Paper:

Vulnerability in impression based strategy against bribery behavior

Takuya Otoguro[1] Kazushi Okamoto[1] Atsushi Shibata[2]

Graduate School of Informatics and Engineering, The University of Electro-Communications[1]
Graduate School of industrial Technology, Advanced Institute of industrial Technology[2]
E-mail[1]: {otoguro,kazushi}@uec.ac.jp
E-mail[2]:shibata-atsushi@aiit.ac.jp
[Received 00/00/00; accepted 00/00/00]

Abstract. Electronic marketplaces have adopted centralized reputation systems that aggregate the ratings given by reviewers. This type of systems is used to express the overall evaluation of the products or sellers. Malicious sellers or users may manipulate the ratings to build false evaluations using thise systems. Therefore, approaches to maintain high robustness against such attacks and to correctly evaluate products or sellers are being devised. Simultaneously, new types of reputation attacks are increasing. This study shows that an impression-based defensive strategy is vulnerable against a bribery behavior in which a seller offers bribes to and demands excellent evaluations from reviewers who have made transactions. The results of experiment suggest that this strategy fails to identify approximately half or more of the dishonest sellers under the ratio of bribe-giving manipulators and bribed ratings exceed by 80% and 30% respectively.

Keywords: bribery behavior・centralized reputation system・impression based strategy・reputation attack

1. Introduction

Reputation attacks in electronic markets refer to a behavior in which malicious sellers and reviewers intentionally give positive ratings to cooperative sellers and negative ratings to competitive sellers. In addition, in electronic markets such as Amazon, a bribery behavior in which a seller demands a superior evaluation from a customer in return for sending a voucher or additional items has been confirmed since around 2020 [1]. In bribe manipulations, the sellers do not hire spammers but offer bribes to regular reviewers who have actually made transactions on the e-commerce platform. Therefore, manipulators have the advantage of obtaining the ratings of experienced reviewers regardless of the quality of the stores or items. Although there have been studies about bribery behavior [2,3], there are currently no studies on its effect on impression based strategy (IBS) [4]. IBS is a centralized reputation system that classifies users' behavioral characteristics and estimates the overall evaluation of a seller based on a four step algorithm. Wang et al. [4] applied the impression theory to reviewers in electronic markets and assumed that reviewers who tend to continuously give high or low ratings impress on ussers a lenient or strict impression and this impression is likely to continue in the near future. IBS with unsupervised algorithm has been confirmed to be more robust against reputation attacks than the previously proposed centralized reputation systems [5,6].

This study argues that the assumption in IBS that the seller reliability can be judged from the ratings given by lenient or strict reviewers is not valid under the bribery behavior. This implies that IBS does not consider the existence of the manipulators who bribe reviewers and demand superior ratings. In addition, the bribed ratings affect the subsequent algorithms. As a result, the detection performance of malicious reviewers is degraded under this situation. To demonstrate this argument, it is explained in Section 3 that a manipulator can possiblly create an unfavorable situation against the IBS through bribery. Subsequently, the defense performance of the IBS against AlwaysUnfair [7] is evaluated which is lower than the case in which the manipulator do not exist. The algorithm is evaluated separately using simulations to identify the vulnerable points against the bribe behavior. The simulations have discretely varying ratios of the number of manipulators and the bribed ratings to ensure quantitative evaluation of the impact of the bribery behavior. Based on these evaluations, certain countermeasures that can improve the robustness of IBS have been identified in this study.

2. Related work

2.1. Spam detection in electric commerce

According to Wang et al. [4], there are two major approaches in the spam detection task. The first is filtering methods to detect and eliminate the spam reviews and reviewers. The second is aggregation methods to estimate the overall rating of a seller. The fIltering methods primarily use reviewer behavior and text data [8,9]. This approach often uses the spam review dataset published by Yelp (http://yelp.com) to evaluate the algorithm [9].

Wang et al. [8] proposed an algorithm to identify the cooperative spammer groups and spam reviewers by considering a time series of the reviewer relationship networks. In the aggregation methods, unsupervised learning is the chief approach [5,6,10]. Liu et al.'s algorithm [6] uses reviewer rating behavior and reviewer relationship networks. These methods are often evaluated by creating simulated electronic markets. Furthermore, IBS is an unsupervised learning algorithm and is classified as an aggregation method.

2.2. Framework of IBS

This section provides an overview of IBS. IBS consists of four algorithms and is used to estimate the overall rating of each seller.

Algorithm 1 selects the lenient and strict reviewers by conducting a nearest neighbor search based on the rating mean $\bar{r}_i$ and standard deviation σ_i for each reviewer i. The initial points of the lenient and strict reviewers in the nearest neighbor search are selected respectively from B_{max} = {the first max (5, N× IC)}, B_{min} = {the last max (5, N× IC)}. B_{max}, B_{min} are created by sorting b_i in descending order based on $\bar{r}_i$, where the Impression coefficient (IC) is a parameter that expresses the ratio of selecting the lenient and strict reviewers from the normal reviewers (IC<<1).

Algorithm 2 uses the results of Algorithm 1 and classifies the sellers according to the following two rules.

Rule1: If a seller s_j has traded with $b_i \in B_{strict}$ and b_i has provided a positive rating to s_j (i.e., $r_t^2(b_i,s_j) = 1$) then $s_j \in S_{honest}$;

Rule2: If a seller s_j has traded with $b_i \in B_{Lenient}$ and b_i has provided a negative rating to s_j (i.e., $r_t^2(b_i,s_j) = -1$) then $s_j \in S_{dishonest}$.

The K-point rating is converted to an binary rating

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

where $\overline{K}$ is the overall average of the ratings in the time window T.

Algorithm 3 classifies the reviewers based on the results of Algorithm 2 and according to the following conditions: if $Rep(b_i,S_{honest}) > Rep(B,S) \wedge Rep(b_i,S_{dishonest}) < Rep(B,S)$ then $b_i \in B_{honest}$; else if $Rep(b_i,S_{honest}) < Rep(B,S) \wedge Rep(b_i,S_{dishonest}) > Rep(B,S)$ then $b_i \in B_{dishonest}$; else $b_i \in B_{uncertain}$; where $Rep(B,S), Rep(b_i,S_{honest}), Rep(b_i,S_{dishonest})$ are computed by

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

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

Algorithm 4 estimates the reputation of each seller afiter eliminating the ratings from the dishonest reviewers,

$$Ag_rep(s_j) = (1-w) \times Rep(B_{honest},s_j) + w \times Rep(B_{uncertain},s_j), \quad \ldots \quad (4)$$

$$w = CF \times \frac{|B_{uncertain}|}{|B|}. \quad \ldots \quad (5)$$

where Confidence (CF) is a parameter that expresses the degree of confidence in uncertain reviewers within the market ($0 \leq CF < 1$).

2.3. Reputation attacks and bribe manipulation

There are four primary types of reputation attacks that have been identified so far [7]. AlwaysUnfair (Constant) is the most common attack to measure the performance of these defensive strategies. The attackers always allot low ratings to the honest sellers with good reputations and high ratings to the sellers with bad reputations. Camouflage is an attack in which the attacker initially pretends to be a normal reviewer, and attacks the reputation similar to always unfair after building a good reputation. Whitewashing refers to an attacker's behavior of creating new accounts and repeating the reputation attacks with the purpose of resetting their own reputation. Sybil is an attack in which a malicious reviewer in the market creates multiple accounts and conducts multiple reputation attacks.

Reputation attacks are often performed by sellers in collusion with malicious reviewers. In contrast, in bribe manipulations, a seller offers bribes to and demand a high rating from regular reviewers who have actually carried out a transaction and demand a high rating. Therefore, the manipulators may obtain high ratings from highly reliable reviewers. This manipulation eliminates the effort of creating fraudulent accounts. Furthermore, normal reviewers have the option of refusing any bribes and reporting such sellers depending on the format of the electronic marketplace. Saúde et al. [2] assumed three cases for the bribe manipulations.

Case1: The seller of item i bribes users that already rated the item, so that the buyer changes its rating.

Case2: The seller bribes users that did not buy a product i previously.

Case3: The mixed case.

This study assumes that the manipulators send bribes to reviewers who have already made a transaction with them. Therefore, it corresponds to case 1.

3. Vulnerability in IBS

Wang et al. [4] made the following assumption in IBS: **Once a reviewer is classified as strict or lenient, the reviewer is expected to remain in the same category in the near future.** Based on this assumption, IBS assumes that the low ratings given by the lenient reviewer or the high ratings given by the strict reviewer are reliable, and Algorithm 2 introduced Rule1 and Rule2. However, under the situation where a manipulator exists, it is necessary to consider a case where a seller can obtain a high evaluation regardless of the reviewer's impression. The reliability of the high rating of lenient and strict reviewers is considered to decrease under this situation. From the manipulator's point of view, it is possible to intentionally change their classification on the IBS by using bribing to obtain high ratings from the lenient and strict reviewers.

This study assumes three situations in which a seller is unclassified on IBS.

Situation1: A seller has not been given low ratings by lenient reviewers and high ratings by strict reviewers.

Situation2: A seller has been considered dishonest by the low ratings given by lenient reviewers, however, it has now been given high ratings by strict reviewers.

Situation3: A seller has been considered honest by the high ratings given by strict reviewers, however, it has now been given low ratings by lenient reviewers.

The unclassified in Situations 2 and 3 arise as the sellers are not allowed to have both honesty and dishonesty properties on Algorithm 2. The transaction history of the unclassified sellers cannot determine the honesty of reviewers. Therefore, an increase in the number of unclassified sellers implies an increase in the number of uncertain reviewers in Algorithm 3. As a result, there is a possibility of accepting the ratings from the attackers that should be classified as dishonest because Algorithm 4 aggregates the ratings of both honest and uncertain reviewers following the equation (4). Moreover, the honest sellers in Situation 3 can keep theirs honesty if they have not been given low ratings by lenient reviewers. Therefore, it is necessary to consider the possibility that theirs honesty may be the result of successful bribery of lenient and strict reviewers.

This study confirms that the number of unclassified sellers in Situations 2 and 3 or honest sellers that succeeded in bribing increases under the bribery behavior by applying Algorithm 2 to the simulated market. The experiments in Section4 and Set1 show that it is difficult to judge the reliability of the sellers based on the ratings of lenient and strict reviewers. Set2 investigates how the manner in which the increase in the number of unclassified sellers in Situations 2 and 3 affects the estimation of the overall evaluation of the sellers in Algorithm 4.

Table 1. IBS parameter

	Set1	Set2
IC	0.15	0.15
CF	0.7	0.7

Table 2. Simulation parameter

Rating grades	5
Dishonest reviewers	30
Honest reviewers	70
Dishonest sellers	20
Honest sellers	20
Ratings volume	2500
Manipulator	$0 \sim 80\%$
Bribed ratings volume	$0 \sim 80\%$

4. Experiment

The bribe manipulations following Case 1 are reproduced by simulation. Set1 shows that the number of unclassified sellers corresponding to Situations 2 and 3 or honest sellers that succeeded in bribing increases under bribe manipulations. In Set2, based on the results of Set1, the estimation accuracy of the sellers is evaluated on Algorithm 4. In both Sets 1 and 2, the reputation attacks are AlwaysUnfair (Constant). Further, the parameters of IBS are provided in Table1.

4.1. Datasets

The simulation is based on the Set3 experiment in [4]. The sellers in the simulation are assigned a true reputation value by honesty, $Rel_rep(0 \leq Rel_rep \leq 1)$ as shown in Table 2. Let $Rel_rep(s_j) = 0.8$ for half of the honest sellers, and 1 for the rest. Among the dishonest sellers, set half of them to be $Rel_rep(s_j) = 0.2$ and the rest to be 0.

$s_j \in S_{honest}$

$s_j \in S_{honest}$

Next comes the set of reviewers in the simulation. the honest reviewers do not carry out reputation attacks. Honest reviewers equate to lenient $\cup$ strict $\cup$ normal. The lenient and strict reviewers are selected randomly from the honest reviewers taking, 20% each. The reviewers give ratings to a seller according to the following conditions.

$$r_t(b_i, s_j) = \begin{cases} Rel_rep(s_j) + 0.2 & \text{if } b_i \in B_{lenient} \\ Rel_rep(s_j) & \text{if } b_i \in B_{normal} \\ Rel_rep(s_j) - 0.2 & \text{if } b_i \in B_{strict} \end{cases} \quad (6)$$

The lenient reviewers give $r_t(b_i, s_j) = 1$ for $Rel_rep(s_j) = 1$ and the strict ones give $r_t(b_i, s_j) = 0$ for $Rel_rep(s_j) = 0$. Dishonest reviewers give $r_t(b_i, s_j) = 1$ for $Rel_rep(s_j) \leq 0.2$, and $r_t(b_i, s_j) = 0$ for

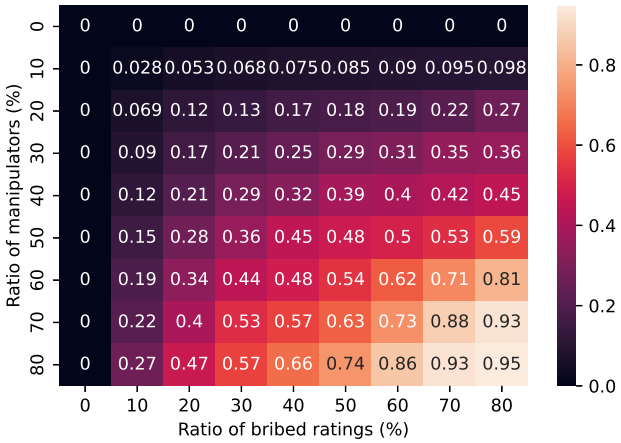


Fig. 1. Ratio of unclassified or honest within the dishonest sellers

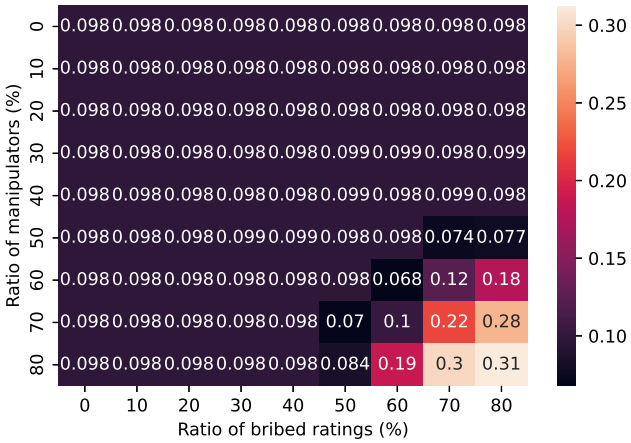


Fig. 2. MAE within the honest sellers

$Rel_rep(s_j) \geq 0.8$. In addition, each seller and reviewer has multiple transactions.

The manipulator is randomly selected from the dishonest sellers in the ratio of $0 \sim 80\%$. The bribed ratings are randomly selected from the transactions satisfying $r_t(b_i, s_j) \leq 0.2$ in the manipulator in the ratio of $0 \sim 80\%$. The selected transactions are changed to $r_t(b_i, s_j) = 1$. Sets 1 and 2 are randomly sampled 100 times, and the average value is used as the evaluation value.

4.2. Evaluation criteria

The estimation accuracy of Algorithm 4 was evaluated with MAE in this experiment.

$$MAE = \frac{\sum_{s_j \in s} \left| Ag_rep(s_j) - Rel_rep(s_j) \right|}{|S|} \quad \ldots \quad (7)$$

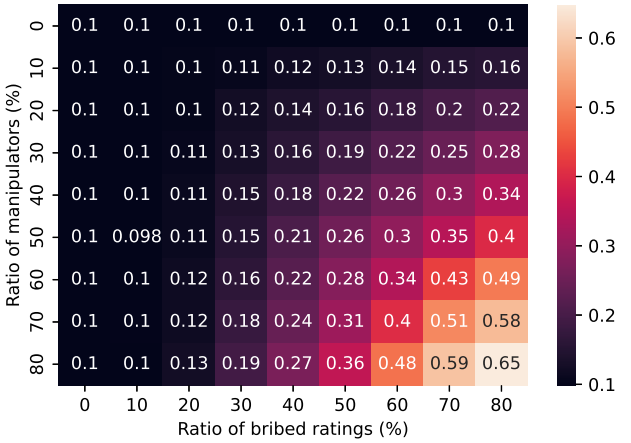


Fig. 3. MAE within the dishonest sellers

5. Experimental results

Fig. 1 shows the ratio of the unclassified sellers that satisfy Situations 2 and 3 or honest sellers that succeeded in bribing in sellers with $Rel_rep(s_j) \leq 0.2$. The IBS fails to identify the dishonest sellers approximately 57% with the manipulator ratio being 80% and the bribed rating being 30%, as depicted in Fig. 1. The result indicates that an increasing proportion of manipulators can falsify their classification on Algorithm 2 with fewer bribed ratings and simple reputation attacks. Therefore, the assumption that the reliability of a seller can be judged by the ratings given by lenient and strict reviewers was demonstrated to be not valid under the bribery behavior with AlwaysUnfair.

Fig. 2 and Fig. 3 show the MAE by seller honesty in Table 2. The MAE of the honest sellers is barely affected unless the ratio of the manipulators and bribed ratings are more than 60%, as shown in Fig. 2. In contrast, Fig. 3 shows a gradual increase of the MAE in the dishonest sellers in proportion to the ratio of the manipulators and bribed ratings. Fig. 4 shows the maximum absolute error for each seller. The protection performance against AlwaysUnfair decreases and the estimation error increases as the unclassified sellers increases, as shown in Fig. 3 and Fig. 4. In particular, for the ratio of each is 30%, the sellers appear with twice the error than without manipulators, as shown in Fig. 4. In addition, Fig. 4 shows the seller appears to have a maximum absolute error of about 0.6 for the ratio of manipulators and bribed ratings are 60% and 50% respectively. However, the sellers with $Rel_rep(s_j) \geq 0.8$ are robust against bribery behavior because the MAE does not exceed 0.4 even when the percentage of each becomes greater than 80% (Fig. 2). In this simulation, five ratings are set. Therefore, these errors of 0.6 and 0.4 is converted to $0.6 \times 5 = 3$, $0.4 \times 5 = 2$. These results indicate that Algorithm 2 is vulnerable to the manipulators, however little impact is observed on the overall

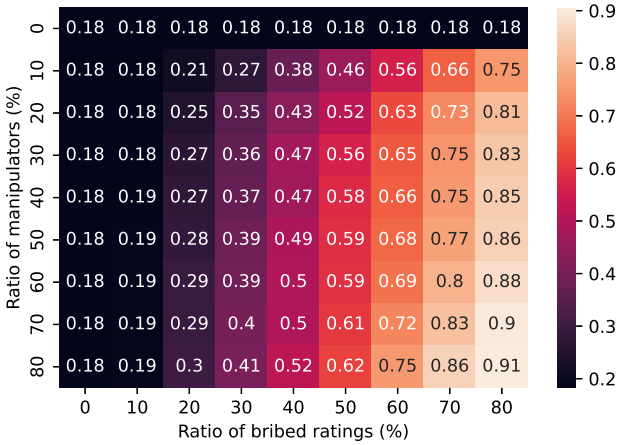


Fig. 4. Maximum absolute error for each seller

market.

In this study, AlwaysUnfair is solely set as the reputation attack and the parameters of IBS is fixed. In addition, the sellers and reviewers uniformly make multi-transactions at random in the simulation. Although these settings and results can be given as one of the challenges in realistic operation of IBS, they are considered limited. Therefore, it is necessary to more investigate cases where single transactions take place and where there is a bias in the volume of transactions by each seller.

6. Conclusion

This study investigated the effect of bribery behavior against IBS and identified an area of vulnerability. IBS is vulnerable to Algorithm 2 against the bribery behavior that dishonest sellers offer bribes to and demand excellent ratings from reviewers who have made transactions. As discussed in Section 3, the sellers are not allowed to have both honesty and dishonesty properties on Algorithm 2. Therefore, Rule1 and Rule2 in the assumption about the reliability of a seller do not validate in the situation where bribery behavior exists, and the number of unclassified or honest sellers judged by IBS increases.

To show the vulnerability in IBS, this study evaluated each algorithm separately by the simulation experiment. Set1 suggests that the ratio of sellers judged as unclassified or honest by IBS increases in proportion to the ratio of manipulators and bribed ratings. This result indicates that if manipulators have the majority of the market, they need relatively few bribed ratings to falsify their classification on Algorithm 2. In Set2, the effect of bribery behavior against Algorithm 4 was investigated based on Set1. The result shows that for the dishonest sellers, MAE increases in proportion to the ratio of manipulators and bribed ratings. In contrast, the respective ratio increases were little effect on MAE in honesty sellers. These results suggest that IBS is vulnerable only to dishonest manipulators.

For future prospects, the economic behavior restrictions of the manipulator and the defensive performance of the IBS against other reputation attacks described in Section 2.3 are investigated. In addition, future work will include increasing the robustness against the bribery behavior by studying the manner in which to deal with unclassified sellers arising in Situations 2 and 3, as shown in this study.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 21H03553, 22H03698.

References:

[1] David Lazarus, “Column: Would you write a 5-star Amazon review in return for a $20 bribe?,” Los Angeles Times, https://www.latimes.com/business/story/2020-02-04/amazon-review-bribe [2022-06-12 accessed].

[2] João Saúde, Guilherme Ramos, Ludovico Boratto, and Carlos Caleiro, “A Robust Reputation-Based Group Ranking System and Its Resistance to Bribery,” ACM Transactions on Knowledge Discovery from Data, vol.16, no.2, pp.1–35, 2022.

[3] Umberto Grandi and Paolo Turrini, “A Network-Based Rating System and Its Resistance to Bribery,” Proceedings of 25th International Joint Conference on Artificial Intelligence (IJCAI-16), pp.301–307, 2016.

[4] Xin Wang, Shu-juan Ji, Yong-quan Liang, Ho-fung Leung, and Dickson K.W. Chiu, “An unsupervised strategy for defending against multifarious reputation attacks,” Applied Intelligence, vol.49, pp.4189–4210, 2019.

[5] Shujuan Ji, Baohua Liu, Benfa Zou, and Chunjin Zhang, “An anti-attack model for centralized C2C reputation evaluation agent,” Proceedings of 2016 IEEE International Conference on Agent (ICA), pp.63–69, 2016.

[6] Siyuan Liu, Jie Zhang, Chuyan Miao, Yin-Leng Theng, and Alex C. Kot, “An integrated clustering based approach to filtering unfair multi-nominal testimonies,” Computational Intelligence, vol.30, no.2, pp.316–341, 2014.

[7] Lizi Zhang, Siwei Jiang, Jie Zhang, and Wee Keong Ng, “Robustness of Trust Models and Combinations for Handling Unfair Ratings“, Proceedings of 6th IFIP International Conference on Trust Management (IFIPTM 2012): Trust Management VI, pp.36–51, 2012.

[8] Zhuo Wang, Runlong Hu, Qian Chen, Pei Gao, and Xiaowei Xu, “ColluEagle: Collusive review spammer detection using Markov random fields,” Data Mining and Knowledge Discovery vol.34, pp.1621–1641, 2020.

[9] Arjun Mukherjee, Vivek Venkataraman, Bing Liu, Natalie Glance, “What yelp fake review filter might be doing?,” Proceedings of 7th International AAAI Conference on Weblogs and Social Media, vol.7, no.1, pp.409–418, 2021.

[10] Jøsang Audun and Walter Quattrociocchi, “Advanced Features in Bayesian Reputation Systems,” Proceedings of 6th International Conference on Trust, Privacy and Security in Digital Business (TrustBus 2009), pp.105-114, 2009.