issue as it takes time to generate trust value. Therefore, the credibility of the ratings increases if the trustworthiness of the rater is high and decrease otherwise. To eliminate the value imbalance, the transaction value (size) and the frequency of the ratings submitted (the number of ratings submitted for a particular time period) are used. The transaction value and how frequent a rater submit ratings are taken into account as it would prevent the dishonest sellers from building up reputation by cooperating in many small transactions and then cheats in a very large transaction (Abawajy and Goscinski, 2006).

An evidential model of reputation management based on Dempster-Shafer theory is discussed in Duma and Shahmehri (2005). The basic idea of this model is that after a few transactions, unfair ratings provided by participants who have low trust value will carry low weight and therefore will not have much influence in reputation assessment. The assumption on which this model is based is unrealistic. The model assumes that all buyers in the system have provided feedbacks for a given period of time. For example,

new users could be treated as bad users and their feedback will carry less weight in trust assessment. Our approach overcomes this problem as it sets a threshold using the majority rules combined with the parameters of rating value, time, and trust value of both the buyers and sellers.

An approach based on the assumption that low trust value participants are more likely to falsify ratings and should be discouraged low trust value buyers is discussed in Chong and Abawajy (2010). The problem with this method is that it assumes all received ratings are truthful. Users can typically increase their reputation values, for example, by paying other raters to falsify ratings. Our approach avoids such shortcomings as the normal ratings are separated from falsified ratings. Also, instead of discarding falsified ratings, a trust metric system is proposed to discourage and reduce the impact of falsified ratings.

Similarity-based filtering technique such as Jøsang and Golbeck (2009) and Whitby et al. (2004) are frequently used to filter out low similarity ratings that are seen as more trustworthy. One of the problems with this approach is that buyers can submit ratings with the same value as many as possible to a seller. On the other hand, we think this similarity-based filtering technique method is unfair to buyers. Sellers who supply a good quality product may not necessarily provide a further different product of similar quality. Most of the proposed schemes depend solely on a user's previous transaction history without distinguish the relevancy of the services. In a reputation trust system, it is necessary to assess the trustworthiness of sellers according to their service relevancy. The proposed approach addresses this problem.

In our approach, the feedback of the seller is grouped into two subsets referred to as relevant and irrelevant products or services. This allows us to select the right subset of ratings for trust evaluation. In other words, we obtained feedback from the relevant group to calculate the trust value. However, when no relevant ratings are found, ratings that were not relevant to the service are required. Initially a buyer and a seller's reputation were set to 0. The reputation of both the buyer and the seller is updated based on the assessment of ratings received about the transaction. This meant both the buyer and the seller built their reputation slowly based on their good performance which was rated by each other after each transaction. If they fail to meet the regulatory requirements, then their reputation suffers.

A multi-attribute trust management model that incorporates trust, transaction costs and product warranties is discussed in Chong and Abawajy (2010). The new trust management system enables potential buyers to determine the risk level of a product before committing to proceed with the transaction. This is useful to online buyers as it allows them to be aware of the risk level and subsequently take the appropriate actions to minimise potential risks before engaging in risky businesses.

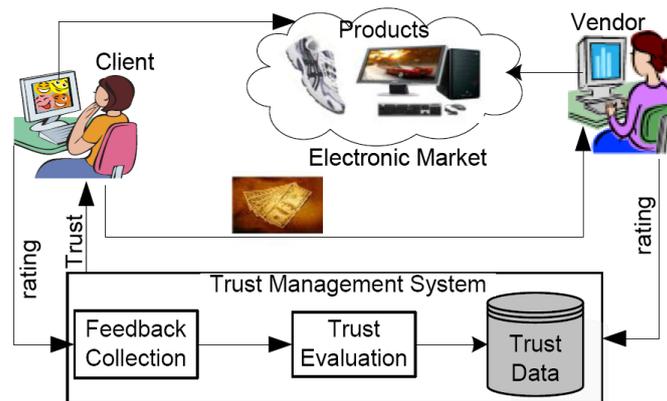
We need to make few observations regarding the proposed techniques. We believe it is possible that malicious participants gain majority ratings through collusion. First, a trustworthy rater is more likely to provide trustworthy feedback (Zhang and Cohen, 2006). Second, the number of transactions is an important factor for comparing the rating in terms of degree of satisfaction among different raters (Abawajy and Goscinski, 2006; Kerr and Cohen, 2007). If the number of ratings submitted to a particular service by the same raters is increased dramatically, these ratings are more likely to be malicious than a scattered rater. It is because a rater could boast the majority rating by submitting as many ratings as possible. Third, the

feedback of a transaction value is another important factor (Gregg and Scott, 2008). A transaction value is the value of a service that a rater paid for. This factor should be incorporated into the evaluation of the quality of feedback for a transaction (Gregg and Scott, 2008). The rationale is that the raters may choose transactions at a lower value of service as often as possible in order to submit ratings in specific period of time which has been identified as costly (Kerr and Cohen, 2007). The behaviour and performance of online market participants' change over time therefore trustworthiness does not remain the same value. Jøsang and Golbeck (2009) pointed out that transactions conducted during a certain period of time can reflect a state of the change in relation to trust. Thus, it is necessary to include time factor to degrade the value as trust value of the sellers and the buyers change overtime. Many approaches, however, assume that the behaviour of both the sellers and the buyers do not change over time and therefore do not take the time factor into account (Gregg and Scott, 2008). In most of the existing approaches, feedbacks suspected or found to be false are usually discarded. In our case, we keep them and evaluate them for later use in determining the trustworthiness of users.

### 3 System model

In this paper, we focus on business to commerce (B2C) model where both the buyer and the seller submit feedback after a business transaction is successfully completed. Figure 1 shows a high-level architecture of an e-commerce system composed of buyers, sellers and products. These components collectively cover most, if not all, phases of e-commerce business transactions such as orders and payments, marketing and distribution. They also enable the sellers to advertise their products and services, deliver goods and services, and provide ongoing customer support. These components also enable the buyers to enquire about products and services, place orders, pay for it and receive goods and services online.

**Figure 1** A generic trust management system (see online version for colours)



We assume that the sellers have website that displays and describes to the customers all of the information about the products, prices, manufacturers, product warranties, etc. The buyers browse the catalogue of the merchandise through a PDA, a mobile phone, etc., to choose one or more products and pay for the order. This acts like an electronic shopping basket and it keeps a record of all of the things that you intend to buy. Once you have chosen all of your items, the payment processing components enable funds to be transferred electronically to anywhere in the world. Your order is then processed by the ecommerce store and sent to you by post. If a successful business transaction occurs, feedback about the service or product is collected from both the customers and the sellers. The collected feedback is then aggregated to produce a trust-level or reputation for both the seller and the buyer. The trust-level is then used to help the potential buyers or sellers to decide whom to trust and subsequently transact with.

Although e-commerce offers enormous opportunities for online trading, the open and anonymous nature of e-commerce presents potential risks to the online buyers. The trust management system will use the feedbacks received from the buyers to determine the trust level of the seller. In this paper, it is assumed that each feedback is uniquely identified by a buyer ID, a product ID, a seller ID, a timestamp and a rating value between 0 and 1. The timestamp is used to verify the originality of the transaction and the actual time the feedback was submitted. Also, a seller/buyer is considered high value if his/her trust level is  $\geq 0.8$  and low value if his/her trust level is  $\leq 0.2$ . Finally, a transaction value is considered high if the transaction amount is  $\geq 0.8$  and the transaction is considered of low value if the amounts to  $\leq 0.2$ .

Transactions in online markets require a great deal of trust among anonymous trading partners. Most online buyers do not have much previous experience dealing with the same trading partner. When there is a lack of personal experience, buyers depend on information from third parties through e-commerce reputation-based trust systems. It is imperative that reliable and effective trust models be in place to enhance the success of e-commerce trust system.

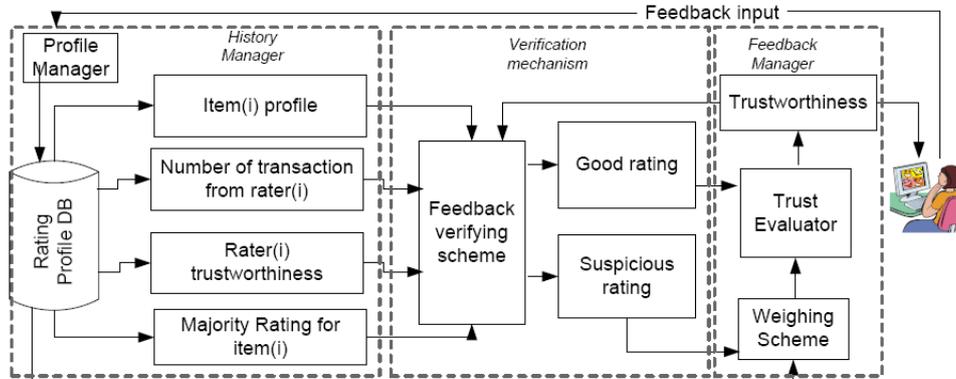
## **4 Feedback verifying strategy**

The reliability of trust system depends largely on the truthfulness of the ratings submitted by the buyers. In this section, we present an approach that will detect falsified feedbacks.

### *4.1 Overview*

The feedback verification mechanism takes the raw feedback and combines it with the information of rater's transaction history which are stored in the transaction record component. A verifying scheme is used to determine if a feedback is genuine or suspicious. Suspicious ratings are maintained for further evaluation to determine the weight of the ratings. Also, both genuine and suspicious ratings have a trust score. Figure 2 shows a high level view of the feedback verifying framework.

The verifier is composed of a 'history manager module' that manages the rating history for all users, a 'feedback verification mechanism module' which is responsible for managing the feedback verification processes and a 'feedback manager module' that is responsible for rating including both good and suspicious ratings.

**Figure 2** Feedback verification framework (see online version for colours)

The system users submit a rating about a service/product after each transaction to the 'profile manager'. The rating contains transaction information including the buyer ID, product ID, seller ID, timestamp (the time at which the ratings were submitted), and the submitted rating with an integer value. This rating can be either a truthful value or a malicious value from the rater. The 'profile manager' manages the profile of all ratings received from users. Profile manager manages this information by using a rating profile database that stores all ratings information including the item profile (information of products), the number of transactions that the rater have done and the majority rating for each item rated. The trust information of both the buyers and the sellers are also available from the rating profile database. All these information will be used by the feedback verifier to verify the credibility of ratings.

The feedback verifier uses a verifying scheme to determine if a feedback is genuine or suspicious one. It first combines all the transaction information including the buyer ID, product ID, and seller ID, timestamp of the rating submitted and the rating value. To separate the suspicious ratings from the genuine ratings, it first examines the majority of the ratings from the raters who have the highest trust value within a given time frame (e.g., a day or a week) depending on the pre-determined system configuration. All ratings within this timeframe fall within the set threshold and are considered good ratings because they satisfy the rules for rating credibility. If the credibility of the ratings is high, it is considered as a good rating otherwise it is considered as a suspicious ratings. The suspicious ratings are then calculated by using the proposed weighing scheme. The feedback manager makes a decision as to how much weight should be given to the ratings based on the information from the 'transaction record' regarding the rater's past transactions. All weighted rating scores are then used by the trust evaluator to determine how trustworthy a rater is. This information is recorded and the trustworthiness of the buyers and the sellers rating is updated. The details of the verifying scheme are discussed in the following subsection.

#### 4.2 *Feedback verification scheme*

The feedback verification scheme uses the k-mean clustering algorithm to group similar ratings together and define the majority rating (Schlosser et al., 2006; Zhang and Cohen, 2006). The scheme assigns each rating in the dataset to the nearest cluster to create the clusters on all reported ratings. The most densely populated cluster is then labelled as the majority cluster and the centroid of the majority cluster is taken as the majority rating. Also, we take into account the service/product value (price) and the quality of the rating which is computed based on the majority ratings, trust value of rater, transaction frequency and transaction value. The trust value of the rater is based on his/her past behaviour and the frequency (number of times) of rating submission. The goal is to verify the suspicious ratings from all of the submitted ratings before determining the credibility of the sellers. Therefore, in order to determine the quality of a rating, we use a trust threshold which designates a minimum value required to establish trust relationship with any entity.

In the first stage, all ratings that fall within the majority cluster are combined with the trust value of the rater, the transaction frequency and the transaction value to determine the credibility of ratings. In this stage, the ratings that are not within the majority cluster are ignored. The trust value of the raters is extracted from the rating profile database if the rater has established the most recent trust value. However, for the raters who do not have trust value assigned to them, their trust value are calculated from their past transaction history. The calculations include time, rating value of transactions, and the frequency of ratings submission. An adjustment scale factor is used in both the transaction value and the frequency value depending on the trustworthiness of the rater. For the raters with a higher level of the trust value, we put less weight on the other two parameters (Velmurugan, 2009; Hung et al., 2012) since low trust value participants are more likely to falsify ratings (Duma and Shahmehri, 2005).

In the second stage, the credibility factors (i.e., trust value of the rater, the transaction frequency and the transaction value) are combined to form a rating verification metric. The filtering mechanism employs this metric to determine the quality of the submitted ratings. The rating verification metric is thus acts as an indirect means to control the performance of the trust management system. The ratings over the threshold value could result in an incorrect trustworthiness value of the rater. However, if the rating value and the value of the verification metric are the same, we consider the rating as a good rating and it is used to evaluate the sellers trust. Otherwise it is considered as a suspicious rating.

In the third stage, all the suspicious ratings are given a value using a weighing metric, which includes a calculation of a rating variance from the value of the good rating. In this stage, the rating from either the majority cluster or away from the majority cluster are combined before the calculation. First, the variance of the rating to the majority rating is calculated. Then the transaction value and the frequency of the ratings are calculated. The weight of a suspicious rating is then assigned based on the rating weighing scheme.

### 5 **Strategy analysis**

This section details each of the analysis features that can be used with the verifying schemes in the feedback verification framework.

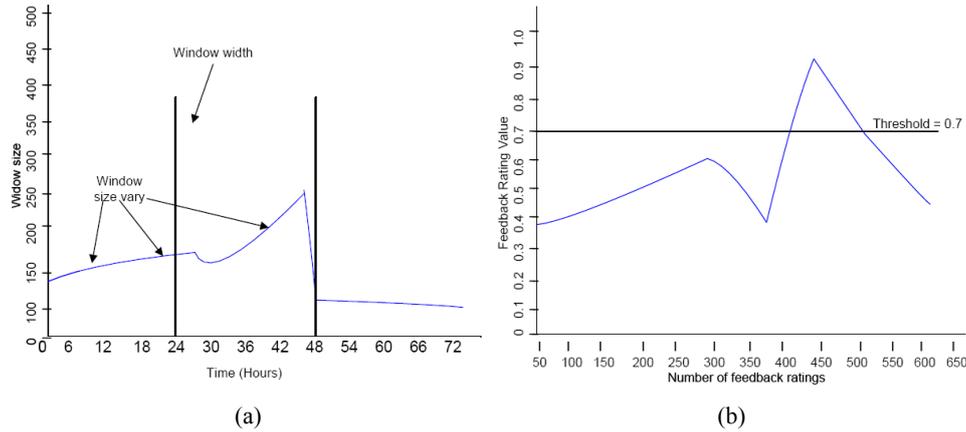
Table 1 provides the description of the symbols used in the rest of the paper.

**Table 1** Symbols and descriptions

<i>Symbol</i>	<i>Description</i>	<i>Symbol</i>	<i>Description</i>
$r$	Rating	$b$	Buyer
$M$	Total feedbacks for a given product/service	$s$	Seller
$f_v$	Rating frequency	$\mathfrak{M}$	Suspicious rating
$\partial$	Weight given to a rating differences	$W$	Window size
$\beta$	Weight given to low value transaction rating	$\tau$	Aging factor
$\lambda$	Weight given to rating frequency	$p$	Product/service
$\mathfrak{S}$	Scale factor for rating submission interval	$t$	Time
$t_i$	Total number of submission of a service	$t_v$	Transaction value
$\mathfrak{N}$	A scale factor for transaction value	$v_i$	Feedback value
$\mathfrak{H}$	A scale factor for frequency	$T_i$	Trust Value
$\Delta t$	Difference between the current time and the recording time of the rating $r_i$	$R$	Ratings
$\Omega$	Difference between a rating submitted by a buyer and the threshold set for a service	$Cr_i$	Credible rating

Let  $r_i$  be a rating submitted by a buyer ( $b_i$ ) for a seller ( $s_i$ ) regarding a product/service ( $p_i$ ) at time ( $t_i$ ). It is assumed that most of the ratings are submitted to a system at different points in time. Therefore, a system will receive  $m$  number of ratings  $R = \{r_i(t_i), r_j(t_j), \dots, r_m(t_m)\} \parallel i = 1, 2, \dots, m$  for a given product. Similarly,  $r_i(t_j)$  is a rating submitted by a rater  $j$  at a time  $i$  for a service ( $p_j$ ) and  $r_m(t_i)$  is a rating submit by a rater  $m$  at a time  $i$  for service ( $p_m$ ).

**Figure 3** (a) Window size and (b) threshold (see online version for colours)

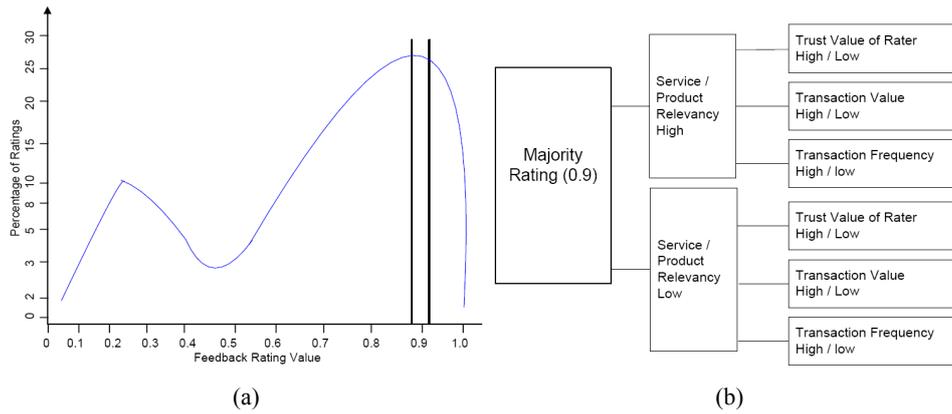


The ratings within a given time  $t_i$  are grouped together using a window of size  $W$ . This window size can be set to a day or a week depending on the needs of the system. The number of ratings in the window is not known in advance and it may vary over time. The window size should be considerably small so that any change in the behaviour of a given

seller is minimal within each element of time. Also, a threshold value is used to differentiate the ratings from the normal ratings. Figure 3(a) and Figure 3(b) show examples of the window size that can be set at every 24 hours for a threshold of 0.8. The threshold is an expected value for the service. This means the ratings is evaluated daily and any rating below or above the threshold are suspicious ratings. In this example, the window size set has a total number of 4,550 ratings. All suspicious ratings become the input to the feedback manager, which determines the degree an individual rating can be trusted.

A rater may rate the same service differently without any malicious intension. Thus, the quality of a rating may change in a number of ways depending on the factors mentioned earlier. Figure 4(a) shows an example of how the quality of a rating is obtained from a majority rating. All ratings received were calculated and the value of 0.9 has the highest number of the total ratings in which the majority ratings is 26% of the total number of the ratings received. Figure 4(b) shows the parameters used to determine the quality of a rating.

**Figure 4** (a) Majority rating (b) Parameters used in determine ratings quality (see online version for colours)



## 5.1 Computing rating credibility

The trust value of a seller is aggregated from the ratings provided by the buyers. The ratings received from the buyers for a seller could be from many different services interacted with. Therefore, the assessment of the trust value of a seller is based on the relevant ratings from the service required.

### 5.1.1 Aging factor

We included an aging factor to degrade the trust value of sellers' overtime (Chong and Abawajy, 2010) as shown in equation (1). The aging factor for a rating is scaled according to the time of the rating received. Let  $\Delta t \mid 0 \leq -\Delta t(r_i) \leq 1$  denote the difference between the current time and the recording time of the rating  $r_i$ . Let  $\tau$  be the aging factor, which is used to decide the level of emphasis given to the past level of trust of the buyer's when calculating the current trust value. Complete distrust is represented by 0 whereas

1 corresponds to the full trust. Similar techniques are used to measure the trust value of a seller. The trust weight for a given rating is determined as shown below:

$$R_i = r_i \cdot e^{\frac{-\Delta t(r_i)}{\tau}} \quad (1)$$

$R_i$  is an indication of the weighted rating assigned by a given rater who has previously conducted business with the seller.

### 5.1.2 Trust value measurement

The trust value is measured by the ratings submitted for a service based on the average of the weighted transaction ratings of that service. Equation (2) shows how the trust value of a seller of service ( $i$ ) is calculated.

$$T_i = \frac{1}{m} \sum_{i=1}^m R_i \quad (2)$$

where  $T_i$  denotes the trust value,  $m$  is the total number of ratings submitted and  $R_i$  is the weighted rating for the seller  $i$ .

### 5.1.3 Transaction values measurement

A transaction value ( $t_v$ ) is the value of a service that a rater paid for and computed as shown in equation (3).

$$t_v = \frac{\sum_{i=1}^v r_i}{\frac{1}{n} \sum_{i=1}^n p_i} \cdot \mathbb{N} \quad (3)$$

The parameters  $r_i$  and  $v$  denote the rater and the transaction value from the rater respectively.  $p_i$  is the average transaction value of the service and  $\mathbb{N}$  is a scale factor used to adjust the transaction value. The weight of the transaction value is measured by its proportion to the value of the transactions. That is the differences between the average transaction value and the rater's total transaction value of a similar service. The larger the difference, the higher possibility that a rater is suspected to be malicious.

### 5.1.4 Transaction frequency measurement

Transaction frequency ( $f_v$ ) is the number of time that the ratings are submitted to a seller by a given rater as compared to the ratings submitted by other raters during a set period of time. Equation (4) shows the measurement of the value of the transaction frequency.

$$f_v = \frac{\sum_{i=1}^k v_i}{\sum_{x=1}^n t_x} \cdot \mathbb{H} \quad (4)$$

where  $v_i$  is the rating value,  $t_x$  denotes the total number of ratings submitted for that service,  $k$  is the total number of ratings from a rater,  $n$  is the total number of ratings submitted for the service and  $\mathbb{H}$  is an adjustment factor scale used to indicate an adjustment value of that service.

### 5.1.5 Credibility of rating measurement

Credibility value ( $Cr_i$ ) is the measurement of a rating submitted for service  $i$ . Equation (5) is to compute the credibility of a rating in a set timeframe (between  $t_1$  and  $t_2$ ). The credibility of a rating is by combining the above parameters and is calculated as follow:

$$Cr_i = \left( r_i \cdot \frac{T_i + f_v + t_v}{3} \right) \quad (5)$$

The result can be used as a trust threshold to compute credibility of the rest of ratings for a particular service in a set timeframe.

## 5.2 Weighing suspicious ratings

As we discussed earlier, suspicious ratings are not discarded, instead a weighing scale is used to weigh the suspicious ratings ( $\mathfrak{M}$ ). In this section, we explain how the weight is determined for suspicious ratings. Once a rating is identified as suspicious, it is then placed in a suspicious group for further weighing. For this purpose, we use the difference between the value of  $Cr_i$  and the suspicious ratings ( $\mathfrak{M}$ ), percentage of the transaction value, the feedback frequency and the suspicious rating value. The suspicious rating is then weighted according to the weight given to all the parameters. An aging scale is later used to scale the value of a suspicious rating. Each suspicious rating is scored between 0 and 1, with a higher value indicating higher suspicion towards a rating.

### 5.2.1 Weighing differences between rating value

The difference between value of  $Cr_i$  and the suspicious ratings ( $\mathfrak{M}$ ) is calculated before weighing. Let  $\Omega$  be the difference between value of  $Cr_i$  and the suspicious ratings ( $\mathfrak{M}$ ).  $r_i$  and  $t_i$  are rating and the total number of submission of service  $i$  respectively.  $\gamma$  is the scale factor and take  $0 \leq \gamma \leq 1$ . The weighing value of  $\partial$  associated with the rating difference is computed as follows:

$$\partial = \mathfrak{M}_{r_i} \cdot \left( e^{-\frac{(\Omega t_i r_i)}{\gamma}} \right) | 0 \leq \Omega t_i r_i \leq 1 | \quad (6)$$

Thus, the higher the value  $\partial$  is, the less weight a rating is given. Note that a rating submitted by a buyer is considered less value rating if the rating deviates from the threshold even though the rating falls within the majority votes.

### 5.2.2 Weighing low value transactions

Ratings submitted by a seller/buyer within a specific time frame is calculated. The total ratings are clustered into individual groups based on the individual rater ID. The transaction value of each rating submitted is also identified. Let  $\eta$  be is the total transaction value submitted by a buyer,  $\mu$  be the percentage of low value transaction and  $l_v$  is the set threshold. The weight of transaction value ( $\beta$ ) is calculated as follows:

$$\beta = \frac{\mu}{\eta} - l_v \quad (7)$$

### 5.2.3 Weighing feedback frequency

Based on the timestamp of every rating submitted, an average time interval of ratings submitted for a seller is obtained during a specific time frame. A scale factor is then used to weigh any rating that is submitted at an abnormal rate of recurrence by a rater as follows:

$$\lambda = e^{-\aleph\vartheta} \mid 0 \leq \aleph\vartheta \leq 1 \quad (8)$$

where  $\vartheta$  is the difference between the average submission time interval of a buyer and the average submission time of buyers to a seller. A scale factor  $\aleph$  is set by the application used to decide how much weight should be given based on the threshold.

### 5.2.4 Weighing trust value

The equation (1) and equation (2) are used to measure the trust value of a seller.

### 5.2.5 Weighing a suspicious rating

The weight of a suspicious rating is based on the four related factors  $\partial$ ,  $\beta$ ,  $\lambda$  and  $T$  are calculated as follows:

$$\mathfrak{M}_{ri} = \frac{\partial w_1 + \beta w_2 + \lambda w_3 + T w_4}{w_1 + w_2 + w_3 + w_4} \quad (9)$$

where  $w_1$  the percentage of participation,  $w_2$  be the percentage of low level transactions,  $w_3$  be the frequency of submissions and  $w_4$  be the trust value of an individual rating. These weighted ratings can be used for supporting the evaluation of the trustworthiness of a seller when there is lack of ratings.

## 6 Performance analysis

In this section, we present the performance analysis of the proposed approach for mitigating malicious feedback attacks against trust management systems through simulation.

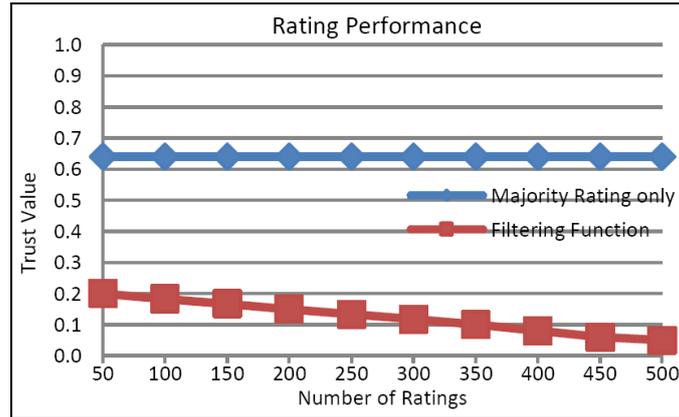
### 6.1 System environment

The performance analysis consists of three sets of experiments. The first set of the experiments, with various setting of parameters, is testing the performance of the credibility of ratings. The second set of experiment is to study the impact of the various weighing scales. We compared the proposed approach with the reputation-based system (Whitby et al., 2004) discussed in Section 2. We focused on the stability of both approaches when the number of untrustworthy sellers increasingly varied in the system. In the simulation, we created 100 sellers selling the same product. The number of buyers and the raters trustworthiness is generated randomly for each buyer and seller in the range of  $[0, 1]$ .

6.2 *Simulation result and discussion*

Figure 5 shows the result of trust value as a function of the number of ratings for the majority ratings proposed credibility filtering function. A large number of malicious raters could affect the majority rating approach as shown in the figure.

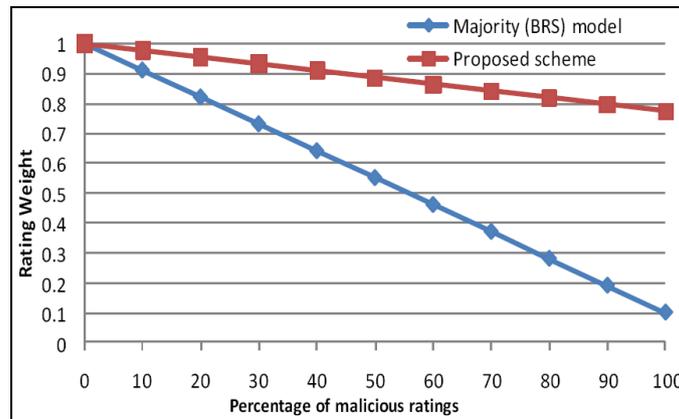
**Figure 5** Comparison of majority and filtering function (see online version for colours)



In the following experiments, we compare the proposed model with the standard Bayesian model using majority votes. In particular, we focus on the stability of both models when the number of untrustworthy raters varied greatly in the system

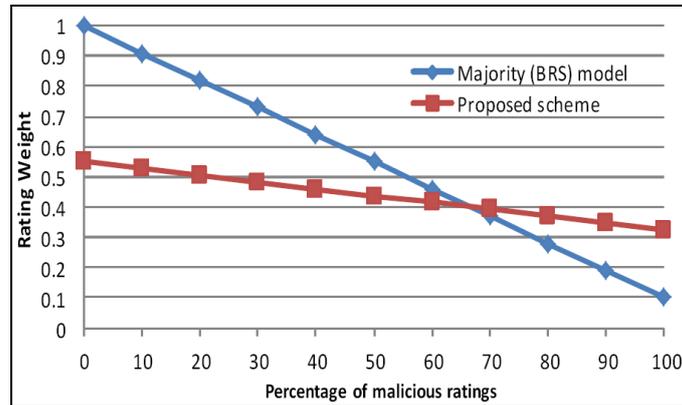
In order to evaluate the performance of our model in different scales, we tested the three parameters  $\beta$ ,  $\lambda$  and  $\varphi$  with different values but were given equal weighing scales. On the other hand, our proposed model not only considers the majority ratings but also the transaction properties discussed in Section 4, which are the value of  $\partial$ ,  $\beta$ ,  $\lambda$  and  $\varphi$ . First, the three parameters were tested with maximum value of 1, then with least value of 0.1. and lastly, one of the parameter  $\beta$  is set as maximum value of 1 and the other two parameter  $\lambda$  and  $\varphi$  0.1. The results are shown in Figures 6, 7 and 8.

**Figure 6** Weighted result (see online version for colours)



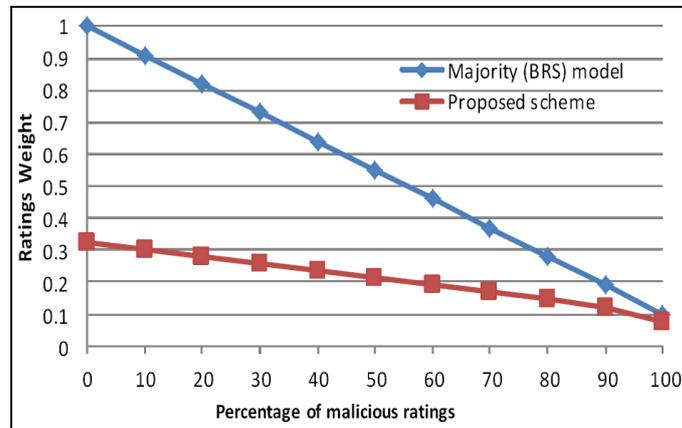
In this experiment, we test the impact of  $\beta = 1$ ,  $\lambda = 1$ ,  $\varphi = 1$  on trust value. Figure 6 shows the results of the experiments. In Figure 6, we can observe that as the percentage of malicious ratings increases, both models show the decrease in the value of ratings.

**Figure 7** Weighted result with minimum value (see online version for colours)



We again observe that the result shown in Figure 7. In this experiment, we test the impact of  $\beta = 0.1$ ,  $\lambda = 0.1$ ,  $\varphi = 0.1$  on trust value. Figure 7 shows the results of the experiments. In Figure 7, we can observe that as the percentage of malicious ratings increases, proposed model shows slight decrease in the value of ratings, whereas majority function decrease significantly in the value of ratings.

**Figure 8** Weighted result with different value (see online version for colours)



In this experiment, we test the impact of  $\beta = 0.1$ ,  $\lambda = 0.1$ ,  $\varphi = 0.1$  on trust value. Figure 8 shows the results of the experiments. The results in Figure 8 shows that our model remains stable although the maximum and minimum values of the parameters are used. When the majority of raters provide ratings, the rating in question will not have a significant influence on the trust value in the proposed model.

Figures 6 to Figure 8 show the experimented results when the values of various parameters changed. The results indicate the proposed weighing metric produces a stable result even though there were increases in malicious ratings. Normally, the results using majority metric remains rigid. From the experimented results, we believe that trustworthiness of the raters and sellers, age of the rating and frequency of the rating are important parameters that should be considered in the design of a rating verifying scheme.

In e-commerce, a rater can register as many identities as he/she likes. It is impossible to know the actual identity of a person or that  $b_1$  is actually  $b_2$ . Most trust models suggest it is safe to have business transactions with those who have higher trust values (Velmurugan, 2009). Although trust value is one important factor, we cannot assume that all trustworthy sellers or buyers provide honest feedback. It is quite often a seller who has been in the e-market business for a long period of time and established a high level of trust who can decide to cheat any given time. The majority function could not predict changes in the behavior of raters behavior and could not indicate the malicious ratings. Furthermore, the proposed model is able to produce results even when a lower number of ratings are received. Trust models that use majority metrics are unable to produce results when ratings are low. Thus, the proposed approach capable for identifying and actioning of falsified feedbacks to make trust management systems robust against rating manipulation attacks.

## **7 Conclusions and future direction**

In this paper, we have discussed the properties and challenges of trading in e-commerce trust management systems. We showed that exiting trust management systems are fallible to strategic manipulation of the feedback attacks and proposed an algorithm to detect suspicious ratings and exclude it from trust calculation in order to improve the reliability of the trust management system. The viability of the proposed approach is studied experimentally and the results of various simulation experiments show that the proposed approach can be highly effective in identifying falsified feedbacks. Thus, improved the accuracy of trust evaluation. We also compared the proposed model against the majority rating model. The result shows that our model is more stable than the majority-based model. Proper trust management will help the users select the provider based on their requirements and trustworthiness.

In the highly dynamic and distributed nature of ecommerce requires that trust management systems be highly scalable in order to efficiently collect feedback and update trust results. Therefore, we believe future research on this topic should include proper scalability and availability techniques which introduces an additional management layer to reach greater sustainability and availability that for trust management systems.

## **Acknowledgements**

The second author would like to acknowledge the support of Maliha Omar and the financial support from the PARADISE Laboratory, School of Information Technology, Deakin University.

## References

- Abawajy, J.H. (2011) 'Establishing trust in hybrid cloud computing environments', *The 10th IEEE International Conference on Trust, Security and Privacy in Computing and Communications (IEEE TrustCom-11)*, 16–18 November 2011, Changsha, China.
- Abawajy, J.H. and Goscinski, A. (2006) *A Reputation-Based Grid Information Service*, Lecture Notes in Computer Science, Vol. 3994, pp.1015–1022, Springer-Verlag, Germany.
- Chong, S.K. and Abawajy, J. (2010) 'Risk-based trust management for e-commerce', in Yan, Z. (Ed.): *Trust Modeling and Management in Digital Environments: From Social Concept to System Development*, pp.332–351, Information Science Reference, New York, USA.
- Duma, C. and Shahmehri, N. (2005) 'Dynamic trust metrics for peer-to-peer systems', *16th International Workshop on Database and Expert Systems Applications*, pp.776–781.
- Gregg, D.G. and Scott, J.E. (2008) 'A typology of complaints about Ebay sellers', *CACM*, Vol. 51, No. 4, p.69–74.
- Habib, S., Hauke, S., Ries, S. and Muhlhauser, M. (2012) 'Trust as a facilitator in cloud computing: a survey', *J. Cloud Comput. Adv. Syst. Appl.*, Vol. 1, No. 1, p.19.
- Hung, M-C., Yang, S-T. and Hsieh, T-C. (2012) 'An examination of the determinants of mobile shopping continuance', *International Journal of Electronic Business Management*, Vol. 10, No. 1, pp.29–37.
- Huynh, T.D., Jennings, N.R. and Shadbolt, N.R. (2006) 'Certified reputation: how an agent can trust a stranger', *AAMAS '06: Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems*, pp.1217–1224.
- Jøsang, A. and Golbeck, J. (2009) 'Challenges for robust of trust and reputation systems', *Proceedings of the 5th International Workshop on Security and Trust Management*.
- Jøsang, A., Ismail, R. and Boyd, C. (2007) 'A survey of trust and reputation systems for online service provision', *Decision Support Systems*, Vol. 43, No. 2, pp.618–644.
- Kerr, R. and Cohen, R. (2007) 'Towards provably secure trust and reputation systems in e-marketplaces', *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multi-agent Systems*.
- Kerr, R. and Cohen, R. (2009) 'Smart cheaters do prosper: defeating trust and reputation systems', *Proceeding AAMAS '09 of the 8th International Conference on Autonomous Agents and Multiagent Systems*, Vol. 2.
- Li, X. (2012) 'Does technology trust substitute interpersonal trust?: Examining technology trust's influence on individual decision-making', *Journal of Organizational and End User Computing*, Vol. 24, No. 2, pp.18–38.
- Mäntymäki, M. (2008) 'Does e-government trust in e-commerce when investigating trust? A review of trust literature in e-commerce and e-government domains', *IFIP International Federation for Information Processing, Towards Sustainable Society on Ubiquitous Networks*, pp.253–264.
- Pittayachawan, S., Singh, M. and Corbitt, B. (2008) 'A multitheoretical approach for solving trust problems in B2C e-commerce', *International Journal of Networking and Virtual Organisations*, Vol. 5, No. 3, pp.369–395.
- Raza, I. and Hussai, S.A. (2008) 'Identification of malicious nodes in an AODV pure ad hoc network through guard nodes', *Computer Communications*, Vol. 31, No. 9, pp.1796–1802.
- Schlosser, A.E., White, T.B. and Lloyd, S.M. (2006) 'Converting web site visitors into buyers: how web site investment increases consumer trusting beliefs and online purchase intentions', *Journal of Marketing*, Vol. 70, No. 2, pp.133–148.
- Slavomír, K. (2013) 'Grid security and trust management overview', *IJCSI International Journal of Computer Science Issues*, Vol. 10, Issue 3, No. 2, pp.1694–0784.
- Trevathan, J. and Read, W. (2006) 'Undesirable and fraudulent behavior in online auction', *SECRYPT'06*, pp.450–458.

- Trevathan, J. and Read, W. (2007) 'A simple shill bidding agent', *Proceedings of the Fourth International Conference on Information Technology*, pp.766–771.
- Velmurugan, M.S. (2009) 'Security and trust in e-business: problems and prospects', *International Journal of Electronic Business Management*, Vol. 7, No. 3, pp.151–158.
- Vijayakumar, V., Wahida Banu, R.S.D. and Abawajy, J.H. (2012) 'An efficient approach based on trust and reputation for secured selection of grid resources', *International Journal of Parallel, Emergent and Distributed Systems*, Vol. 27, No. 1, pp.1–17.
- Wang, W. and Benbasat, I. (2005) 'Trust in and adoption of online recommendations agents', *Journal of the Association for Information Systems*, Vol. 6, No. 3, pp.72–101.
- Wang, W. and Benbasat, I. (2008) 'Attributions of trust in decision support technologies: a study of recommendation agents for e-commerce', *Journal of Management Information Systems*, Vol. 24, No. 4, pp.249–273.
- Whitby, A., Josang, A. and Indulska, J. (2004) 'Filtering out malicious ratings in Bayesian reputation system', *Proceeding 7th Int. Workshop on Trust in Agent Societies*.
- Zhang, J. and Cohen, R. (2006) 'Trusting advice from other buyers in e-marketplaces: the problem of malicious ratings', *Proceedings of the 8th International Conference on Electronic Commerce*.