PAPER Special Section on Data Engineering and Information Management

Detecting Anomalous Reviewers and Estimating Summaries from Early Reviews Considering Heterogeneity

Yasuhito ASANO[†a)], ***Member and*** **Junpei KAWAMOTO**[††b)], ***Nonmember***

SUMMARY Early reviews, posted on online review sites shortly after products enter the market, are useful for estimating long-term evaluations of those products and making decisions. However, such reviews can be influenced easily by anomalous reviewers, including malicious and fraudulent reviewers, because the number of early reviews is usually small. It is therefore challenging to detect anomalous reviewers from early reviews and estimate long-term evaluations by reducing their influences. We find that two characteristics of *heterogeneity* on actual review sites such as Amazon.com cause difficulty in detecting anomalous reviewers from early reviews. We propose ideas for consideration of heterogeneity, and a methodology for computing reviewers' degree of anomaly and estimating long-term evaluations simultaneously. Our experimental evaluations with actual reviews from Amazon.com revealed that our proposed method achieves the best performance in 19 of 20 tests compared to state-of-the-art methodologies.

key words: *early reviews, heterogeneity, data mining, bipartite graph*

1. Introduction

How can we find evaluations of products from people's reviews as quickly as possible? Various web sites such as online shopping sites, movie databases, and online recipes collect people's reviews. Users of such web sites are influenced by those reviews when they make decisions such as buying a product or not. It is therefore important to analyze reviews and identify correct evaluations of products rapidly to take advantage of reputation. However, early reviews posted on those web sites shortly after products come onto the market can be affected easily by anomalous reviewers, including malicious and fraudulent reviewers, because early reviews are few. After reviews become sufficiently numerous, a few anomalous reviews can be simply ignored. Conventional intrusion detection systems can find them if huge numbers of anomalous reviews are posted. Most web sites also have a system to report anomalous reviews. Administrators can delete them after obtaining a certain number of reports. Therefore, it is a challenging problem to detect anomalous reviewers from early reviews and reduce such influences of anomalous reviewers to estimate long-term summary evaluations, which closely reflect the opinions of ordinary people.

In most cases, early reviews are reviews for new products. For that reason, knowledge learned from old products is ineffective to analyze those new reviews. That fact suggests that context-free and unsupervised methodologies are expected to analyze early reviews. Fraud Eagle [1] and FRAUDAR [2] are such methodologies. They have been proposed to detect anomalous reviewers and fraudulent reviewers. Those algorithms classify reviewers into two categories: fraudulent or not. However, binary classification is insufficient in typical rating systems such as the five star rating system because a review takes various values. Therefore, the degree of a reviewer's anomaly should take various values as well.

Computing the degree of anomaly has also been studied, but such studies do not address heterogeneity in real reviews. We have investigated reviews posted on real web sites such as Amazon.com, which has revealed the importance of considering two characteristics of heterogeneity:

1. the relation between reviews' degree of anomaly and rarity is neither obvious nor linear; and
2. the numbers of reviews of products and their dispersion are widely diverse.

As described herein, we introduce a novel methodology that can accommodate heterogeneity. Moreover, it achieves both detection of anomalous reviewers and estimation of long-term summaries simultaneously. Our methodology has two key ideas to address the heterogeneity as described above: *deviation rarity* measures how rarely large the deviation of a rating is to solve heterogeneity of the first kind; and *controversiality* measures how important a product is for computing the degree of anomaly of reviewers by calculating the number of product reviews and their variance to solve the second kind of heterogeneity.

We compared the performance of our proposal with two state-of-the-art algorithms: Fraud Eagle and FRAUDAR. For this evaluation, we used actual reviews collected from Amazon.com and synthetic reviews. Then we investigated how correctly each methodology can detect anomalous reviewers and can estimate long-term summaries from early reviews. We used four metrics and five datasets, about 20 kinds of evaluations in all. Our proposal achieved the best results in 19 of them.

We present a summary of our contributions below.

- We address the novel problem of detecting anomalous

Manuscript received June 20, 2017.
Manuscript revised November 6, 2017.
Manuscript publicized January 18, 2018.

[†]The author is with Graduate School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.
[††]The author is with Graduate School and Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka-shi, 819–0395 Japan.
a) E-mail: asano@i.kyoto-u.ac.jp
b) E-mail: junpei.kawamoto@acm.org

DOI: 10.1587/transinf.2017DAP0006

reviewers and estimating long-term summaries simultaneously from early reviews.
- We clarify the importance of heterogeneity in anomalous reviewer detection on actual review sites.
- We propose an unsupervised context-free method by combining the principle of repeated improvement with two novel key ideas to accommodate the heterogeneity: *controversiality* and *deviation rarity*.

2. Related Work

Anomalous reviewers is a broad concept. It includes not only fraudulent reviewers such as social spammers and crowd turfing workers, but also early adopters and experts. Opinions of experts are often considerably useful for products targeting heavy users, although they can be inappropriate for products targeting light users. For example, if a person desires to buy a digital camera for daily use, then reviews from novices might be more suitable than those from experts because experts usually recommend cameras that are too expensive or too heavy for casual users.

Most previous methodologies related to detecting anomalous reviewers aim to find spam reviewers. They depend on large amounts of training data or review texts [3]–[6]. However, early reviews, posted on online review sites shortly after products are on sale, have insufficient amounts of training data. It is not readily apparent that knowledge learned from reviews of old products can function for reviews of new products. Consequently, context-free and unsupervised methods are desired for analyzing early reviews.

Fraud Eagle [1] and FRAUDAR [2] are context-free and unsupervised state-of-the-art methodologies for detecting anomalous reviewers and fraudulent reviewers. Fraud Eagle assumes that reviewers, reviews, and products are classifiable into two categories: reviewers are honest or fraudulent, reviews are positive or negative, and products are good or bad. Roughly speaking, we can say a reviewer who posts many negative reviews to good products is fraudulent with high probability, based on those assumptions. Fraud Eagle models reviewers, reviews, and products using a graph structure and introduces a classification system based on loopy belief propagation. However, Fraud Eagle must be extended to classify reviews and products into at least five categories to apply it to typical web sites that employ a five-star rating system. This extension is neither easy nor trivial because it is not readily apparent how anomalous a reviewer is who assigns four stars to products that mostly receive three stars. Such extension persists as an open problem. FRAUDAR, which is also a graph-based algorithm, is designed to find camouflaged fraudulent reviewers who post malicious reviews to their targets but post reviews to other products in a manner similar to normal reviewers. Camouflaged fraudulent reviewers generally have obvious attack targets. FRAUDAR can detect them by relying on specific graph structures. However, anomalous reviewers more generally exhibit anomalous behavior without having a readily identifiable target. Because the graph structures of those anomalous reviewers might resemble those of normal reviewers, FRAUDAR does not seem to distinguish them.

Detection of spam review groups [7]–[9] is a kind of context-free and unsupervised studies to find anomalous reviewers. It is specialized to identify numerous colluding spammers. This approach is promising for massive crowdturfing, although it would not be suitable for early reviews, which consist of a few reviews. Xie et al. [10] proposed a method that is specialized to detect spammers with a single review. In other words, their method does not address anomalous reviewers with multiple reviews. Their method would have been more useful if it were combined with other methods, even our method. Feng et al. [11] proposed a method for detecting anomalous reviewers using statistical behavior data of such reviewers. Their method regards reviewers who have posted at least 10 reviews as trustworthy, and computes the difference between the evaluation of trustworthy reviewers to a product and that of others. If the difference is large, then the product is regarded as suspect. This method assumes numerous reviews for a product. However, some reviews, such as early reviews, usually do not include reviews posted by such trustworthy reviewers. Consequently, their method cannot incorporate heterogeneity in this problem on actual review sites sufficiently.

Several context-free and unsupervised methods exist to find anomalous reviewers. A representative approach is an application of the principle of repeated improvement [12] to a bipartite graph model of a review site. In the graph, one vertex set represents reviewers. The other vertex set represents products. Each edge, which represents a review for a product from a reviewer, has a quantified value that we call a *rating*. Some review analyses have adopted repeated improvement [13]–[15]. However, these methods do not fully use *heterogeneity* in the anomalous reviewer detection problem on actual review sites. The opposite situation arises in a ranking of reviewers according to leniency on peer review of academic papers [13]. Every paper is assigned to the same number of reviewers. All reviewers are fundamentally naive. Strict reviewers and lenient reviewers are symmetric. However, actual review sites have some products that have attracted many reviews. Some have few reviews, including early reviews. Therefore, actual review sites have a high degree of heterogeneity. In addition, anomalous reviewers and normal reviewers are asymmetric. Anomalous reviewers are far fewer than normal reviewers, according to their nature. Therefore, some heterogeneity exists in anomalous reviewer detection problems.

We finally discuss sentiment analysis. Several studies have been undertaken to quantify polarities or sentiments represented in text documents including review sentences [16]–[19]. Although our method uses ratings in reviews only as numerical values, such quantified polarities or sentiments are applicable with no modification if only we can define the distance separating them.

Table 1 Summary of notation.

Symbol	Definition
R	Set of reviewers ($\{r_1, r_2, \cdots, r_n\}$)
P	Set of products ($\{p_1, p_2, \cdots, p_m\}$)
E	Set of reviews, i.e. edges
rate(r, p)	Rating of reviewer r to product p
a(r)	Anomalous score of reviewer r
s(p)	Summary of reviews assigned to product p
deviation(r, p)	$\lvert\text{rate}(r, p) - \text{s}(p)\rvert$
P_r	Set of product reviewer r reviews
R_p	Set of reviewers reviewing product p

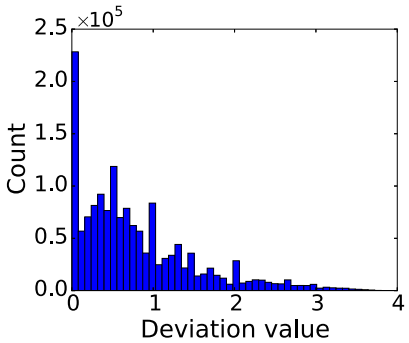

Fig. 1 Distribution of deviation values in the Amazon dataset used in our experiments.

3. Preliminaries: Bipartite Graph Model

We represent reviewers, products, and reviews using a bipartite graph model.

Definition 1 (Bipartite graph): Let $G = (R, P, E, \text{rate})$ be a bipartite graph; $R = \{r_1, r_2, \cdots, r_n\}$ is the set of nodes representing n reviewers, $P = \{p_1, p_2, \cdots, p_m\}$ be the set of nodes representing m products, each edge $(r, p) \in E$ denotes the review of $r \in R$ to $p \in P$, and the rating of edge (r, p) is denoted by rate(r, p).

We assume that a reviewer is able to post, at most, one review for a product and that each rating can be normalized in $[0, 1]$. We assign each reviewer r anomalous score a(r) defined in $[0, 1]$. Simultaneously, each product p has summary s(p) of review ratings. Because we assume that ratings are in $[0, 1]$, summaries are also in $[0, 1]$. We also define three notations to facilitate later discussion: deviation(r, p) denotes the deviation of rate(r, p) to s(p), i.e. , $|\text{rate}(r, p) - \text{s}(p)|$. P_r denotes the set of products reviewer r reviews, i.e., $P_r = \{p|(r, p) \in E\} \subset P$, and R_p denotes the set of reviewers who review product p, i.e., $R_p = \{r|(r, p) \in E\} \subset R$. Table 1 presents a summary of notation.

4. Repeated Improvement Considering Heterogeneity

In this section, we first introduce the outline of our method based on the bipartite graph model and the principle of repeated improvement. We then introduce the two key ideas of *deviation rarity* and *controversiality* to handle the heterogeneity, and finally present the details of our algorithm, *repeated improvement considering heterogeneity (RIH)*.

4.1 The Outline of RIH

The outline of RIH is the following:

1. initializing the summary of each product to the mean of its ratings,
2. updating the anomalous score of each reviewer r by accumulating *partial anomality* of reviewer r to each product $p \in P_r$, which is the degree how rare the rating of r to p is.
3. updating summaries of products considering the anomalous scores,
4. repeating steps 2 and 3 until reaching termination conditions.

RIH updates the anomalous scores and summaries alternately and repeatedly. This is based on a bipartite graph model and the *principle of repeated improvement* [12] used in many graph analyses such as HITS [20]. We adopt the following assumptions [14] between anomalous scores and review summaries to apply the principle for our case.

Assumption 1: i) An anomalous reviewer is expected to be a reviewer posting a different rating to each of many products from their review summary. ii) A review summary is expected to approximate the ratings of normal reviewers.

In the step 3, we simply use a weighted average of ratings as the summaries to consider the anomalous scores of reviewers. More precisely, we use $1 - \text{a}(r)$ as the weight of reviewer r. If anomalous scores are high for some reviewers, then ratings from those reviewers have only slight effects on the summaries.

Definition 2 (Weighted summary): The weighted summary of product p, s(p), is defined as

$$\text{s}(p) = \frac{\sum_{r \in R_p} (1 - \text{a}(r)) \times \text{rate}(r, p)}{\sum_{r \in R_p} 1 - \text{a}(r)}.$$

The most difficult part is the step 2. A straightforward approach to calculate and accumulate partial anomality does not work well because of two kinds of heterogeneity, named *heterogeneity (I) and (II)*, on actual review sites. We will introduce two key ideas to deal with them below.

4.2 Deviation Rarity

The idea behind the deviation rarity is to compute how rare given rating rate(r, p) to s(P) is by considering the distribution of deviation(r, p), because the heterogeneity (I) caused by a bias of the distribution.

Figure 1 illustrates the distribution of deviation which uses the mean of ratings of each product p as s(p) (i.e. $\forall r, \text{a}(r) = 0$) in the Amazon dataset used in our experiments. The simplest approach to measure how rare given rating rate(r, p) is by the difference between the value of

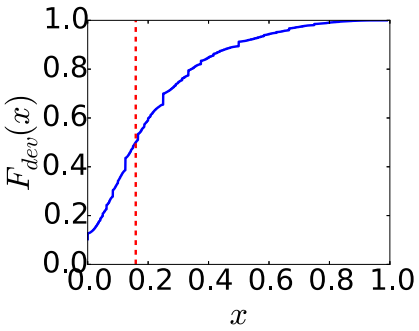

Fig. 2 CDF of the deviation distribution in the Amazon dataset used in our experiments.

deviation(r, p) and the average of deviations for all r and p; a positive difference is expected to represent a rare rating, and vice versa. However, this approach penalizes normal reviewers excessively. Because the average deviation is much smaller than the median, the range of negative values (corresponding to common ratings) of is narrower than that of positive values (anomalous ratings). We designate that fact as *heterogeneity (I)*.

This heterogeneity becomes problematic under circumstances in which anomalous reviewers account for most early reviews of a product. In this situation, the summary of a product initially becomes close to the ratings of anomalous ones. As a result, each normal reviewer corresponding to the minority for the product receives a positively large difference. The original purpose of the repeated improvement is to propagate the influences of the products to which the normal reviewers assigning ratings and to reverse this result. However, the heterogeneity explained above prevents the purpose. The positively large difference could not be alleviated by the influences of a few products, although the normal reviewer receives negatively small differences from them.

To handle *heterogeneity (I)*, we have to consider judge whether a given deviation(r, p) is rarely large considering the distribution of deviation in a probabilistic way so that we can deal with normal reviewers and anomalous ones equally. Therefore, we utilize the cumulative distribution function (CDF) of the deviations. Figure 2 presents CDF of the deviation distribution shown in Fig. 1. The CDF function $F_{\mathrm{dev}}(x)$ is equal to the ratio of the number of ratings whose deviations are at most x to the number of all ratings. Then, we can measure how rare deviation(r, p) is compared to the mean $\bar{\delta}$ of deviations by using this function. Note that $\bar{\delta}$ is depicted as a red dotted line and $F_{\mathrm{dev}}(\bar{\delta})$ is depicted as the cross point of the dotted line and the blue curve in Fig. 2; $F_{\mathrm{dev}}(\bar{\delta}) = 0.5$ theoretically because it is the ratio of ratings whose deviations are at most the mean.

Definition 3 (Deviation rarity): Letting $\mathrm{dr}(r, p)$ be the deviation rarity of the rating from reviewer r to product p, it is defined as

$$\mathrm{dr}(r, p) = F_{\mathrm{dev}}(\mathrm{deviation}(r, p)) - F_{\mathrm{dev}}(\bar{\delta}),$$

where $\bar{\delta} = \frac{1}{|E|} \sum_{(r,p)\in E} \mathrm{deviation}(r, p)$.

If $\mathrm{dr}(r, p)$ is positive, then x is rarely large compared to $\bar{\delta}$; if $\mathrm{dr}(r, p)$ is negative, then x is commonly small compared to $\bar{\delta}$. We will explain how to use the deviation rarity in the calculation of partial anomality in Sect. 4.4.

4.3 Controversiality

Deviations obtained from products cannot be treated equivalently for computing the partial anomality because the number of reviews and the distribution of ratings to a prodcut are usually different from those to another product. To handle this heterogeneity (II), we should consider how controversial the ratings to each product is.

Let us consider the following three typical cases sharing a common situation: reviewer r_1 posts one star to product p_1 and the mean of ratings to p_1 is 3.0. However, the partial anomality of r_1 concerning p_1 would be different in the three cases because the number of reviews to p_1 and their distribution are different.

(Case 1) Most ratings to p_1 are close to three stars, and as a result the mean is 3.0. The rating of r_1 is quite rare, and r_1 is suspected as anomalous with regard to p_1. In this case, the reviews can be regarded as almost not controversial.

(Case 2) About half of reviewers posted one star to p_1, and another half posted five stars; the mean is 3.0 although few reviewers posted three stars actually. In this case, the review of r_1 does not seem anomalous, and the reviews are regarded as controversial.

(Case 3) There is another reviewer r_2 other than r_1 posted five stars to p_1, then the mean is 3.0. The variance of ratings is quite large, although this case cannot be controversial as much as the case 2; a few additional reviews could change the variance drastically.

From the observation of the three cases above, controversiality has to reflect the number of ratings posted to the product and their variance. Because we employ the weighted summary $\mathrm{s}(p)$ (Definition 2), we should use it to compute their variance instead of the mean of the ratings. Furthermore, we should use anomalous scores as weights to compute the variance in order to maintain the consistency with the calculation of summary in Definition 2.

Definition 4 (Weighted variance of ratings): The weighted variance of ratings posted to product p, $\mathrm{wvar}(p)$, is

$$\mathrm{wvar}(p) = \sum_{r\in R_p} \frac{(1-\mathrm{a}(r))(\mathrm{rate}(r, p)-\mathrm{s}(p))^2}{|R_p|}.$$

Similarly to the deviation rarity, we use CDF to evaluate how distant the weighted variance for a product is from the average of weighted variances for all products. Let $F_{\mathrm{wvar}}(x)$ be the ratio of the number of products for which

weighted variances are at most x to the number of all the products. We then define controversiality using a sigmoid function so that the value is within $[0, 1]$.

Definition 5 (Controversiality): The controversiality of product p, $\mathrm{cont}(p) : P \to [0, 1]$, is

$$\mathrm{cont}(p) = \begin{cases} 0.5 & \text{if } |R_p| = 1, \\ 1 - \left(1 + |R_p|^{\alpha(F_{\mathrm{wvar}}(\mathrm{wvar}(p)) - 0.5)}\right)^{-1} & \text{otherwise,} \end{cases}$$

where α is a positive parameter to adjust the range of controversiality (see Sect. 5.4).

One can examine this definition of controversiality is consist with the observations of the three cases above. $\mathrm{cont}(p)$ is small because $|R_p|$ is big but $\mathrm{wvar}(p)$ is small in case 1, $\mathrm{cont}(p)$ is big because both $|R_p|$ and $\mathrm{wvar}(p)$ are big in case 2, and $\mathrm{cont}(p)$ takes a middle value because $|R_p|$ is too small in case 1.

4.4 Anomalous Score

Based on the controversiality and deviation rarity, we introduce partial anomality $\mathrm{pa}(r, p)$, which measures how anomalous the review which reviewer r posts to product p is.

We first consider an ideal case in which the controversiality of every product is equal to 1, i.e., every product is equivalent. Then, the partial anomality $\mathrm{pa}(r, p)$ should be an increasing linear function of the deviation rarity. That is, if the deviation rarity is positively large, then the partial anomality $\mathrm{pa}(r, p)$ should be high because the rating is quite rarely large. In contrast, if the deviation rarity is negatively large, i.e., the rating is quite commonly small, then $\mathrm{pa}(r, p)$ should be low.

Let us then turn to actual review sites having various values of controversiality. If $\mathrm{cont}(p)$ is high, we should not make $\mathrm{pa}(r, p)$ large even if the deviation rarity is positively large. Furthermore, when we calculate the anomalous score of r by integrating $\mathrm{pa}(r, p)$, $\forall p \in E_r$, we should not assign a higher weight to $\mathrm{pa}(r, p)$ than $\mathrm{pa}(r, p')$ for another product p' which r posts reviews if $\mathrm{cont}(p) > \mathrm{cont}(p')$. In this way, the controversiality represents how important a product is to compute $\mathrm{pa}(r, p)$ and $\mathrm{a}(r)$. Therefore, $(1 - \mathrm{cont}(p))$ should work as an amplifier of the deviation rarity and a weight for calculating $\mathrm{a}(r)$. The following definitions of $\mathrm{pa}(r, p)$ and $\mathrm{a}(r)$ implements the idea explained here naturally.

Definition 6 (Partial anomality): The partial anomality of the review which reviewer r posts to product p, $\mathrm{pa}(r, p) : R \times P \to [0, 1]$ is

$$\mathrm{pa}(r, p) = \left(1 + e^{-\beta \times (1 - \mathrm{cont}(p)) \times \mathrm{dr}(r, p)}\right)^{-1},$$

where β is a parameter that is used to adjust the effects of controversiality and deviation rarity.

Definition 7 (Anomalous score): Letting $\mathrm{pa}(r, p)$ be the partial anomality of reviewer r and product p, the anomalous score of r, $\mathrm{a}(r)$, is

$$\mathrm{a}(r) = 1 - \left(1 - \frac{\sum_{p \in P_r} (1 - \mathrm{cont}(p)) \mathrm{pa}(r, p)}{N_r}\right)^{\gamma},$$

where $\gamma \geq 1$ is a given parameter to control effects from partial anomality of low-controversiality products.

Algorithm 1 RIH.

Require: Graph $G = (R, P, E, \mathrm{rate})$, parameters α, β, γ.
  for each product p **do**
    $\mathrm{s}(p) \leftarrow \sum_{r \in R_p} \mathrm{rate}(r, p)/|R_p|$
  end for
  repeat
    Compute CDF $F_{\mathrm{dev}}(x)$ and $F_{\mathrm{wvar}}(x)$.
    for each reviewer r **do**
      $\mathrm{a}(r) \leftarrow 1 - \left(1 - \frac{\sum_{p \in P_r} (1 - \mathrm{cont}(p)) \mathrm{pa}(r, p)}{N_r}\right)^{\gamma}$
    end for
    for each product p **do**
      $\mathrm{s}(p) \leftarrow \frac{\sum_{r \in R_p} (1 - \mathrm{a}(r)) \times \mathrm{rate}(r, p)}{\sum_{r \in R_p} 1 - \mathrm{a}(r)}$
    end for
  until change of all $\mathrm{a}(r)$ and $\mathrm{s}(p)$ is negligible.
  return $\mathrm{a}(\cdot)$ and $\mathrm{s}(\cdot)$.

4.5 Algorithm

Algorithm 1 summarizes our **RIH**, where a left arrow ($\leftarrow$) denotes an assignment of a new value. **RIH** receives a part of a bipartite graph G i.e., reviewers R, products P, reviews E, and rate function, and then assigns anomalous scores and summaries for reviewers and products as results. **RIH** also receives three parameters: α, β, and γ.

RIH initializes the review summary of each product to the average of ratings posted to the product, and then starts the repeated improvement. In each iteration, **RIH** computes two CDFs first. It then updates the anomalous score of each reviewer by Definition 7, and subsequently updates the review summary of each product by Definition 2. The iteration is continued until the updates of anomalous scores and summaries become negligible. After that, **RIH** returns the final anomalous scores and summaries. It is noteworthy that, in this algorithm, the computational cost of one iteration is $O(|E|)$, which is the same as that Fraud Eagle.

5. Experiment

5.1 Dataset

For this experimental evaluation, we use a review dataset provided by Jindal et al. [4] that consists of reviews posted to Amazon.com before June, 2006. We extract reviews for products in the *book category*. We refer to book reviews extracted from Amazon.com as the *Amazon dataset*.

One objectives of this experiment is to evaluate whether these methods can estimate long-term summaries from early reviews. Therefore, we first divide these book

reviews into two sets: reviews posted before 2005 and reviews posted after 2005. Then, we apply each method to the former set (i.e. older reviews only) for computing anomalous scores and summaries, and compare the obtained summaries with summaries calculated as the simple average of all ratings for each book using both the sets (i.e. including newer reviews). If the difference between these two kinds of summaries is small, then the performance of estimation is regarded as good. The Amazon dataset includes 239,371 books, with 1,555,315 reviews posted before 2005, and 613,265 reviews posted after 2005; the number of reviewers posting reviews before 2005 is 730,667.

The other objective of this experiment is to examine whether each method can detect anomalous reviewers in early reviews. For this objective, we require a ground truth indicating which reviewers are anomalous and which are not. However, two important difficulties arise in preparing such a ground truth. First, the number of reviewers in the Amazon dataset is too huge to investigate whether every reviewer is anomalous manually. Second, most past fraudulent reviewers were deleted from Amazon because most E-commerce sites delete spam reviews for the customers' benefit. That means the Amazon dataset has less anomalous reviewers than real world reviews. Therefore, we decided to add review data which simulate a possible way for fraudulent reviewers to attack early reviews.

It is considered that there are the following two ways for colluding fraudulent reviewers to attack a product having only early reviews: (1) posting numerous negative (or positive) reviews to reduce the influence of positive (or negative, respectively) reviews of early adopters, or (2) posting a small but sufficient number of reviews for occupying the majority in early reviews. Because the reviewers who take the first way (1) repeatedly are quite easily identifiable using simple methods, we specifically examine the second way (2). Even if reviewers take the second way (2), they would be identified easily if their ratings were always extreme. Therefore, we construct a dataset of added reviewers to simulate fraudulent reviewers and anomalous reviewers who have various rating tendencies, as follows.

Here, we assume the five-star rating system in which each reviewer chooses a rating from $\{1, 2, 3, 4, 5\}$. For positive integer parameters S_a and S_n, the added reviewers consist of S_a groups of anomalous reviewers and S_n groups of normal reviewers. To simulate the situation (2), let the number of reviewers in each group be chosen randomly from six to nine; all reviewers in each group give ratings to t target products which has fewer reviews than the number of reviewers in the group, for a specified positive integer t. To simulate the various rating tendencies explained above, we divide each anomalous reviewer group into two subgroups 1 and 2, each of which consists of three to six reviewers (determined randomly); we assume that (a) the ratings of reviewers in each subgroup are the same value, which differs from the original average at least one, (b) the difference between ratings of the two subgroups in each group is at least two. Therefore, if the original average of a product is

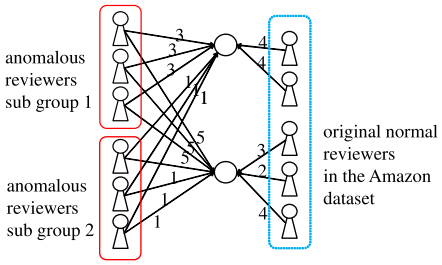

Fig. 3 Group of added anomalous reviewers.

greater than four, then the respective ratings of subgroup 1 and 2 are three and one. Otherwise, the respective ratings of subgroups 1 and 2 are five and one. Figure 3 presents an example of an added anomalous reviewers group and their ratings.

We are able to determine the rating values of each reviewer more randomly instead of the values described here, although it would affect the result only slightly because possible values are quite limited in the five-star rating system. The average of ratings of added normal reviewers is expected to be close to the original average of their target product. Because ratings are limited to integers but because averages can be real numbers in the five-star rating system, we determine the proper integers for their ratings using a simple calculation. The graph structures of anomalous and normal reviewers groups are fundamentally identical. Therefore, we can examine whether a method can distinguish them well.

5.2 Competitors

We explain other methodologies that can be compared to our method (**RIH**). **ONE** [5] employs an anomalous index. This method uses no repeated improvement. It computes anomalous scores just one time. **MRA** [14] uses repeated improvement, but does not address heterogeneity. The remaining two are Fraud Eagle [1], abbreviated to **FEA**, and FRAUDAR [2], abbreviated to **FRA**. Also, **MRA**, **FEA**, and **RIH** are iterative methods. The numbers of iterations for **MRA**, **FEA**, and **RIH**, are 10, 20, and 10. After these numbers of iterations, the results of these methods are almost unchanged. We use Google Cloud Platform (compute engine with 2 cores 7.5 GB memory) to conduct experiments.

5.3 Measurements

We use four measurements, "AUCa", "AUCe", "Diff1", and "Diff2", to evaluate the five methods introduced above.

AUCa and AUCe.
The purpose of the first two measurements is to examine whether each method could detect added anomalous reviewers. We employ the area under curve (AUC) of the ROC curve for detecting added anomalous reviewers among all reviewers. For a method that assigns each reviewer an anomalous score of a real number between 0 and 1, we assume that the method regards reviewer r as anomalous if

a(r) > θ where real number θ is a given threshold. By varying θ from 0 to 1, we can calculate the AUC.

In such a case, **RIH**, **ONE**, and **MRA** actually assign anomalous scores to all reviewers. Therefore, we can calculate the AUC as explained above. However, **FRA** and **FEA** do not assign anomalous scores. **FRA** is a binary classification of reviewers, and judges whether each reviewer is a fraud or not. Therefore, by regarding reviewer r "judged as fraud" as "a(r) = 1" and r "judged as not fraud" as "a(r) = 0", we can calculate the AUC according to the explanation presented above. **FEA** is a binary classification of reviewers, although it computes the probability that each reviewer is a fraud. We can calculate AUC by regarding this probability as an anomalous score of each reviewer.

Using the AUC, we prepare two measurements: AUC(all), abbreviated to AUCa, and AUC(early), abbreviated to AUCe. AUC(all) is the AUC calculated using all the reviewers including the original reviewers in the Amazon dataset and the added anomalous and normal reviewers. AUC(early) is the AUC calculated using the added anomalous and normal reviewers only, who post reviews to products having only early reviews.

Diff1 and Diff2.
The remaining two measurements are used for evaluating the results of estimation of long-term summaries. As we explained in Sect. 5.1, we apply each method to reviews posted before 2005, and regard the computed summaries as a result of estimation. Then, we calculate the difference between the computed summaries and "correct" long-term summaries, which are average ratings posted before and after 2005. Although some methods normalize ratings to $[0, 1]$, we rescale them to $[1, 5]$ so that we can easily compare results with those obtained using other methods. We use the mean of the absolute value of the difference for each product to prepare the following two measurements. A mean larger than 1.0 for a product corresponds to a difference that is greater than one star in a five-star rating system. It would be sufficient to affect customers' buying choices.

Diff1 is the mean calculated by targeting products to which the added anomalous reviewers post reviews. Similarly, the target products of Diff2 are those to which both the added anomalous and normal reviewers post reviews. The added reviewers post reviews to products having only early reviews. Therefore, these measurements evaluate the estimation results obtained from long-term summaries from early reviews. **RIH**, **ONE**, and **MRA** actually assign anomalous scores to all reviewers. Then we can calculate summaries according to Definition 2. The result of reviewer classification of **FRA** is converted into anomalous score 0 or 1 in the manner explained above. We use this score to calculate summaries according to Definition 2. Because **FEA** classifies each product as a good or bad one, we can convert good to five stars and bad to one star. Our preliminary experiments, however, revealed that this approach is much worse than the approach that incorporates a probability assigned to each reviewer as the anomalous score and which calculates summaries according to Definition 2. Therefore, we use the approach using the probability for **FEA**.

Table 2 Best parameters for respective methods and results.

(a) S_a = 10000, S_n = 10000, t = 2.

Method	Param	AUCa	AUCe	Diff1	Diff2
ONE	-	0.694	0.699	1.35	0.727
MRA	-	0.743	0.756	1.10	0.609
FEA	0.05	0.585	0.690	0.863	0.497
FRA	20	0.791	0.633	0.810	0.558
RIH	6, 3, 11	**0.869**	**0.891**	**0.621**	**0.428**

(b) S_a = 10000, S_n = 20000, t = 2.

Method	Param	AUCa	AUCe	Diff1	Diff2
ONE	-	0.708	0.741	1.35	0.520
MRA	-	0.758	0.793	1.10	0.443
FEA	0.05	0.600	0.691	0.864	0.375
FRA	20	0.836	0.662	0.647	0.450
RIH	6, 3, 11	**0.876**	**0.899**	**0.622**	**0.365**

(c) S_a = 10000, S_n = 30000, t = 2.

Method	Param	AUCa	AUCe	Diff1	Diff2
ONE	-	0.720	0.760	1.35	0.417
MRA	-	0.771	0.810	1.09	0.359
FEA	0.05	0.612	0.690	0.863	**0.314**
FRA	20	0.811	0.657	0.643	0.425
RIH	6, 3, 11	**0.882**	**0.903**	**0.621**	0.332

(d) S_a = 20000, S_n = 20000, t = 2.

Method	Param	AUCa	AUCe	Diff1	Diff2
ONE	-	0.689	0.698	1.35	0.720
MRA	-	0.739	0.754	1.11	0.609
FEA	0.05	0.600	0.678	0.859	0.493
FRA	20	0.839	0.690	0.645	0.497
RIH	6, 3, 11	**0.875**	**0.890**	**0.617**	**0.427**

(e) S_a = 30000, S_n = 30000, t = 2.

Method	Param	AUCa	AUCe	Diff1	Diff2
ONE	-	0.685	0.693	1.35	0.786
MRA	-	0.734	0.747	1.127	0.669
FEA	0.05	0.598	0.649	0.857	0.529
FRA	20	0.844	0.745	0.652	0.498
RIH	6, 3, 11	**0.875**	**0.888**	**0.626**	**0.453**

5.4 Evaluation

Tables 2 (a) to 2 (e) present the results. Tables 2 (a) to 2 (c) present the results for added datasets fixing S_a, the number of added anomalous reviewer group, to 10,000 and increasing S_n, the number of added normal reviewer group, from 10,000 to 30,000. Tables 2 (a), 2 (d) and 2 (e) show the results for added datasets fixing $S_a/S_n = 1$ and increasing S_a and S_n from 10,000 to 30,000. We fix $t = 2$ because larger t generates too many added reviews compared to original reviews. Furthermore, we ascertained that there are only slight differences among the results for $t \in \{1, 2, 3\}$ in our preliminary experiments. For each table, we generate 10 added datasets randomly with the same parameter values which are described in the captions of respective tables. The result value in each cell is the mean of the results for the 10 datasets.

Parameters of each method.
The bold number(s) in columns represent the best value for the column among the methods used. Larger values are better for AUCa and AUCe, although smaller values are better for Diff1 and Diff2. The column "Param" shows the used value(s) of parameter(s) for each method. Our method **RIH** takes two parameters β and γ. The values are written in this order. For example, numbers "6, 3, 11" in the "Param" column for **RIH** mean that $\alpha = 6, \beta = 3$, and $\gamma = 11$.

We explain how we determined the values of parameters used by **FEA**, **FRA** and **RIH**. We first discuss α used in **RIH**. Let a *limit product* denote a product with controversiality of exactly 0 or 1. To use the range of the controversiality [0, 1] without waste, it is not desired that there be too many limit products or that there be no limit product. Therefore, we can compute proper values of α by binary search from the review data. Results show that values around 6.0 are sufficient for the Amazon dataset. Therefore, we set $\alpha = 6.0$ through the experiments. For the other parameters for **FEA**, **FRA** and **RIH**, we conducted preliminary experiments using a grid search approach; we obtained candidate values of parameters for good performance. The candidates used for **FEA** were $\{0.01, 0.05, 0.07, 0.1\}$. Those for **FRA** were $\{10, 20, 30, 40, 50, 60, 70, 80, 90\}$. The candidates of β in **RIH** are $\{1, 3, 5, 10\}$, and those of γ are $\{9, 11, 13, 15\}$. We chose values that perform the best for each dataset and for each method. However, in our experiments, every method takes the same values of the parameters for all datasets.

It is noteworthy that the original paper [2] of **FRA** describes that the method has no parameter, although the authors' implementation of **FRA** can accommodate an optional parameter. Setting the parameter to one is identical to the method proposed in the paper. We tried such a setting in preliminary experiments, although its result was much worse than the parameters used above.

Performance of each method.
ONE is the simplest method, although its performance is far from the best. Especially, it performs the worst for Diff1 and Diff2 in most cases (except Diff2 in Table 2 (c)).

MRA performs better than **ONE** for all measurements and all datasets. Furthermore, the AUCe values of **MRA** are higher than those of the other methods except **RIH**. This result indicates the effectiveness of the concept of the repeated improvement, especially for detecting anomalous reviewers in early reviews. Unfortunately, **MRA** does not consider the heterogeneity that we claimed in this work. This is a primary reason why **MRA** performs worse than **RIH** for all measurements and all datasets.

The AUCa and AUCe of **FEA** are lower than those of the other methods, except for Table 2 (c). For every table, Diff1 of **FEA** is better than **ONE** and **MRA**, although it is worse than **FRA** and **RIH**. The values for Diff2 of **FEA** are close to the best ones. Especially, in Table 2 (c), Diff2 of **FEA** is the best. Therefore, **FEA** performs better for Diff2 than for Diff1.

We discuss the reason for this phenomenon. We denote an error by which "a reviewer who is presumed to have a very low anomalous score is assigned a very high anomalous score" as a type I error, and denote an opposite error as a type II error. Type II errors can be extremely important for Diff1 and Diff2 because the ratings of misjudged anomalous reviewers obtain heavy weights to summaries. Type I errors cause ignorance misjudged normal reviewers to calculate summaries, although this does not affect summaries much if there are other reviewers who are assigned low anomalous scores and who have similar ratings to the ignored ones. The target of Diff2 includes books to which the added normal reviewers review, in addition to the target of Diff1. Even if a method misjudges many added normal reviewers (type I errors) and both AUCa and AUCe are consequently bad, Diff2 of the method becomes good if it can reduce type II errors. It is considered that **FEA** exactly matches this situation.

FEA is fundamentally a binary classification method. However, degree of anomaly of reviewers is diverse in actual review sites including the added reviewers in this work. If a binary classification method forces to classify reviewers with a low degree of anomaly as anomalous ones, the situation explained above would occur. We found that Diff2 of **FEA** becomes better as S_n increases (from Tables 2 (a) to 2 (c)). This tendency of **FEA** supports the explanation presented above.

FRA performs well for AUCa but not well for AUCe. The value of AUCe becomes better as S_a increases, although it is the lowest in each of Tables 2 (a), 2 (b), and 2 (c), which set S_a to 10,000. This result implies that **FRA** could not distinguish added reviewers well, although AUCa is not affected strongly because the number of added reviewers is much smaller than that of all reviewers. Therefore, **FRA** is considered to be weak against anomalous reviewers in early reviews. Because **FRA** only uses the graph structure of reviews, the method could not distinguish two reviewers reviewing the same product but the tendencies of rating differ such as the added reviewers in our experiment. Diff1 of **FRA** is very close to the best, that of **RIH**, although Diff2 of **FRA** is inferior to that of **RIH**. Diff2 is related to both the added anomalous and normal reviewers in early reviews. Therefore, this fact also supports the explanation above that **FRA** seems accurate overall but it is weak against early reviews.

RIH performs the best or very close to the best for all measurements and datasets. Actually, it achieves the best for 19 items among 20 (4 measurements $\times$ 5 datasets). Calculation of the t-test confirmed that the difference between the best value and the second one is significant at the $p = 0.05$ level in every column in every table. Especially, **RIH** outperforms the other methods with respect to AUCe. From these results, we infer that our two ideas for dealing with the heterogeneity demonstrate the expected effectiveness for early reviews.

6. Conclusion

We have proposed a methodology for detecting anomalous reviewers and for estimating long-term review summaries from early reviews by considering heterogeneity on actual review sites. We have clarified heterogeneity of two kinds in actual review sites such as Amazon.com. To address heterogeneity, we have introduced two novel ideas for our method: *deviation rarity* and *controversiality*. We have ascertained the effectiveness of our method through the experiments using a real review dataset from Amazon.com and compare them with other methods including state-of-the-art methods: Fraud Eagle and FRAUDAR.

The authors who proposed FRAUDAR claimed that their method can deal with *camouflage* merely by using the graph structure of reviews, although our experimentally obtained results reveal an important difficulty for FRAUDAR to distinguish reviewers in early reviews. However, no method, including our method RIH, employs both the graph structure and ratings to address the difficulties posed by camouflage and early reviews well. That remains as an open problem related to construction of such methods.

References

[1] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion fraud detection in online reviews by network effects," Proc. 7th International Conference on Weblogs and Social Media, pp.2–11, 2013.
[2] B. Hooi, H.A. Song, A. Beutel, N. Shah, K. Shin, and C. Faloutsos, "FRAUDAR: Bounding graph fraud in the face of camouflage," Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.895–904, 2016.
[3] J. Liu, Y. Cao, C. Lin, Y. Huang, and M. Zhou, "Low-Quality Product Review Detection in Opinion Summarization," Proc. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.334–342, 2007.
[4] N. Jindal and B. Liu, "Opinion spam and analysis," Proc. 2008 International Conference on Web Search and Data Mining, pp.219–230, Feb. 2008.
[5] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H.W. Lauw, "Detecting Product Review Spammers using Rating Behaviors," Proc. 19th ACM International Conference on Information and Knowledge Management, pp.939–948, ACM Press, Oct. 2010.
[6] M. Ott, Y. Choi, C. Cardie, and J.T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," Proc. 49th Annual Meeting of the Association for Computational Linguistics, pp.309–319, 2011.
[7] A. Mukherjee, B. Liu, J. Wang, N.S. Glance, and N. Jindal, "Detecting group review spam," Poster Proc. 20th International Conference on World Wide Web, pp.93–94, 2011.
[8] A. Mukherjee, B. Liu, and N. Glance, "Spotting Fake Reviewer Groups in Consumer Reviews," Proc. 21st International Conference on World Wide Web, pp.191–200, 2012.
[9] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting Burstiness in Reviews for Review Spammer Detection," Proc. 7th International Conference on Weblogs and Social Media, pp.175–184, 2013.
[10] S. Xie, G. Wang, S. Lin, and P.S. Yu, "Review spam detection via temporal pattern discovery," Proc. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.823–831, 2012.
[11] S. Feng, L. Xing, A. Gogar, and Y. Choi, "Distributional Footprints of Deceptive Product Reviews," Proc. Sixth International Conference on Weblogs and Social Media, pp.98–105, 2012.
[12] D. Easley and J. Kleinberg, Networks, Crowds, and Markets: Reasoning About a Highly Connected World, Cambridge University Press, 2010.
[13] H.W. Lauw, E. Lim, and K. Wang, "Summarizing review scores of "unequal" reviewers," Proc. Seventh SIAM Conference on Data Mining, pp.539–544, 2007.
[14] K. Tawaramoto, J. Kawamoto, Y. Asano, and M. Yoshikawa, "A Bipartite Graph Model and Mutually Reinforcing Analysis for Review Sites," Proc. 22nd International Conference on Database and Expert Systems Applications, vol.6860, pp.341–348, 2011.
[15] G. Wang, S. Xie, B. Liu, and P.S. Yu, "Review Graph Based Online Store Review Spammer Detection," Proc. 11th IEEE International Conference on Data Mining, pp.1242–1247, Dec. 2011.
[16] D. Ikeda, H. Takamura, L.A. Ratinov, and M. Okumura, "Learning to shift the polarity of words for sentiment classification," Proc. Third International Joint Conference on Natural Language Processing, pp.296–303, 2008.
[17] N. Jindal and B. Liu, "Analyzing and Detecting Review Spam," Proc. Seventh IEEE International Conference on Data Mining, pp.547–552, Oct. 2007.
[18] A. Neviarouskaya, H. Prendinger, and M. Ishizuka, "Affect analysis model: Novel rule-based approach to affect sensing from text," Nat. Lang. Eng., vol.17, no.1, pp.95–135, 2011.
[19] M.S. Pera, R. Qumsiyeh, and Y.-K. Ng, "An Unsupervised Sentiment Classifier on Summarized or Full Reviews," Proc. 11th International Conference on Web Information Systems Engineering, pp.142–156, Dec. 2010.
[20] J.M. Kleinberg, "Authoritative sources in a hyperlinked environment," J. ACM, vol.46, no.5, pp.604–632, Sept. 1999.

Yasuhito Asano received the BS, MS, and DS degrees in information science from the University of Tokyo in 1998, 2000, and 2003 respectively. In 2003–2005, he was a research associate in the Graduate School of Information Sciences, Tohoku University. In 2006–2007, he was an assistant professor in the Department of Information Sciences, Tokyo Denki University. He joined Kyoto University in 2008. He currently serves as an associate professor in the Graduate School of Informatics. His research interests include web mining and network algorithms. He is a member of the IEICE, IPSJ, DBSJ and OR Soc. Japan.

Junpei Kawamoto received his B.Eng. in 2007, Master of Informatics in 2008, and Ph.D. in Informatics in 2012 all from Kyoto University, Japan. He was also a research fellow of the Japan Society for the Promotion of Science between 2009–2011, an expert researcher of Japan's National Institute of Information and Communications Technology in 2012, then a Postdoctoral Fellow at the University of Tsukuba between 2012–2013. He is currently an Assistant Professor at Kyushu University in Fukuoka, Japan and concurrently working with the Institute of Systems and Information Technologies and Nanotechnologies, Japan. He is interested in database security and privacy preserving data mining.