Abstract

Many web services like eBay, TripAdvisor, Epinions, etc, provide historical product ratings so that users can evaluate the quality of products. Product ratings are important since they affect how well a product will be adopted by the market. The challenge is that we only have "partial information" on these ratings: Each user provides ratings to only a "small subset of products". Under this partial information setting, we explore a number of fundamental questions: What is the "minimum number of ratings" a product needs so one can make a reliable evaluation of its quality? How users' misbehavior (such as cheating) in product rating may affect the evaluation result? To answer these questions, we present a formal mathematical model of product evaluation based on partial information. We derive theoretical bounds on the minimum number of ratings needed to produce a reliable indicator of a product’s quality. We also extend our model to accommodate users’ misbehavior in product rating. We carry out experiments using both synthetic and real-world data (from TripAdvisor, Amazon and eBay) to validate our model, and also show that using the "majority rating rule" to aggregate product ratings, it produces more reliable and robust product evaluation results than the "average rating rule".

1 Introduction

Nowadays, it is common in many web services that users can contribute their opinions in the form of ratings or reviews. For example, we see ratings or reviews in content sharing sites (i.e., Flickr, YouTube, etc), online recommendation systems (i.e., Amazon, MovieLens, etc), product review sites (i.e., Epinions, TripAdvisor, etc), and online e-commerce systems (i.e., eBay, etc), etc. With these ratings and product reviews, one can perform information search or make purchase decisions by taking advantage of the opinions of other users (aka "wisdom of the crowd"). Web-based online rating systems can be briefly described
as follows: There are a number of items (either products or services), each user provides ratings or reviews to a subset of items, and ratings of an item will be available to the public.

Based on how these ratings are being used, online rating systems can be classified into two categories. The first category interprets ratings as public assessment on products, where the main interest is in assisting users to evaluate products’ quality. For example, in eBay, ratings are used to reflect the quality a product or the reputation of a seller. Other examples of this type of web services include Tripadvisor, Epinions, Wikipedia, etc. For these web services, users want to know whether the quality of a product is as good as its declared historical ratings, or whether the reputation of a seller is fully reflected by its historical ratings. The second category interprets ratings as users’ preference information and they are used in users’ preference modeling. This is the view adopted by researchers in online recommendation systems, i.e., in Amazon and MovieLens, users’ ratings are used to infer users’ preferences as to make personalized recommendations.

In this paper, we focus on the first category of online rating systems, in which ratings are interpreted as product quality assessment. Although there are wide deployment of product ratings in web services, people only have partial information on these ratings: each user only expresses ratings to a small subset of products. Hence, it is important to understand the “accuracy” and “effectiveness” of online rating systems. Little attention has been paid to this fundamental question. The goal of this paper is to fill this void. We first provide a mathematical model for a general online rating system, then we explore a number of fundamental questions, i.e., (a) What is the minimum number of ratings a product needs so as to have a reliable reflection on its quality? (b) How users’ misbehavior may affect the the accuracy? To the best of our knowledge, this is the first paper that provides a formal mathematical model and analysis on such problem. Our contributions are:

- We propose a mathematical model to specify the rating behavior of users to a set of products. Our model is general enough to represent both honest and misbehaving users in product rating. We analyze the model and derive the minimum number of ratings we need to produce a reliable reflection on a product’s quality. We also derive the minimum fraction of misbehaving users such that they can manipulate an online rating system (please refer to Section 2 and 3).

- We propose an inference algorithm to infer parameters of our model from partial information, say available products’ ratings (please refer to Section 4).

- We show that the majority rule is more robust and insensitive to users’ misbehavior as compared with the average scoring rule (please refer to Section 4 and Section 5).

- We perform experiments using both synthetic data and real-world data (data set from TripAdvisor, Amazon and eBay) to validate our framework.
and to examine various factors that may affect product quality evaluation. We find a number of interesting observations, i.e., the system can resist a small fraction of misbehaving users. For TripAdvisor, Amazon, and Ebay, we also show why the majority rule is more robust and reliable than the average scoring rule in evaluating products’ quality. (please refer to Section 4 and Section 6).

This is the outline of the paper. In § 2, we present the mathematical model of product rating systems. In § 3, we show the analysis and derive theoretical results for two rating aggregation rules under normal or misbehaving setting. In § 4, we present an inference algorithm to infer the parameters of our model. In § 5, we present our experimental results using synthetic data. In § 6, we present our experimental results using real-world data. Related work is given in § 7 and § 8 concludes.

2 Mathematical Model

Let there be a finite set of $N$ products denoted by $P_1, \ldots, P_N$. We have $M$ users denote by $U_1, \ldots, U_M$. Each user only expresses ratings to a subset of products on an $m$-level cardinal metric where ratings are drawn from $\{1, \ldots, m\}$, i.e., a 3-level cardinal metric could be: $\{1 =$ "poor", 2 = "good", 3 = "excellent"\}. Ratings from users are independent. Product $P_i$ is rated by $n_i \leq M$ users. Let $r_i = \{r_{i,1}, \ldots, r_{i,M}\}$ denote a set of $M$ ratings of product $P_i$, where $r_{i,j} \in \{1, \ldots, m\}$ if user $U_j$ rates product $P_i$, otherwise $r_{i,j} = 0$. Higher rating implies higher quality. We treat the available ratings as partial information because some $r_{i,k} = 0$. To make a purchase decision, a user evaluates the quality of products relying on historical product ratings. The objective of this work is to answer how various factors can influence the evaluation accuracy.

Let $A$ denote a rating aggregation rule which is used to summarize all ratings of a product. Two commonly used rating aggregating rules are the majority rule and the average scoring rule.

• Majority rule. It evaluates the quality of a product via the majority ratings. Let $\ell_i \in \{1, \ldots, m\}$ represents the label which reflects the true quality of product $P_i$ under this policy. Let $\hat{\ell}_i$ denote the evaluated label of product $P_i$ produced by applying the majority rule. The label $\hat{\ell}_i$ can be computed as follows. Let $n_{i,k}$ be the number of ratings of $P_i$ that is of level $k$, where $k \in \{1, \ldots, m\}$. Then $\hat{\ell}_i = \arg \max_k \{n_{i,k}\}$. There are a number of interesting questions to explore, i.e., How many ratings, or $n_i$, do we need so as to have a strong guarantee in revealing the true label, or $\Pr[\hat{\ell}_i = \ell_i]$ with high probability?

• Average scoring rule. It evaluates the quality of a product via averaging all ratings of that product. Let $\gamma_i \in [1, m]$ denote the true quality of $P_i$ under this policy. Let $\hat{\gamma}_i = \sum_j r_{i,j} / n_i$ denote the evaluated label of $P_i$ using the average scoring rule. Again, there are number of interesting questions to explore, i.e., how many ratings, or $n_i$ do we need, such that $\hat{\gamma}_i$ accurately reflects $\gamma_i$?
Remark: There should be no confusion between $\ell_i$ and $\gamma_i$. They show two different ways in evaluating the quality of products either by the ratings of the majority population, or by averaging the ratings on the whole population respectively.

Because each user may rate a product differently. To describe a user’s rating behavior, let us first present the probabilistic model in product rating.

### 2.1 Model for rating behavior

When rating a product, a user needs to evaluate its quality. We consider two most important factors that may affect the evaluation: (a) intrinsic quality of products, and (b) preference of users on a product. The rating behavior of different users can be modeled by a random variable in which one can vary its mean or variance to reflect the above factors. Specifically, a higher mean implies that the product has a higher intrinsic quality, while a smaller variance implies that a user prefers a particular branding.

To illustrate, consider the user $U_j$ who rates product $P_i$, and the rating is $r_{i,j}$. The rating $r_{i,j}$ is a random variable, and its probability mass function (pmf) is:

$$\Pr[r_{i,j} = k] = \rho_k, \quad k = 1, \ldots, m,$$

where $\rho_k \geq 0$ and $\sum_{k=1}^{m} \rho_k = 1$. One can set the mean or variance by varying the probability mass distribution $\{\rho_1, \ldots, \rho_m\}$ to reflect different intrinsic quality or users preference on product $P_i$. To model the collective rating behavior of the whole user population over product $P_i$, let

$$S = \left\{ (\theta_1, \ldots, \theta_m) \bigg| \sum_{i=1}^{m} \theta_i = 1, \theta_i \geq 0, \forall i \right\}$$

denote the space over all the possible probability distributions of a rating. We assume that there is an underlying distribution, say $D(P_i)$, over the space $S$, that defines the collective rating behavior of the whole user population over product $P_i$. In this study, $D(P_i)$ is a Dirichlet distribution $Dirichlet(\alpha_i)$ with density function:

$$p(\theta_1, \ldots, \theta_m) = \frac{\Gamma(\sum_{k=1}^{m} \alpha_{i,k})}{\prod_{k=1}^{m} \Gamma(\alpha_{i,k})} \prod_{k=1}^{m} \theta_{i,k}^{\alpha_{i,k}-1}.$$

where $\alpha_i = (\alpha_{i,1}, \ldots, \alpha_{i,m})$, $\sum_{k=1}^{m} \alpha_{i,k} = 1$ and $\alpha_{i,k} > 0, \forall k$.

Remark: The choice of Dirichlet distribution is validated in [3].

### 3 Theoretical Analysis

In this section, we analyze two most representative rating aggregation rules: majority rule and average scoring rule, that are widely used in online rating systems. The central question we seek to answer is how many ratings of a given product we need so that the aggregate rating is statistically accurate to reflect
the ground truth of that product’s quality? Furthermore, we also analyze the impact of misbehavior in product rating, e.g., when users intentionally rate a product beyond its ground truth level. In this section, we assume that the model parameters $\alpha_i, \forall i$, are given. In Section 4, we will show how to infer these parameters from historical ratings.

3.1 Majority rule

Majority rule is suggested as a good rule to aggregate ratings[1]. This rule helps a user to understand how receptive the majority of the masses are toward a given product. Note that we only have partial information, and this makes it challenging to extract the true label of products. There are a number of interesting questions to explore, i.e., how many ratings do we need so that we can extract the true label with high probability? What is the effect of misbehavior in ratings? Let us begin our exploration by a simple case, say all users rate products honestly.

Analysis for Honest Rating: In this case, we assume that with a large enough number of ratings, one can successfully extract the true label of products. Recall that $\hat{\ell}_i$ is the label of product $P_i$ produced by using the majority rule on its ratings, and $\ell_i$ is the true label of product $P_i$. Intuitively, the true label of product $P_i$ occurs when $\ell_i = \lim_{n_i \to \infty} \hat{\ell}_i$. Recall that $P_i$ has $n_i$ ratings and its rating set is $r_i$, where missing ratings are denoted by zeros. For convenience, here we use $r^+_i = \{r^+_{i,1}, \ldots, r^+_{i,n_i}\}$ to denote a set of all positive (or observed) ratings of $r_i$. Let us state the probability mass function (pmf) of $r^+_{i,j}$, $j = 1, \ldots, n_i$ in the following lemma.

Lemma 1 The pmf of the rating $r^+_{i,j}$ is:

$$\Pr[r^+_{i,j} = k] = \alpha_i, k, \quad \text{for } k = 1, \ldots, m,$$

where $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$.

Proof: Please refer to the appendix for derivation.

Lemma 1 states that the probability of product $P_i$ receiving a rating $k$ is $\alpha_{i,k}$. It follows that $\ell_i = \lim_{n_i \to \infty} \hat{\ell}_i = \arg \max_k \{\alpha_{i,k}\}$. Let $\bar{\alpha}_i$ denote the second largest value among $\alpha_{i,1}, \ldots, \alpha_{i,m}$. Now we state the main theorem for the honest rating case.

Theorem 1 (Honest Rating Case) Suppose all users rate honestly, and product $P_i$ has at least

$$n_i \geq n'_i = \frac{12\alpha_i, \ell_i}{(\alpha_i, \ell_i - \bar{\alpha}_i)^2} \ln \frac{m}{\delta}$$

ratings, then one can claim with high probability that the true label of product $P_i$ is $\hat{\ell}_i$. Mathematically, we have: $\Pr[\hat{\ell}_i = \ell_i] \geq 1 - \delta$. 

5
Proof: Please refer to the appendix for derivation.

Table 1 shows some numerical results on the minimum number of ratings we need when all users rate honestly. It depicts the level of rating metric $m$, the success probability $1 - \delta$ and the lower bound of the minimum number of ratings denoted by $n'_i$ respectively. From the table, we see that if we increase the success probability $(1 - \delta)$, we also need to increase $n'_i$.

$$m \alpha_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$$

| $m$ | $\alpha_i$ | $1 - \delta$ | $n'_i$ |
|-----|-------------|---------------|--------|
| 5   | $\left( \frac{4}{35}, \frac{2}{35}, \frac{3}{35}, \frac{2}{35}, \frac{1}{35} \right)$ | 0.7 | 67   |
| 5   | $\left( \frac{4}{35}, \frac{2}{35}, \frac{3}{35}, \frac{2}{35}, \frac{1}{35} \right)$ | 0.8 | 77   |
| 5   | $\left( \frac{4}{35}, \frac{2}{35}, \frac{3}{35}, \frac{2}{35}, \frac{1}{35} \right)$ | 0.9 | 93   |

Table 1: Honest rating: minimum number of ratings.

Remark: Theorem 1 gives an lower bound on the minimum number of ratings $n_i$. One may ask whether this bound is sufficiently tight. The following theorem answers this question.

Theorem 2 (Tightness) Suppose all users rate honestly and $\tilde{\alpha}_i \geq \frac{100}{101} \alpha_{i,\ell_i}$. Then there exist a positive constant $\eta$, such that for any $\delta \leq \eta$, if product $P_i$ has at most

$$n_i = O \left( \frac{\alpha_{i,\ell_i}}{(\alpha_{i,\ell_i} - \tilde{\alpha}_i)^2 \ln \frac{1}{\delta}} \right)$$

ratings, then we will fail to extract the true label with probability at least $\Omega(\delta)$.

Proof: Please refer to the appendix for derivation.

Remark: Theorem 2 states the lower bound derived in Theorem 1 is asymptotically tight in general.

Analysis of Rating under Misbehavior: Let us now explore the effect of misbehavior in ratings. We consider the following typical cases of misbehavior:

- **Random misbehavior.** A random misbehavior implies that a user assigns a random rating to a product. This is one typical misbehavior, because sometimes a user may not want to spend time to evaluate the quality of a product so that user just generates random ratings.

- **Biased misbehavior:** A biased misbehavior implies that a user is biased toward one particular rating, i.e., a user may be hired by a company to assign the lowest rating to a competitor’s product, or assign the highest rating to his employer’s product. This type of misbehavior has been observed in many web services, i.e., reported in TripAdvisor [1].
In the remaining of this sub-section, we *model* the ratings of misbehaving users as noise. We specify the collective rating behavior of honest users to product $P_i$ using our model with parameter $\alpha_i$. Again, $\ell_i = \arg \max \{ \alpha_{i,k} \}$ is the true label of product $P_i$ and $\tilde{\alpha}_i$ is the second largest value among $\alpha_{i,1}, \ldots, \alpha_{i,m}$.

The following theorem states the effect of random misbehavior.

**Theorem 3 (Resist Random Misbehavior)** Let $f_i < 1$ denote the fraction of random misbehaving users who rate product $P_i$. If product $P_i$ has $n_i \geq n'_i$ ratings, then one can claim with high probability that the true label of product $P_i$ is $\hat{\ell}_i$. Mathematically, we have: $\Pr[\hat{\ell}_i = \ell_i] \geq 1 - \delta$.

**Proof:** Please refer to the appendix for derivation.  

**Remark:** Theorem 3 states that if the fraction of random misbehavior is less than 1, then one can always extract the true label of a product with a large enough number of ratings.

Table 2 shows some numerical results of the minimum number of ratings when some ratings are assigned by random misbehaving users. Table 2 depicts $m$, $\alpha_i$, $1 - \delta$, $n'_i$, and the fraction of random misbehaving users $f_i$ respectively.

We can see that if we increase $f_i$, we also need to increase the minimum number of ratings.

| $m$ | $\alpha_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$ | $1 - \delta$ | $f_i$ | $n'_i$ |
|-----|---------------------------------|--------------|------|-------|
| 5   | $(\frac{4}{35}, \frac{22}{35}, \frac{1}{35}, \frac{1}{35}, \frac{1}{35})$ | 0.8          | 0.1  | 88    |
| 5   | $(\frac{4}{35}, \frac{22}{35}, \frac{1}{35}, \frac{1}{35}, \frac{1}{35})$ | 0.8          | 0.2  | 102   |

Table 2: Random misbehavior: minimum number of ratings.

We now consider the effect of biased misbehavior. Let $\ell'_i \in \{1, \ldots, m\}$ denote the rating that biased misbehaving users bias toward. The goal of the biased users is to make the extracted label of a product to be the label that they bias toward. Then, how many biased users is needed such that they will achieve their goal with high probability? The following theorem answers this question.

**Theorem 4 (Biased Misbehaving Users Win)** Let $f'_i$ denote the fraction of biased misbehaving users who rate product $P_i$. If the fraction of biased misbehaving users is no less than $f'_i > (\alpha_{i,\ell_i} - \alpha_{i,\ell'_i})/(1 + \alpha_{i,\ell_i} - \alpha_{i,\ell'_i})$, then with at least

$$n_i \geq \begin{cases} 
12 \left( \frac{f'_i + (1 - f'_i)(\alpha_{i,\ell'_i})}{(f'_i + (1 - f'_i)(\alpha_{i,\ell'_i} - \alpha_{i,\ell_i})^{2}} \ln \frac{m}{\delta}, & \ell'_i \neq \ell_i \\
12 \left( \frac{f'_i + (1 - f'_i)(\alpha_{i,\ell_i})}{(f'_i + (1 - f'_i)(\alpha_{i,\ell_i} - \tilde{\alpha}_i))^{2}} \ln \frac{m}{\delta}, & \ell'_i = \ell_i 
\end{cases}$$
ratings, one can claim with high probability that the biased misbehaving users win. Mathematically, we have: \( \Pr [\hat{\ell}_i = \ell_i'] \geq 1 - \delta. \)

**Proof:** Please refer to the appendix for derivation.

**Remark:** Theorem 4 states two conditions for biased misbehaving users to win: a large enough fraction of biased misbehaving users and a large enough number of such ratings.

Table 3 shows some numerical results of the minimum fraction of biased misbehaving users so that they can distort the evaluation process. Table 3 depicts \( m, \alpha_i, n_i' \), and the label that the biased misbehaving users bias toward \( \ell_i' \) respectively. From Table 3 we could see that if biased misbehaving users biased toward 5 (or 1), and to ensure they can distort, the minimum fraction of biased misbehaving users we need is 0.407 (or 0.375).

| \( m \) | \( \alpha_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5}) \) | \( \ell_i' \) | \( f_i' \) |
|---|---|---|---|
| 5 | \( (\frac{4}{35}, \frac{25}{35}, \frac{1}{35}, \frac{2}{35}, \frac{1}{35}) \) | 5 | 0.407 |
| 5 | \( (\frac{4}{35}, \frac{25}{35}, \frac{1}{35}, \frac{2}{35}, \frac{1}{35}) \) | 1 | 0.375 |

Table 3: Biased misbehavior: minimum fraction of misbehaving users to distort the evaluation.

**Proposition 1** Suppose that the misbehaving users bias against the ground truth label, or \( \ell_i' \neq \ell_i \). If the fraction of biased misbehaving users who rate product \( P_i \) is at least \( f_i' > (\alpha_{i,\ell_i} - \alpha_{i,\ell_i'})/(1 + \alpha_{i,\ell_i} - \alpha_{i,\ell_i'}) \), then it is impossible to extract the true label with high probability no matter how many ratings we have.

The following theorem states that when the fraction of biased misbehaving users is small, we can resist this type of anomaly.

**Theorem 5 (Resist Biased Misbehavior)** Suppose that the biased misbehaving users bias against the ground truth label, say \( \ell_i' \neq \ell_i \). If the fraction of biased misbehaving users who rate product \( P_i \) is at most \( f_i' < (\alpha_{i,\ell_i} - \alpha_{i,\ell_i'})/(1 + \alpha_{i,\ell_i} - \alpha_{i,\ell_i'}) \), then with at least

\[
 n_i \geq n_i' = \frac{12(1 - f_i')\alpha_{i,\ell_i} \ln m}{((1 - f_i')\alpha_{i,\ell_i} - \max \left \{ f_i' + (1 - f_i')\alpha_{i,\ell_i'}, (1 - f_i')\alpha_{i,\ell_i} \right \})^2}
\]

ratings, then one can claim with high probability that the true label of product \( P_i \) is \( \hat{\ell}_i \). Mathematically, we have: \( \Pr [\hat{\ell}_i \neq \ell_i] \geq 1 - \delta. \)

**Proof:** The derivation is similar to Theorem 4.
Table 4 shows some numerical examples of the minimum number of ratings we need to resist biased misbehavior, when the fraction of biased misbehaving users is small. From Table 4 we could see that when we increase the fraction of biased misbehaving users, we also need to increase the minimum number of ratings.

| $m$ | $\alpha_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$ | $1 - \delta$ | $\epsilon'_i$ | $f'_i$ | $n'_i$ |
|-----|---------------------------------|-------------|-------------|--------|--------|
| 5   | $\left(\frac{4}{35}, \frac{25}{35}, \frac{3}{35}, \frac{7}{35}, \frac{1}{35}\right)$ | 0.8         | 5           | 0      | 77     |
| 5   | $\left(\frac{4}{35}, \frac{35}{35}, \frac{3}{35}, \frac{3}{35}, \frac{1}{35}\right)$ | 0.8         | 5           | 0.1    | 93     |
| 5   | $\left(\frac{4}{35}, \frac{35}{35}, \frac{35}{35}, \frac{1}{35}, \frac{1}{35}\right)$ | 0.8         | 5           | 0.2    | 182    |

Table 4: Biased misbehavior: minimum number of ratings to resist biased misbehavior.

### 3.2 Average scoring rule

Average scoring rule evaluates the quality of a product by taking average on all ratings of that product. There are number of interesting question to explore, i.e., how many ratings do we need so that we can produce an average rating that reflects the quality of a product accurately? What is the effect of misbehavior in ratings? Let us explore these questions here.

**Analysis for Honest Rating:** We assume that for the honest rating case, with a large enough number of ratings, we can produce an average rating that reflects the quality of a product accurately. Recall that the average rating of product $P_i$ is $\hat{r}_i = \frac{1}{n_i} \sum_{j} r_{i,j}$, and $\gamma_i$ is the ground truth value that reflects the true quality of product $P_i$. Intuitively, $\gamma_i$ can be computed by $\gamma_i = \lim_{n_i \to \infty} \hat{r}_i$. From Lemma $[1]$ we have that $\Pr[r_{i,j} = k] = \alpha_{i,k}$, where $k = 1, \ldots, m$ and $j = 1, \ldots, n_i$. Thus, we have $\gamma_i = \sum_k k \alpha_{i,k}$. Let us first explore the case that all users rate honestly. We present the main theorem for the honest rating case as follows.

**Theorem 6 (Honest Rating Case)** Suppose all users rate honestly. If product $P_i$ has at least

$$n_i \geq n'_i = \frac{3}{\epsilon^2} \ln \frac{2m}{\delta}$$

ratings, then

$$|\hat{r}_i - \gamma_i| \leq \epsilon \sqrt{m \gamma_i} + m \epsilon^2$$

holds with probability at least $1 - \delta$.

**Proof:** Please refer to the appendix for derivation. \(\blacksquare\)

**Remark:** The physical meaning of Theorem is that with a large enough number of ratings, we can produce an accurate estimation of $\gamma_i$ with an arbitrarily small error.
Table 5 shows some numerical results on the minimum number of ratings we need to produce an accurate estimation of $\gamma_i$ when all users rate honestly. Table 5 depicts $m$, $\alpha_i$, $1 - \delta$, $n'_i$ and the absolute error bound between average rating and its ground truth value, or $E_r = \epsilon \sqrt{m\gamma_i} + me^2$ respectively. We could see that increasing the minimum number of ratings will decrease the error. It is also interesting to observe that using the average score rule, we need at least 221 ratings, while one only needs 77 ratings using the majority rule (please refer to Table 1).

### Table 5: Honest rating: minimum number of ratings (average scoring rule).

| $m$ | $\alpha_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$ | $1 - \delta$ | $E_r$ | $n'_i$ |
|-----|-------------------------------------------------|------------|-------|-------|
| 5   | $\left( \frac{4}{35}, \frac{2}{35}, \frac{3}{35}, \frac{2}{35}, \frac{1}{35} \right)$ | 0.8        |       | 221   |
| 5   | $\left( \frac{4}{35}, \frac{2}{35}, \frac{3}{35}, \frac{2}{35}, \frac{1}{35} \right)$ | 0.8        | 0.75  | 366   |
| 5   | $\left( \frac{4}{35}, \frac{2}{35}, \frac{3}{35}, \frac{2}{35}, \frac{1}{35} \right)$ | 0.8        | 0.5   | 716   |

### Analysis of Rating under Misbehavior:

We consider random misbehavior and biased misbehavior specified in last sub-section. We treat the ratings of misbehaving users as noise. The collective rating behavior of the honest users to product $P_i$ is specified by our model with parameter $\alpha_i$ and $\gamma_i = \sum_k k\alpha_{i,k}$ is the ground truth value of average rating of product $P_i$. The following theorems state the effect of random misbehavior and biased misbehavior respectively.

**Theorem 7 (Random Misbehavior)** Let $f_i$ denote the fraction of random misbehaving users who rate product $P_i$. If product $P_i$ has $n_i \geq 3 \ln(2m/\delta)/\epsilon^2$ ratings, then

$$
\left| \frac{m}{2} - \gamma_i f_i - \epsilon \sqrt{m(\gamma_i + mf_i/2 - \gamma_i f_i)} - me^2 \right| \leq |\bar{r}_i - \gamma_i|
$$

holds with probability at least $1 - \delta$.

**Proof:** Please refer to the appendix for derivation.

**Theorem 8 (Biased Misbehavior)** Let $\ell'_i$ denote the rating that the biased misbehaving users bias toward. Let $f'_i$ denote the fraction of biased misbehaving users who rate product $P_i$. If product $P_i$ has $n_i \geq 3 \ln(2m/\delta)/\epsilon^2$ ratings, then

$$
\left| \ell'_i - \gamma_i f'_i - \epsilon \sqrt{m(\gamma_i + f'_i \ell'_i - \gamma_i f'_i)} - me^2 \right| \leq |\bar{r}_i - \gamma_i|
$$

holds with probability at least $1 - \delta$.  

10
Proof: Please refer to the appendix for derivation.

Remark: The physical meaning of Theorem 7 and 8 is that average scoring rule is sensitive to random and biased misbehavior. If the fraction of misbehaving users is small, then we can still reflect the true quality of products with a high probability, otherwise it is impossible to accurately reveal the ground truth.

4 Inferring Model Parameters

In last section, we presented the formal analysis of our model with the assumption that the model’s parameters $\alpha_i, \forall i$ are known. In this section, we show how to obtain these parameters’ values by presenting a maximum likelihood inference algorithm to infer these parameters $\alpha_i, \forall i$ from partial information, or historical ratings. With this inference algorithm, we can apply our framework to analyze and improve the applications of online rating systems in web services, i.e., eBay, TripAdvisor, etc.

Recall that $r_i^+ = \{r_{i,1}^+, \ldots, r_{i,n_i}^+\}$ represents a set of all observed ratings of product $P_i$. We seek to infer $\alpha_i$ from $r_i^+$. From Lemma 1 we could write down the probability mass function (pmf) of $r_i^+$ as follows: $Pr[r_{i,j}^+ = k] = \alpha_{i,k}$, where $j = 1, \ldots, n_i$ and $k = 1, \ldots, m$. Let $n_{i,k}$ denote the number of ratings of product $P_i$ that equal to $k$. Then the likelihood of the parameter $\alpha_i$ given a set of observed ratings $r_i^+$ can be expressed as:

$$L(\alpha_i) = p(r_i^+ | \alpha_i) = \prod_{j=1}^{n_i} p(r_{i,j}^+ | \alpha_i) = \prod_{k=1}^{m} (\alpha_{i,k})^{n_{i,k}}.$$

The remaining issue is to derive the maximum likelihood estimation of the parameter $\alpha_i$, which is denoted by $\hat{\alpha}_i$, via maximizing $L(\alpha_i)$. This is equivalent to maximize the log likelihood function:

$$L(\alpha_i) = \log L(\alpha_i) = \sum_{k=1}^{m} n_{i,k} \log \alpha_{i,k}$$

$$= \sum_{k=1}^{m-1} n_{i,k} \log \alpha_{i,k} + n_{i,m} \log(1 - \sum_{k=1}^{m-1} \alpha_{i,k}).$$

By maximizing $L(\alpha_i)$, we obtain the maximum likelihood estimation of $\alpha_i$ as follows:

$$\hat{\alpha}_{i,k} = \frac{n_{i,k}}{n_i}, \text{ for } k = 1, \ldots, m.$$

The inference algorithm is outlined in Algorithm 1.

Remark: The running time of this algorithm is $\Theta(\sum_{i=1}^{N} |r_i^+|)$, or the running time is linear with respect to the number of ratings.

5 Experiments on Synthetic Data

Let us first use synthetic data to examine various factors that influence the accuracy of product quality evaluation. This will help us gain some insights on
Algorithm 1 Inference algorithm for model parameters

Input: A set of ratings of product $P_i$: $r_i^+ = \{r_{i,1}^+, \ldots, r_{i,n_i}^+\}$

Output: $\hat{\alpha}_i$

1: $n_{i,k} = |\{r_{i,j}^+ | r_{i,j}^+ \in r_i^+, r_{i,j}^+ = k\}|$, $k = 1, \ldots, m$
2: for $k = 1$ to $m$ do
3: $\hat{\alpha}_{i,k} = n_{i,k}/n_i$
4: end for

improving applications of online rating systems. We consider an online rating system with a 5-level cardinal rating metric, say $m = 5$. We start our evaluation from the context where the majority rule is used to aggregate ratings. Then we discuss the context corresponding to the average scoring rule.

5.1 Majority rule

We start with the case that all users are honest, then explore when some users misbehave in their ratings.

**Experiment 1: Effect of Model Parameter $\alpha_i$ and Success Probability.** In this experiment, we study the honest ratings case where all users rate honestly. To examine the effect of model parameters, we synthesize the ratings of two products, say $P_i$ and $P_j$, using our model with parameter $\alpha_i = (\frac{4}{35}, \frac{5}{35}, \frac{3}{35}, \frac{2}{35}, \frac{1}{35})$ and $\alpha_j = (\frac{6}{35}, \frac{4}{35}, \frac{3}{35}, \frac{2}{35})$ respectively. Let $n'$ denote the minimum number of ratings we need to extract the true label of a product with success probability of at least $1-\delta$. To examine the effect of the success probability, we vary $1-\delta$ from 0.7 to 0.9. The numerical results of $n'$ are shown in Fig. 1, where the horizontal axis represents the success probability, or $1-\delta$, and the vertical axis shows the corresponding minimum number of ratings, or $n'$. From Fig. 1 we have the following observations. First, when we increase the success probability, or $1-\delta$, the minimum number of ratings increased. In other words, to improve the success probability, we need a larger number of ratings. Note that the minimum number of ratings increases only with a small rate. To extract the true label with the same probability, we need far more ratings for product $P_j$ (with a higher variance in rating scores than $P_i$) than $P_i$.

**Lessons learned:** Increasing the success probability increases the minimum number of ratings. The variance on rating scores of a product has a significant impact on the minimum number of ratings. So it is interesting to explore further, i.e., are these results sensitive to misbehavior? Let us continue to explore.

**Experiment 2: Effect of Random Misbehavior.** We extend the online rating system specified in Exp. 1 to a system where the user population consists of honest users and random misbehaving users. We synthesize the ratings of product $P_i$ from the honest users using our model with parameter $\alpha_i = (\frac{4}{35}, \frac{5}{35}, \frac{3}{35}, \frac{2}{35}, \frac{1}{35})$. Here $n'$ represents the same measure as specified in Exp. 1. To examine the effect of random misbehavior, we vary the fraction of random misbehaving users from 0 to 0.8. The numerical results of $n'$ are shown in Fig. 2 in which three curves correspond to success probability of at least 0.7,
0.8, 0.9 respectively. We can observe that when the fraction of misbehavior is less than 0.6, increasing the fraction of random misbehaving users only increases the minimum number of ratings slightly. When the fraction of misbehavior is larger than 0.6, a small increase in the fraction of random misbehavior can increase the minimum number of ratings remarkably.

**Lessons learned:** When random misbehaving users exist, one can still extract the true label of a product with a price of increasing the minimum number of ratings. A small fraction of random misbehavior only increases the minimum number of ratings slightly. This shows the robustness of *majority rule*. But when this fraction is increased beyond a certain threshold, we need a lot more ratings to compensate. Next, we explore the impact of biased misbehavior in users’ ratings.

**Experiment 3: Effect of Biased Misbehavior.** We extend the online rating system specified in Exp. 1 to the system where the user population consists of honest users and biased misbehaving users who bias toward one specific rating. We synthesize the ratings of product $P_i$ from the honest users using our model with parameter $\alpha_i = (4/35, 5/7, 3/35, 2/35, 1/35)$. The biased misbehaving users bias toward rating 5. From Theorem 4 and 5, one can still extract the true label with a high probability if the fraction of biased misbehaving users is less than 0.407. But if that fraction is larger than 0.407, then it is impossible to extract the true label with high probability. We vary the fraction of biased misbehaving users from 0 to 0.3. Here $n'$ represents the same measure as specified in Exp. 1. The numerical results of $n'$ are shown in Fig. 3. To compare the impact of random misbehavior and biased misbehavior, we draw another curve corresponding to random misbehavior in Fig. 3 also. From Fig. 3 we have the following observations. When the fraction of biased misbehaving users is small, say less than 0.1, increasing the fraction of biased misbehaving users only increases the minimum number of ratings slightly. When that fraction is larger than 0.1, a small increase in the fraction of biased misbehaving users can remarkably increase the minimum number of ratings needed. The *majority rule*
is more robust against random misbehavior compared to biased misbehavior, since the curve corresponding to random misbehavior is quite flat.

**Lessons learned:** *Majority rule* is robust and can resist a small fraction, say less than 0.1, of biased misbehaving users. Also, the majority rule is robust against random misbehavior. The biased misbehavior is more disruptive than the random misbehavior in distorting the evaluation process.

![Figure 3: Impact of biased misbehavior.](image)

**5.2 Average scoring rule**

We start our study with a simple case, say all users rate honestly.

**Experiment 4: Effect of Model Parameters $\alpha_i$ and Success Probability.** We study the case where all users rate honestly. To examine the effect of model parameters, we synthesize the ratings of two products, say $P_i$ and $P_j$, using our model with parameter $\alpha_i = (4/35, 5/7, 3/35, 2/35, 1/35)$ and $\alpha_j = (6/35, 4/7, 4/35, 3/35, 2/35)$ respectively. Let $E_r$ denote the absolute error bound between the average rating and its ground truth value. We set $E_r = 0.5$. Let $n'$ denote the minimum number of ratings a product needs so as to guarantee that the absolute error is less or equal to $E_r$ with probability of at least $1 - \delta$. We vary the success probability $1 - \delta$ from 0.7 to 0.9. The numerical results of $n'$ are shown in Fig. 4, where the horizontal axis represents the success probability, or $1 - \delta$, and the vertical axis shows the corresponding minimum number of ratings, or $n'$. We have the following observations. The minimum number of ratings is insensitive to the given parameters $\alpha_i$ and $\alpha_j$, since these two curves in Fig. 4 are similar. When we increase the success probability, we increase the minimum number of ratings slightly.

Now we compare the performance of the two score aggregation rules we study. We compare the minimum number of ratings for the *majority rule* and the *average scoring rule* and the results are shown in Table 6. One can see that to achieve the same level of confidence, we need a lot more ratings for average scoring rule.
Table 6: Comparison of minimum number of ratings for majority rule and average scoring rule.

| $1 - \delta$ | 0.7 | 0.75 | 0.8 | 0.85 | 0.9 |
|--------------|-----|------|-----|------|-----|
| $n'$ (majority) | 67 | 70 | 77 | 83 | 93 |
| $n'$ (average score) | 642 | 675 | 716 | 769 | 843 |

Table 7: Impact of misbehavior on two rating aggregation rules

| Random misbehavior |
|---------------------|
| fraction of bad users | 0.1 | 0.15 | 0.2 | 0.25 | 0.3 |
| $n'$ (majority) | 89 | 94 | 102 | 112 | 122 |
| $n'$ (average score) | 842 | 882 | 947 | 1018 | 1099 |

| Biased misbehavior |
|---------------------|
| fraction of bad users | 0.1 | 0.15 | 0.2 | 0.25 | 0.3 |
| $n'$ (majority) | 93 | 125 | 181 | 296 | 259 |
| $n'$ (average score) | 1107 | 1452 | 2027 | 3110 | 5600 |

Lessons learned: To obtain a reliable evaluation of product quality using the average scoring rule, we need remarkably more ratings than using the majority rule. The minimum number of ratings is insensitive to model parameters or success probability. This implies that the majority rule requires less ratings to guarantee the same level of confidence, so it is the preferred rating aggregation rule.

Effect of Misbehavior. To examine the effect of misbehavior, we extend the online rating system as specified in Exp. 4: one consists of honest users and random misbehaving users, and the other one consists of honest users and biased misbehaving users. We synthesize the ratings of product $P_i$ from the honest users using our model with parameter $\alpha_i = (\frac{1}{35}, \frac{1}{35}, \frac{a}{35}, \frac{b}{35}, \frac{c}{35})$. The biased misbehaving users bias toward rating 5. For both extensions, we vary the fraction of misbehaving users from 0.1 to 0.3. Here $n'$ and $E_r$ (set to 0.5) represents the same measure as specified in Exp. 4. The numerical results of $n'$ are shown in Table 7. To compare the robustness of the majority rule and the average scoring rule, we show the numerical results of $n'$ corresponding to the majority rule in the same table. From Table 7 we could see that the majority rule is more robust against misbehavior (either random misbehavior or biased misbehavior), since to resist the same fraction of misbehaving users, the average scoring rule requires a lot more ratings.

6 Experiments on real data

In this section, we show the experimental results on three large datasets which we obtain via web crawling: hotel ratings from TripAdvisor, product ratings from Amazon, and seller ratings from Ebay. We first validate our model with
these three datasets, and then explore a number of important questions, i.e., what’s the minimum number of ratings a hotel/product/seller needs to receive so that one can evaluate its quality accurately?

6.1 Datasets

**TripAdvisor.** It is one of the most popular travel websites which assists customers in booking hotels, restaurants, etc, and post travel-related opinions in the form of ratings (or reviews) to hotels, restaurants, etc. We crawled the historical ratings of 11,540 hotels. Table 8 shows the overall statistics of the data set.

**Amazon.** Amazon is one of the most popular recommendation systems that assists customers in product adoption, and one can post opinions on products in the form of ratings (or reviews). We crawled the historical ratings of 32,888 products. Table 8 shows the overall statistics of the data set.

**Ebay.** Ebay is an E-commerce system that assists customers in online product purchasing, and customers can post opinions of sellers in the form of ratings (or reviews) to reflect the reputation of a seller. We crawled the historical ratings of 4,586 sellers. Table 8 shows the overall statistics of the data set.

|                | TripAdvisor | Amazon | Ebay  |
|----------------|-------------|--------|------|
| Number of items| 11,540      | 32,888 | 4,586|
| Total number of ratings | 3,114,876 | 5,066,070 | 19,217,083 |
| Maximum / Minimum on number of ratings | 9930/1 | 24,195/1 | 117,100/1 |
| Mean / Median on number of ratings | 269.9/179 | 154/47 | 4190/1437 |
| Rating metric: \{1, ..., m\} | 1, ..., 5 | 1, ..., 5 | 1, ..., 12 |

Table 8: Statistics for three rating data sets.

6.2 Model validation

We validate our model by showing that if an item meets the requirement on minimum number of ratings, then the evaluation of its quality is indeed reliable.

**Extract minimum number of ratings.** To extract the minimum number of ratings, we need to first estimate model parameters $\alpha_i, \forall i$. Since the accuracy depends on the number of ratings, therefore, we select those items with a large number of ratings, i.e., those with at least 400 ratings. Table 9 shows the number of items in the dataset satisfy this selection criteria. From Table 9 we could see that in total we select 2368 + 2396 + 3307 = 8071 items out. We map the ID of these 8071 items to 1, ..., 8071 and use $P_1, \ldots, P_{8071}$ to denote them. For each selected item $P_i, i \in \{1, \ldots, 8071\}$, we apply Algorithm 1 on their rating sets, so as to obtain an estimation of its model parameter denoted by $\hat{\alpha}_i$. We use
the inferred parameters $\hat{\alpha}_i$ to compute the minimum number of ratings for item $P_i, i \in \{1, \ldots, 8071\}$, setting the success probability $1-\delta=0.8$, and $E_r = 0.5$. We use $n'$ to denote the desired minimum number of ratings an item needs.

**Model validation algorithm design.** The dataset is with time stamps. Based on this, our validation procedure can be stated as follows. For each time stamp we first check if the number of ratings of a hotel (up to that time stamp) meets the requirement on minimum number of ratings. If yes, we then apply a rating aggregation rule to evaluate its quality (based on the historical ratings up to that time stamp), and check if the evaluation is reliable. After we finished the whole process, we compute the faction of evaluations which are reliable. We outline our model validation algorithm in Algorithm 2. We use the following notations in the algorithm. For each hotel, we sort its ratings based on ratings’ times tamps. Let $\tau_i = \{\tau_{i,1}, \ldots, \tau_{i,n_i}\}$ denote a set of sorted ratings of hotel $P_i$, where $i=1, \ldots, 8071$, $n_i$ is the number of ratings of $P_i$, and $\tau_{i,1}$ is the earliest rating. Let $N_{test}$ denote the number of time stamps that an item meets the desired minimum number of ratings, $N_{reliable}$ denote the number of reliable evaluations, and the faction of reliable evaluations is denoted by $f_{reliable}$.

| # of selected items | TripAdvisor | Amazon | Ebay |
|---------------------|-------------|--------|------|
|                     | 2368        | 2396   | 3307 |

Table 9: Number of selected items.

We apply Algorithm 2 on TripAdvisor, Amazon and Ebay datasets respectively, and we compute the faction of reliable evaluations $f_{reliable}$ for each of them separately. The statistics of $n'$ across items are shown in Table 10 which depicts rating aggregation rules, the number of time stamps a hotel meets the requirement of minimum number of ratings $N_{test}$, the number of reliable evaluations $N_{reliable}$, and the faction of reliable evaluations $f_{reliable}$. From Table 10 we have two observations. First, if a product meets the requirement on minimum number of ratings, then the evaluation of its quality is indeed reliable, because the value of $f_{reliable}$ is around 99%. This shows the correctness of our model and our mathematical framework. Secondly, it shows the robustness of the majority rule since we have more time stamps (or ratings) that satisfy the minimum requirement. In fact, the majority rule has around three times (836088 / 296309 $\approx$ 2.82, TripAdvisor), two times (2246804 / 981886 $\approx$ 2.29, Amazon), and 1.28 times (16667935 / 13008919 $\approx$ 1.28, Ebay) the number of ratings that can be used for reliable evaluation. We explore why this is the case in the next subsection.

6.3 Examine minimum number of ratings

We consider the same set of items and we use the same settings specified in the previous sub-section. Again, $n'$ denotes the required minimum number of ratings a product (hotel) needs. The statistics of $n'$ across items are shown in Table 11 where $f_{[0,400]}$ denotes the fraction of items with $n' \in [0, 400]$, and we
Algorithm 2 Algorithm for model validation

Input: $\tau_{i_1}, \ldots, \tau_{i_n}, \hat{\alpha}_{i_1}, \ldots, \hat{\alpha}_{i_n}$, rating aggregation rule $A$

Output: $N_{test}, N_{reliable}, f_{reliable}$

1: $N_{test} \leftarrow 0, N_{reliable} \leftarrow 0$
2: for $\kappa \in \{i_1, \ldots, i_n\}$ do
3: Extract the true quality of $P_i$, say if $A$ is majority rule, then the true quality is $\gamma_i = \arg\max_k \{\hat{\alpha}_{i,k}\}$, and if $A$ is average score rule, then the true quality is $\gamma_i = \sum_k k\hat{\alpha}_{i,k}$
4: $n' \leftarrow$ minimum number of ratings $P_i$ needs
5: for $j = n'$ to $|\tau_\kappa|$ do
6: $N_{test} \leftarrow N_{test} + 1$
7: if $A(\{\tau_{i_1}, \ldots, \tau_{i,j}\})$ reflects the true quality of $P_i$ then
8: $N_{reliable} \leftarrow N_{reliable} + 1$
9: end if
10: end for
11: end for
12: $f_{reliable} = N_{reliable}/N_{test}$

use MR and ASR to denote majority rule and average score rule respectively for brevity. From Table 11 we have the following observations. For the average scoring rule, nearly all items need more than 800 ratings. In fact 99.45% hotels (TripAdvisor), 99.08% products (Amazon), and 99.97% sellers (Ebay) need more than 800 rating. And a large fraction of them need more than 1200 ratings, say 47.55% hotels (TripAdvisor), 61.56% products (Amazon) and 98.34% sellers (Ebay). But for the majority rule, more than half the hotels only need less than 800 ratings, say 58.79% hotels, 88.22% products, 82.8% sellers need less than 800 ratings. What’s more, a large fraction of them need less than 400 ratings, i.e., around 39.74% hotels, 80.51% products and 58.06% sellers. In other words, the minimum number of ratings trend to be of smaller values for the majority

| Aggregation Rule | $N_{test}$ | $N_{reliable}$ | $f_{reliable}$ |
|------------------|-----------|----------------|----------------|
| TripAdvisor      |           |                |                |
| Majority rule    | 836088    | 835822         | 99.97%         |
| Average scoring  | 296309    | 295854         | 99.85%         |
| Amazon           |           |                |                |
| Majority rule    | 2246804   | 2241196        | 99.75%         |
| Average scoring  | 981886    | 971987         | 98.99%         |
| Ebay             |           |                |                |
| Majority rule    | 16705440  | 16667935       | 99.78%         |
| Average scoring  | 13008919  | 12990199       | 99.86%         |

Table 10: Fraction of reliable evaluations when items meet the requirement on minimum number of ratings.
Evaluating rating aggregation rules. We compare robustness of majority rule and average score rule by examining the distribution of minimum number of ratings corresponding to them. We use notation $\Pr[n' \geq n]$ to denote the fraction of items with minimum number of ratings larger or equal to $n$. Fig. 5, 6 and 7 show the numerical results of $\Pr[n' \geq n]$ corresponding to TripAdvisor, Amazon and Ebay dataset respectively, where the horizontal axis shows the number of ratings $n$, and the vertical axis shows the corresponding value of $\Pr[n' \geq n]$. From Fig. 5 we could observe that the distribution curve corresponding to majority rule lies under the curve corresponding to average score rule. In other words, majority rule require less ratings than average score rule. This also holds for Fig. 6 and 7. These results show that majority rule is more robust than average score rule.

Case studies. We examine the following question: among TripAdvisor, Amazon and Ebay, whose ratings are more reliable? We answer this question by comparing the distribution of minimum number of ratings across items for them. Here we fix the rating aggregation rule to be majority rule, since it is more robust than average score rule. Again, we use notation $\Pr[n' \geq n]$ to denote the fraction of items with the minimum number of ratings larger or equal to $n$. The numerical results of $\Pr[n' \geq n]$ are shown in Fig. 8, where the horizontal axis shows the number of ratings $n$, and the vertical axis shows the corresponding value of $\Pr[n' \geq n]$. From Fig. 8 one can observe that, as the web sites varies from Amazon, Ebay, and TripAdvisor, the corresponding distribution curve move up. In other words, the minimum number of ratings corresponding to Amazon is the least and that corresponding to TripAdvisor is the highest. This shows that Amazon requires less ratings than TripAdvisor and Ebay, and TripAdvisor requires more ratings than Amazon and Ebay. Thus, the ratings on Amazon is more reliable than TripAdvisor and Ebay.

Here, we study the fraction of items (products, hotels, sellers) which have sufficient number of ratings. Assume our data sets stated in Table 8 are representative samples from TripAdvisor, Amazon and Ebay. Based on this assumption we explore, the fraction of items which have sufficient number of ratings. We set the desired minimum number of ratings a product/hotel/seller needs as the median of the value of minimum number of ratings obtained in the beginning of this section. Then we test the fraction of products in the data sets stated in Table 8 satisfy the minimum number of ratings. Let $f_s$ denote the fraction of hotels/products/sellers that satisfy the requirement on minimum number of ratings. The numerical results $f_s$ are shown in Table 12, where $N_s$ denote the number of hotels/products/sellers that meet the requirement. From Table 12, we have the following observations: Only 12.8% hotels on TripAdvisor, 21.6% products on Amazon, 74.1% sellers on Ebay, are with sufficient number of ratings. Ratings on Ebay is more sufficient than TripAdvisor and Amazon.

Lessons learned & tips. Assume we use the majority rule to aggregate ratings, we need at least hundreds of ratings for a hotel (TripAdvisor) / product (Amazon) / seller (Ebay), so as to produce a reliable evaluation on the quality of a hotel/product or reputation of a seller. If the number of ratings of a
hotel/product/seller is lower than 100, then we may be in danger of produce unreliable evaluations. If we use the average scoring rule to aggregate ratings, the system needs more ratings (around one thousand of ratings for each hotel/product/seller) than using the majority rule. The majority rule is more robust and reliable than the average scoring rule. The rating on Amazon is more reliable than TripAdvisor and Ebay, since to produce a reliable evaluation, Amazon needs less rating than TripAdvisor and Ebay. Ratings on Ebay are more sufficient than TripAdvisor and Amazon.

6.4 Inferring minimum number of ratings

In real world applications, it may be expensive or sometimes even impossible to obtain the exact desired minimum number of ratings for a product. Here we tackle this challenge by designing an efficient algorithm to infer the desired minimum number of ratings, stated in Algorithm 3. We show the effectiveness of Algorithm 3 by performing empirical experiments on TripAdvisor, Amazon and Ebay dataset.
Table 12: Fraction of hotels satisfy the requirement of minimum number of ratings.

|      | \( n' \) | \# of items | \( N_s \) | \( f_s \) |
|------|----------|-------------|----------|--------|
| TripAdvisor | 573      | 11540       | 1481     | 12.8%  |
| Amazon    | 146      | 32888       | 7119     | 21.6%  |
| Ebay      | 356      | 4586        | 3397     | 74.1%  |

Inferring minimum number of ratings algorithm design. In real world web services, ratings are with time stamps. For a given time stamp, we infer the desired minimum number of ratings for a product based on the historical ratings (up to the given time stamp) of that product. To infer the desired minimum of ratings, we first infer the model parameter by applying Algorithm 1 on the the historical ratings. Then based on the inferred model parameter, we compute the desired minimum number of ratings. We outline our algorithm for inferring minimum number of ratings in Algorithm 3 where \( \tau_i(t) = \{ \tau_{i,1}, \ldots, \tau_{i,t} \} \) denote a set of historical ratings (up to time stamp \( t \)) of hotel \( P_i \), and \( \tau_{i,1} \) is the earliest rating. We use \( \hat{n}' \) to denote the inferred minimum number of ratings.

Empirical evaluations. We illustrate the effectiveness of Algorithm 3 by showing that if a product meets the inferred minimum number of ratings, then the evaluations of its quality is indeed reliable. We work on the same dataset and use the same notations, say with Section 6.2. The datasets are with time stamps. For each time stamp, we first infer the minimum number of ratings of that product by performing Algorithm 3 on its historical ratings (up to that time stamp). And then we check whether the number of ratings is larger than the inferred minimum number of ratings. If yes, then the number of test time stamps \( N_{test} \) will be increased by one, and at the meantime, we further evaluate the quality of that product based on its ratings up to that time stamp, if the evaluated quality reflects the true quality (obtained by step 3 of Algorithm 2), then the number of reliable evaluations \( N_{reliable} \) is increased by one. We consider
Here. We perform experiments on TripAdvisor, Amazon and Ebay datasets. The experiment results of $N_{test}$ and $N_{reliable}$ are shown Table 13 where $f_{reliable} = N_{reliable}/N_{test}$. From Table 13 we could observe that if a products meets the requirement on the inferred minimum number of ratings, then the evaluation of its quality is reliable because the value of $f_{reliable}$ is 97.27% for TripAdvisor, 98.50% for Amazon, 99.56% for Ebay. This shows the effectiveness of our algorithm for inferring minimum number of ratings.

**Lessons learned.** Our algorithm of inferring the minimum number of ratings is effective in producing reliable evaluations of product quality. Our algorithm can be deployed into real world applications of web services.

### Algorithm 3 Infer minimum number of ratings

**Input:** A set of historical ratings $\tau_i(t) = \{\tau_{i,1}, \ldots, \tau_{i,t}\}$, rating aggregation rule $A$, $\delta$, $Er$

**Output:** Inferred minimum of ratings $\hat{n}'$.

1. Infer model parameter $\hat{\alpha}_i = \{\hat{\alpha}_{i,1}, \ldots, \hat{\alpha}_{i,m}\}$ by performing Algorithm 1 on the historical rating set $\tau_i(t)$

2. if $A$ is majority rule then

   3. $\hat{\alpha}_{i,\ell} \leftarrow \max\{\hat{\alpha}_{i,1}, \ldots, \hat{\alpha}_{i,m}\}$
   
   4. $\hat{\alpha}_i' \leftarrow \max\{\hat{\alpha}_{i,1}, \ldots, \hat{\alpha}_{i,m}\} \setminus \hat{\alpha}_{i,\ell}$
   
   5. $\hat{n}' \leftarrow \frac{12E_i^\ell}{(\hat{\alpha}_{i,\ell} - \hat{\alpha}_i')^2} \ln \frac{m}{\delta}$

3. end if

4. if $A$ is average score rule then

   5. $\hat{\gamma}_i \leftarrow \sum_k \hat{\alpha}_{i,k}$, $\epsilon \leftarrow \frac{-\sqrt{m\gamma_i} + \sqrt{m\gamma_i + 4mEr}}{2m}$
   
   6. $\hat{n}' \leftarrow \frac{3}{\epsilon} \ln \frac{2m}{\epsilon}$

7. end if

8. Compute the minimum number of ratings $\hat{n}'$ based

| Source     | $N_{test}$ | $N_{reliable}$ | $f_{reliable}$ |
|------------|------------|----------------|----------------|
| TripAdvisor| 907810     | 883030         | 97.27%         |
| Amazon     | 2262036    | 2228173        | 98.50%         |
| Ebay       | 16516962   | 16444179       | 99.56%         |

Table 13: Fraction of reliable evaluations when products meet the requirement on inferred minimum number of ratings.

## 7 Related Work

Several works have investigated the problem of rating aggregation in online rating systems. For example, in [19][4], authors proposed reputation based weighted average score rules that compute the average rating of products weighted on the reputation score of users. Other representative works on rating aggregation can
be found in [13, 9, 10, 20]. However, in real online rating system based web services, *average scoring rule* and *majority rule* are two most widely deployed rating aggregation rules, because they are simple, easy to deploy and more importantly, tested to perform well for many web services.

Several works have investigated fraud detection. For example, authors in [11] explored review spam detection in online reviewing systems. In [5, 21] authors studied fraud detection in trading communities. Also, fraud detection in recommendation systems was investigated in [12, 16]. In our paper, we provide a probabilistic model and formal analysis of users misbehavior in online ratings systems, and show the condition in which online rating systems may not be able to reflect the true quality of products.

Online rating systems are widely deployed in recommendation systems [18, 2]. Recommendation systems were introduced since the seminal work on collaborative filtering [8, 17]. In general, recommendation systems interpret ratings as preferences of users, and try to make personalized recommendations by taking into account the preferences of users. Researchers investigate various algorithmic and complexity issues in designing recommendation systems [7, 18, 2]. The important distinction of our work is that we treat rating as product quality assessment and we consider the condition where we can reveal the true quality with high probability.

## 8 Conclusion

We present a mathematical framework to analyze online rating systems that are widely deployed in web services such as TripAdvisor, Amazon, eBay, etc. We first present a novel mathematical model to specify users’ rating behavior. We then formally analyze a general model of online rating systems. Through this analysis, we derived theoretical bounds on the desired minimum of the ratings we need to produce a reliable evaluation on the quality of products. We extend our model to accommodate users’ misbehavior (i.e., cheating) in product rating, and show that the *majority rule* is more robust and insensitive to misbehavior as compared with the *average scoring rule*. Based on our framework, we also propose an practical algorithm to infer the minimum number of ratings, which is effective in produce reliable product quality evaluations in real-world. We perform experiments on both synthetic data and three *large* real-world data sets (from TripAdvisor, Amazon, Ebay). Number of interesting observations were found, e.g., generally, hundreds of ratings are sufficient to reflect the true quality of a hotel / product on Amazon / Ebay, or the true reputation of a seller on Ebay. Ratings on Ebay are more sufficient than TripAdvisor and Amazon. We believe that our models and methodology can be used as important building blocks to refine and improve online rating systems for various web services.
References

[1] Easily rigged travel review sites labeled untrustworthy. http://www.tecca.com/columns/travel-review-site-scams-travel-tech/, 2012.

[2] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE TKDE, 17(6):734–749, 2005.

[3] C. Bishop. Pattern recognition and machine learning. springer New York, 2006.

[4] M. Chen and J. Singh. Computing and using reputations for internet ratings. In Proc. of ACM EC, 2001.

[5] C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In Proc. of ACM EC, 2000.

[6] W. Feller. Generalization of a probability limit theorem of cramer. Trans. Amer. Math. Soc, 54(3):361–372, 1943.

[7] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl. Evaluating collaborative filtering recommender systems. ACM TOIS, 22(1):5–53, 2004.

[8] W. Hill, L. Stead, M. Rosenstein, and G. Furnas. Recommending and evaluating choices in a virtual community of use. In Proc. of ACM CHI, 1995.

[9] R. Jin and L. Si. A study of methods for normalizing user ratings in collaborative filtering. In Proc. of ACM SIGIR, 2004.

[10] R. Jin, L. Si, C. Zhai, and J. Callan. Collaborative filtering with decoupled models for preferences and ratings. In Proc. of ACM CIKM, 2003.

[11] N. Jindal and B. Liu. Analyzing and detecting review spam. In Proc. of IEEE ICDM, 2007.

[12] S. Lam and J. Riedl. Shilling recommender systems for fun and profit. In Proc. of WWW, 2004.

[13] H. W. Lauw, E.-P. Lim, and K. Wang. Quality and leniency in online collaborative rating systems. ACM Trans. Web, 6(1):4:1–4:27, 2012.

[14] J. Matoušek and J. Vondrák. The probabilistic method.

[15] M. Mitzenmacher and E. Upfal. Probability and computing: Randomized algorithms and probabilistic analysis. Cambridge University Press, 2005.

[16] B. Mobasher, R. Burke, and J. Sandvig. Model-based collaborative filtering as a defense against profile injection attacks. In Proceedings of the National Conference on Artificial Intelligence, 2006.
[17] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the ACM conference on Computer supported cooperative work, 1994.

[18] P. Resnick and H. Varian. Recommender systems. Communications of the ACM, 40(3):56–58, 1997.

[19] T. Riggs and R. Wilensky. An algorithm for automated rating of reviewers. In Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, pages 381–387. ACM, 2001.

[20] J. Traupman and R. Wilensky. Collaborative quality filtering: Establishing consensus or recovering ground truth? Advances in Web Mining and Web Usage Analysis, pages 73–86, 2006.

[21] J. Zhang and R. Cohen. Trusting advice from other buyers in e-marketplaces: the problem of unfair ratings. In Proc. of ACM EC, 2006.

Appendix

Proof of Lemma 1

Let \( \rho = (\rho_1, \ldots, \rho_m) \) denote the rating distribution of the user who assigns score \( r_{i,j}^+ \). Recall that the probability distribution of \( \rho \) is Dirichlet(\( \alpha_i \)), and \( \Pr[r_{i,j}^+ = k | \rho] = \rho_k \), where \( k \in \{1, \ldots, m\} \). Then we have:

\[
\Pr[r_{i,j}^+ = k] = \int p(\rho) \Pr[r_{i,j}^+ = k | \rho] d\rho = \int \prod_{k=1}^{m} \frac{\Gamma(\sum_{\kappa=1}^{m} \alpha_{\kappa})}{\Gamma(\alpha_{\kappa})} \rho_{\kappa}^{\alpha_{\kappa} - 1} \rho_{j,1} \ldots \rho_{j,m} d\rho = \alpha_{i,k}, \text{ for } k = 1, \ldots, m,
\]

so this proof is completed.

Proof of Theorem 1

Let us state a theorem on bounding the tail probability, that will be used in later derivation.

**Theorem 9 (Chernoff Bound [15])** Let \( X_1, \ldots, X_n \) be \( n \) independent random variables, with \( X_i = 1 \) with probability \( p_i \) and 0 otherwise. Let \( X = \sum_{i=1}^{n} X_i \) and let \( \mu = E[X] \). Then for each \( \epsilon \geq 0 \), we have

\[
\Pr[X \geq (1 + \epsilon)\mu] \leq \exp \left( - \min(\epsilon, \epsilon^2) \mu / 3 \right),
\]

\[
\Pr[X \leq (1 - \epsilon)\mu] \leq \exp \left( - \epsilon^2 \mu / 3 \right).
\]
Now we can prove Theorem 1 by applying Theorem 9 as follows. Let $n_{i,k} = | \{ r_{i,j} | r_{i,j} \in r_i, r_{i,j} = k \} |$ denote the number of ratings that equal to $k$. By basic probability arguments, we can derive the probability of failing to extract the true label as follows:

\[ \Pr[ \hat{\ell}_i \neq \ell_i, n_{i,\ell_i} < \alpha_{i,\ell_i} + \tilde{\alpha}_i n_i ] + \]
\[ \Pr[ \hat{\ell}_i \neq \ell_i, n_{i,\ell_i} \geq \alpha_{i,\ell_i} + \tilde{\alpha}_i n_i ] \]
\[ \leq \Pr[ n_{i,\ell_i} < \frac{\alpha_{i,\ell_i} + \tilde{\alpha}_i}{2} n_i ] + \]
\[ \Pr[ \exists k \neq \ell_i \text{ that } n_{i,k} > n_{i,\ell_i}, n_{i,\ell_i} \geq \frac{\alpha_{i,\ell_i} + \tilde{\alpha}_i}{2} n_i ] \]
\[ \leq \Pr[ n_{i,\ell_i} < \frac{\alpha_{i,\ell_i} + \tilde{\alpha}_i}{2} n_i ] + \sum_{k \neq \ell_i} \Pr[ n_{i,k} > \frac{\alpha_{i,\ell_i} + \tilde{\alpha}_i}{2} n_i ] . \tag{5} \]

Let us now proceed to individually derive these two terms of the above inequality.

One can apply the Chernoff Bound to bound the first term of Inequality (5). Observe that $E[n_{i,k}] = n_i \alpha_{i,k}$. Based on this observation, we can write down the first term of Inequality (5) as follows:

\[ \Pr[ n_{i,\ell_i} < \frac{\alpha_{i,\ell_i} + \tilde{\alpha}_i}{2} n_i ] = \Pr[ n_{i,\ell_i} < E[n_{i,\ell_i}] \left( 1 - \frac{\alpha_{i,\ell_i} - \tilde{\alpha}_i}{2\alpha_{i,\ell_i}} \right) ] , \]

by applying Chernoff Bound we have

\[ \Pr[ n_{i,\ell_i} \leq \frac{\alpha_{i,\ell_i} + \tilde{\alpha}_i}{2} n_i ] \leq \exp \left( -\frac{(\alpha_{i,\ell_i} - \tilde{\alpha}_i)^2}{3(2\alpha_{i,\ell_i})^2 E[n_{i,\ell_i}]} \right) \]
\[ = \exp \left( -\frac{(\alpha_{i,\ell_i} - \tilde{\alpha}_i)^2}{12\alpha_{i,\ell_i}} n_i \right) , \]

by substituting $n_i$ with Inequality (3), we have

\[ \Pr[ n_{i,\ell_i} \leq \frac{\alpha_{i,\ell_i} + \tilde{\alpha}_i}{2} n_i ] \leq \frac{\delta}{m} . \tag{6} \]

The remaining issue is to derive the last term of Inequality (5). Since $E[n_{i,k}] = n_i \alpha_{i,k}$. Based on this observation, we can write down the last term of Inequality (5) as follows:
as follows:

\[
\sum_{k=1}^{m} \Pr \left[ n_{i,k} > \frac{\alpha_{i,\ell_i} + \bar{\alpha}_i}{2} n_i \right]
\]

\[
= \sum_{k=1}^{m} \Pr \left[ n_{i,k} > E[n_{i,k}] \left( 1 + \frac{\alpha_{i,\ell_i} + \bar{\alpha}_i - 2\alpha_{i,k}}{2\alpha_{i,k}} \right) \right]
\]

\[
\leq \sum_{k=1}^{m} \exp \left( - \frac{E[n_{i,k}]}{3} \right) \cdot \min \left\{ \frac{\alpha_{i,\ell_i} + \bar{\alpha}_i - 2\alpha_{i,k}}{2\alpha_{i,k}}, \frac{(\alpha_{i,\ell_i} + \bar{\alpha}_i - 2\alpha_{i,k})^2}{(2\alpha_{i,k})^2} \right\},
\]

(7)

where the last step is obtained by applying Chernoff Bound. Since \( \alpha_{i,\ell_i} + \bar{\alpha}_i - 2\alpha_{i,k} \geq \alpha_{i,\ell_i} - \bar{\alpha}_i \geq 0 \), thus we have

\[
\min \left\{ \frac{\alpha_{i,\ell_i} + \bar{\alpha}_i - 2\alpha_{i,k}}{2\alpha_{i,k}}, \frac{(\alpha_{i,\ell_i} + \bar{\alpha}_i - 2\alpha_{i,k})^2}{(2\alpha_{i,k})^2} \right\} \geq \frac{\alpha_{i,\ell_i} - \bar{\alpha}_i}{2\alpha_{i,k}} \min \left\{ 1, \frac{\alpha_{i,\ell_i} - \bar{\alpha}_i}{2\bar{\alpha}_i} \right\}, \quad \forall k \neq \ell_i,
\]

applying this inequality to Inequality (7), we obtain

\[
\sum_{k=1}^{m} \Pr \left[ n_{i,k} > \frac{\alpha_{i,\ell_i} + \bar{\alpha}_i}{2} n_i \right]
\]

\[
\leq \sum_{k=1}^{m} \exp \left( - \frac{(\alpha_{i,\ell_i} - \bar{\alpha}_i)n_i}{6} \cdot \min \left\{ 1, \frac{\alpha_{i,\ell_i} - \bar{\alpha}_i}{2\bar{\alpha}_i} \right\} \right).
\]

(8)

We can further simplify this inequality under the following two cases.

**Case 1:** \( \bar{\alpha}_i < \alpha_{i,\ell_i}/3 \). In this case, \( (\alpha_{i,\ell_i}/\bar{\alpha}_i)/2 > 1 \). Therefore \( \min \left\{ 1, (\alpha_{i,\ell_i} - \bar{\alpha}_i)/(2\bar{\alpha}_i) \right\} = 1 \). Applying this to Inequality (8) and substituting \( n_i \) with Inequality (3), we have

\[
\sum_{k=1}^{m} \Pr \left[ n_{i,k} > \frac{\alpha_{i,\ell_i} + \bar{\alpha}_i}{2} n_i \right] \leq \sum_{k=1}^{m} \exp \left( - \frac{2\alpha_{i,\ell_i}}{\alpha_{i,\ell_i} - \bar{\alpha}_i} \ln \frac{m}{\delta} \right)
\]

\[
\leq \sum_{k=1}^{m} \exp \left( - 2 \ln \frac{m}{\delta} \right) = \frac{m - 1}{m^2 \delta^2}.
\]

(9)
Case 2: $\tilde{\alpha}_i \geq \alpha_{i,\ell}/3$. In this case, $(\alpha_{i,\ell} - \tilde{\alpha}_i)/(2\tilde{\alpha}_i) \leq 1$. We have $\min\{1, (\alpha_{i,\ell} - \tilde{\alpha}_i)/(2\tilde{\alpha}_i)\} = (\alpha_{i,\ell} - \tilde{\alpha}_i)/(2\tilde{\alpha}_i)$. Applying this equation to Inequality (5) and substituting $n_i$ with Inequality (3), we have

$$m \sum_{k=1}^{m} \Pr \left[ n_{i,k} > \frac{\alpha_{i,\ell} + \tilde{\alpha}_i}{2} n_i \right] \leq \sum_{k=1}^{m} \exp \left( -\frac{\alpha_{i,\ell} \ln \frac{m}{\delta}}{\tilde{\alpha}_i} \right) \leq \sum_{k=1}^{m} \exp \left( -\ln \frac{m}{\delta} \right) = \frac{m - 1}{m} \delta. \quad (10)$$

Combining Inequality (9) and Inequality (10), we have

$$m \sum_{k=1}^{m} \Pr \left[ n_{i,k} > \frac{\alpha_{i,\ell} + \tilde{\alpha}_i}{2} n_i \right] \leq \frac{m - 1}{m} \delta. \quad (11)$$

Applying Inequality (6) and Inequality (11) to Inequality (5), we complete the proof.

**Proof of Theorem 2**

Let us state a theorem on bounding the tail probability, that will be used in later derivation.

**Theorem 10** ([14, 6]) Let $X$ be a sum of independent random variables, each attaining values in $[0, 1]$, and let $\sigma = \sqrt{\text{Var}[X]} \geq 200$. Then for all $t \in [0, \frac{\alpha_2}{100}]$, we have

$$\Pr[X \geq E[X] + t] \geq c \exp \left( -\frac{t^2}{3\sigma^2} \right),$$

for a suitable constant $c > 0$.

Now we can prove Theorem 2 by applying Theorem 10 as follows. Recall that $n_{i,k}$ is the number of ratings that equal $k$. Let $\tilde{\ell}_i = \arg \min_k \{|\alpha_{i,k} - \tilde{\alpha}_i|\}$. We can derive a general lower bound of the probability that we fail to extract the true label as follows:

$$\Pr \left[ \hat{\ell}_i \neq \ell_i \right] \geq \Pr \left[ n_{i,\tilde{\ell}_i} > n_{i,\ell_i} \right].$$

Let $R_{i,j}, j \in \{1, \ldots, n_i\}$ denote a set of random variables with

$$R_{i,j} = \begin{cases} 1, & \text{with probability } \Pr[r_{i,j} = \tilde{\ell}_i] \\ 0, & \text{with probability } \Pr[r_{i,j} = \ell_i] \\ 1/2, & \text{otherwise} \end{cases},$$

where the probability mass function of $r_{i,j}$ is derived in Lemma 1. Observe that $R_i > n_i/2$ if and only if the number of ratings equal $\ell_i$ is smaller than the number
of ratings equal \( \bar{\ell}_t \), or \( n_{i,\bar{\ell}_t}, n_{i,\bar{\ell}_t} \). Thus we have \( \Pr[\ell_i \neq \ell_t] \geq \Pr[R_i > \frac{n_i}{2}] \). In the following, we seek to complete the proof by showing \( \Pr[R_i > n_i/2] \geq \Omega(\delta) \).

Here we apply Theorem 10 to give a lower bound of \( \Pr[R_i > n_i/2] \). First we can express the expectation and variance of \( R_i \) as follows:

\[
E[R_i] = n_i/2 - (\alpha_{i,\bar{\ell}_t} - \bar{\alpha}_i)n_i/2, \\
\text{Var}[R_i] = (\alpha_{i,\bar{\ell}_t} + \bar{\alpha}_i - (\alpha_{i,\bar{\ell}_t} - \bar{\alpha}_i)^2) n_i/4.
\]

Let \( t = n_i(\alpha_{i,\bar{\ell}_t} - \bar{\alpha}_i)/2 + \epsilon \), where \( 0 < \epsilon < 10^{-10} \). Observe that \( E[R_i] + t = n_i/2 + \epsilon > n_i/2 \), thus \( \Pr[R_i > \frac{n_i}{2}] \geq \Pr[R_i > E[R_i] + t] \). Before applying Theorem 10 to derive a lower bound of \( \Pr[R_i > E[R_i] + t] \), we need to check whether some conditions specified in Theorem 10 are satisfied.

Let us first check whether \( \sqrt{\text{Var}[R_i]} \geq 200 \) holds. Since

\[
\frac{100}{101} \alpha_{i,\bar{\ell}_t} \leq \bar{\alpha}_i \Leftrightarrow \alpha_{i,\bar{\ell}_t} - \bar{\alpha}_i \leq \frac{1}{201}(\alpha_{i,\bar{\ell}_t} + \bar{\alpha}_i).
\]  

Based this observation, we can derive a lower bound of \( \alpha_{i,\bar{\ell}_t} + \bar{\alpha}_i - (\alpha_{i,\bar{\ell}_t} - \bar{\alpha}_i)^2 \) as follow:

\[
\alpha_{i,\bar{\ell}_t} + \bar{\alpha}_i - (\alpha_{i,\bar{\ell}_t} - \bar{\alpha}_i)^2 \geq \alpha_{i,\bar{\ell}_t} + \bar{\alpha}_i - \left(\frac{\alpha_{i,\bar{\ell}_t} + \bar{\alpha}_i}{201}\right)^2 \\
\geq \frac{200 \times 202}{201^2} \left(\alpha_{i,\bar{\ell}_t} + \frac{100}{101} \alpha_{i,\bar{\ell}_t}\right) \geq 1.99 \alpha_{i,\bar{\ell}_t}.
\]

By applying Inequality 13, the lower bound of \( \text{Var}[R_i] \) is

\[
\text{Var}[R_i] \geq \frac{1.99}{4} n_{i,\bar{\ell}_t} \alpha_{i,\bar{\ell}_t} = O\left(\frac{1.99}{4} \left(\frac{\alpha_{i,\bar{\ell}_t}^2}{\alpha_{i,\bar{\ell}_t} - \bar{\alpha}_i)^2\right) \ln \frac{1}{\delta}\right) \\
\geq O\left(\frac{1.99 \times 101^2}{4} \ln \frac{1}{\delta}\right).
\]

By making the left side of this inequality larger or equal to 40000, we guarantee that \( \sqrt{\text{Var}[R_i]} \geq 200 \), or \( \delta \leq O(e^{-7.9}) \). Hence, by letting \( \eta = O(e^{-7.9}) \), then for any \( \delta \leq \eta \), we have \( \sqrt{\text{Var}[R_i]} \geq 200 \).

Let us continue to check whether \( t \in [0, \text{Var}[R_i]/100] \). By apply Inequality 12, we could derive an upper bound of \( t/\text{Var}[R_i] \) as follows:

\[
\frac{t}{\text{Var}[R_i]} = \frac{2(\alpha_{i,\bar{\ell}_t} - \bar{\alpha}_i)}{\alpha_{i,\bar{\ell}_t} + \bar{\alpha}_i - (\alpha_{i,\bar{\ell}_t} - \bar{\alpha}_i)^2} + \frac{\epsilon}{\text{Var}[R_i]} \\
\leq 2 \left(\frac{\alpha_{i,\bar{\ell}_t} + \bar{\alpha}_i}{\alpha_{i,\bar{\ell}_t} - \bar{\alpha}_i} - (\alpha_{i,\bar{\ell}_t} - \bar{\alpha}_i)\right)^{-1} + 10^{-14} \\
\leq 2 \left(\frac{201}{201} - \frac{1}{201}(\alpha_{i,\bar{\ell}_t} + \bar{\alpha}_i)\right)^{-1} + 10^{-14} \leq \frac{1}{100}.
\]

Thus we have that \( t \in [0, \text{Var}[R_i]/100] \).
Finally, we can apply Theorem 10 to derive a lower bound of \( \Pr \left[ R_i > \frac{n_i}{2} \right] \) as follows:

\[
\Pr \left[ R_i > \frac{n_i}{2} \right] \geq \Pr \left[ R_i \geq E[R_i] + t \right] \geq c \exp \left( -\frac{t^2}{3 \text{Var}[R_i]} \right)
\]

\[
= c \exp \left( -\frac{1}{3} \frac{n_i(\alpha_{i,\ell_i} - \bar{\alpha}_i)^2}{\alpha_{i,\ell_i} + \bar{\alpha}_i - (\alpha_{i,\ell_i} - \bar{\alpha}_i)^2} \right)
\]

\[
\geq c \exp \left( -\frac{1}{3} \frac{n_i(\alpha_{i,\ell_i} - \bar{\alpha}_i)^2}{1.99 \alpha_{i,\ell_i}} \right) \geq \Omega(\delta)
\]

Thus, \( \Pr \left[ \ell_i \neq \ell_i \right] \geq \Pr \left[ R_i > \frac{n_i}{2} \right] \geq \Omega(\delta) \).  

**Proof of Theorem 3**

We prove this theorem by extending Theorem 1. First let us present a probabilistic interpretation of the random misbehaving users as follows: with probability \( f_i \) we generate a misbehaving user to rate product \( P_i \), and with probability \( 1 - f_i \) we generate an honest user to rate product \( P_i \). This interpretation describes a new generative process for ratings with random misbehaving users. Let \( r_i \) be a rating generated by that new generative process. Observe that \( \Pr[r_i = k | \text{random misbehavior}] = \frac{1}{m} \) and \( \Pr[r_i = k | \text{sincere}] = \alpha_{i,k}, \) where \( k = 1, \ldots, m \). Then we can compute the probability mass function of \( r_i \) as follows:

\[
\Pr[r_i = k] = \frac{f_i}{m} + (1 - f_i)\alpha_{i,k}, \quad \text{for } k = 1, \ldots, m.
\]

It follows that \( \Pr[r_i = \ell_i] = \arg \max \{ \Pr[r_i = k] \} \) and \( f_i/m + (1 - f_i)\bar{\alpha}_i \) are the largest and second largest among \( \Pr[r_i = 1], \ldots, \Pr[r_i = m] \) respectively. Then, by a similar derivation with Theorem 1, we finish this proof.

**Proof of Theorem 4**

We prove this theorem by extending Theorem 1. First let us present a probabilistic interpretation of the bias misbehaving users as follows: with probability \( f'_i \) we generate a biased misbehaving user to rate product \( P_i \), and with probability \( 1 - f'_i \) we generate a honest user to rate product \( P_i \). This interpretation describes a new generative process for ratings with biased misbehaving users. Let \( r'_i \) be a rating generated by that new generative process. Observe that \( \Pr[r'_i = k | \text{biased misbehavior}] = 1 \) with \( k = \ell'_i \) and 0 otherwise, and \( \Pr[r'_i = k | \text{sincere}] = \alpha_{i,k}, \forall k \). Then we have:

\[
\Pr[r'_i = k] = \begin{cases} f'_i + (1 - f'_i)\alpha_{i,\ell'_i}, & \text{for } k = \ell'_i \\ (1 - f'_i)\alpha_{i,k}, & \text{otherwise} \end{cases}
\]

We have two cases to consider:

**Case 1**: \( \ell'_i \neq \ell_i \). Observe that

\[
f'_i > \frac{\alpha_{i,\ell_i} - \alpha_{i,\ell'_i}}{1 + \alpha_{i,\ell_i} - \alpha_{i,\ell'_i}} \Leftrightarrow f'_i + (1 - f'_i)\alpha_{i,\ell'_i} > (1 - f'_i)\alpha_{i,\ell_i}.
\]
Thus \( \Pr[r'_i = \ell'_i] = f'_i + (1 - f'_i)\alpha_{i,\ell'_i} \) and \( \Pr[r'_i = \ell_i] = (1 - f'_i)\alpha_{i,\ell_i} \) are the largest and second largest among \( \Pr[r'_i = 1], \ldots, \Pr[r'_i = m] \) respectively. Therefore, with a large enough value of \( n_i \), the number of ratings that equal to \( \ell'_i \) will be the largest with high probability. Then by a similar derivation with Theorem 1, we could finish the proof.

**Case 2:** \( \ell'_i = \ell_i \). Observe that \( \Pr[r'_i = \ell_i] = f'_i + (1 - f'_i)\alpha_{i,\ell_i} \) and \( (1 - f'_i)\tilde{\alpha}_i \) are the largest and second largest among \( \Pr[r'_i = 1], \ldots, \Pr[r'_i = m] \) respectively. Thus we extract the label \( \ell_i \) with a high probability with large enough value of \( n_i \). Finally, by a similar derivation with Theorem 1, one completes the proof.

### 8.1 Proof of Theorem 6

Let \( \beta_{i,k} = \sum_{j=1}^m \alpha_{i,j} \). Observe that \( \gamma_i = \sum_{k=1}^m k\alpha_{i,k} = \sum_{k=1}^m \beta_{i,k} \). Recall that we use \( n_{i,k} \), \( k = 1, \ldots, m \), to denote the number of ratings that equal to \( k \). Let \( n'_{i,k} = \sum_{j=1}^m n_{i,j} \), \( k = 1, \ldots, m \), denote the number of ratings which are no less than \( k \). Then we have

\[
\sum_{j=1}^m r_{i,j} = \sum_{k=1}^m k n_{i,k} = \sum_{k=1}^m n'_{i,k}.
\]

Then it follows that our problem can be formulated to bound

\[
|\hat{r}_i - \gamma_i| = \left| \sum_{k=1}^m \frac{n'_{i,k}}{n_i} - \sum_{k=1}^m \beta_{i,k} \right|.
\]

In the following, we will first seek to bound \( |n'_{i,k}/n_i - \beta_{i,k}| \), and then we apply this bound to complete our proof.

We bound \( |n'_{i,k}/n_i - \beta_{i,k}| \) by the following claim:

- **Claim 1:** For each \( k \in \{1, \ldots, m\} \), if Inequality (14) holds, then

\[
\left| \frac{n'_{i,k}}{n_i} - \beta_{i,k} \right| \leq \begin{cases} 
\epsilon^2, & \beta_{i,k} \leq \epsilon^2 \\
\epsilon \sqrt{\beta_{i,k}}, & \text{otherwise}
\end{cases}
\]

holds with probability at least \( 1 - \delta/m \).

Let us prove this claim here. Let \( I_{j,k}, j = 1, \ldots, n_i, k = 1, \ldots, m \) be a set of indicator random variables such that

\[
I_{j,k} = \begin{cases} 
1, & \text{if } r^+_{i,j} \geq k \\
0, & \text{otherwise}
\end{cases}
\]

From Lemma 1 we have that \( \Pr[r^+_{i,j} = k] = \alpha_{i,k} \), where \( k = 1, \ldots, m \) and \( j = 1, \ldots, n_i \). Thus \( \Pr[r^+_{i,j} \geq k] = \beta_{i,k} \). Therefore, we could have \( \Pr[I_{j,k} = 1] = \beta_{i,k} \).
It follows that $E \left[ \sum_{j=1}^{n_i} I_{j,k} \right] = n_i \beta_{i,k}$. Observe that $n_{i,k}' = \sum_{j=1}^{n_i} I_{j,k}$. Let us consider a special case, $\beta_{i,k} = 0$ first. For this special case, by a simple check, we could see that claim 1 holds. Then we consider the case, $\beta_{i,k} \geq \epsilon^2$. Observe that for this case the inequality $0 < \epsilon / \sqrt{\beta_{i,k}} < 1$ holds. Based on this observation, we can further show:

$$
\Pr \left[ \left\| \frac{n_{i,k}'}{n_i} - \beta_{i,k} \right\| \leq \epsilon \sqrt{\beta_{i,k}} \right] \\
\geq 1 - \Pr \left[ \left\| \frac{n_{i,k}'}{n_i} - \beta_{i,k} \right\| \geq \epsilon \sqrt{\beta_{i,k}} \right] \\
= 1 - \Pr \left[ \sum_{j=1}^{n_i} I_{j,k} - n_i \beta_{i,k} \right] \geq \frac{\epsilon}{\sqrt{\beta_{i,k}}} n_i \beta_{i,k} \right] \\
\geq 1 - 2 \exp \left( -\frac{n_i \epsilon^2}{3} \right) \geq 1 - \frac{\delta}{m},
$$

where the last two steps are obtained by applying Chernoff Bound and by substituting $n_i$ with Inequality (14) respectively. Finally, we consider the case, $0 < \beta_{i,k} < \epsilon^2$. Observe that for this case the inequality $\epsilon / \sqrt{\beta_{i,k}} > 1$ holds. Then by a derivation as the previous case, we obtain

$$
\Pr \left[ \left\| \frac{n_{i,k}'}{n_i} - \beta_{i,k} \right\| \leq \epsilon^2 \right] \geq 1 - \frac{\delta}{m}
$$

holds. Hence we proved Claim 1.

Let $E_k$ denote the event that Inequality (14) holds for a specific value $k$. Then based on claim 1, we have $\Pr [E_k] \geq \delta / m$. Then we can derive the probability that events $E_1, \ldots, E_m$ all holds as follows:

$$
\Pr \left[ \bigcap_{k=1}^{m} E_k \right] = 1 - \Pr \left[ \bigcup_{k=1}^{m} \overline{E_k} \right] \geq 1 - \sum_{k=1}^{m} \Pr [ \overline{E_k} ] \\
\geq 1 - \sum_{k=1}^{m} \frac{\delta}{m} \geq 1 - \delta.
$$

Suppose that events $E_1, \ldots, E_m$ all holds then,

$$
|\tilde{r}_i - \gamma| = \left| \sum_{k=1}^{m} \frac{n_{i,k}'}{n_i} - \sum_{k=1}^{m} \beta_{i,k} \right| \\
\leq \sum_{k=1}^{m} \left| \frac{n_{i,k}'}{n_i} - \beta_{i,k} \right| \\
\leq \epsilon \sum_{k=1}^{m} \max \{ \sqrt{\beta_{i,k}}, \epsilon \} \\
\leq \epsilon \sum_{k=1}^{m} \sqrt{\beta_{i,k}} + m \epsilon^2
$$

32
by applying Cauthy’s Inequality we have:

\[ |\hat{r}_i - \gamma_i| \leq \epsilon \sqrt{m \sum_{k=1}^{m} \beta_{i,k} + m \epsilon^2} = \epsilon \sqrt{m \gamma_i + m \epsilon^2} \]

The proof can be completed by recalling that the probability that events \( E_1, \ldots, E_m \) all holds is at least \( 1 - \delta \).

### 8.2 Proof of Theorem 7

When there is \( f_i \) fraction of random misbehaving users, the average rating \( \hat{r}_i \) of product \( P_i \) converges to \((1 - f_i)\gamma_i + f_i m/2 = \gamma_i + (m/2 - \gamma_i) f_i\). Since \( n_i \geq 3 \ln(2m/\delta)/\epsilon^2 \) holds, then with a similar derivation of Theorem 6, we could obtain that:

\[
\left| \sum_j r_{i,j}/n_i - (\gamma_i + m/2 - \gamma_i) f_i \right| \leq \epsilon \sqrt{m (\gamma_i + (m/2 - \gamma_i) f_i) + m \epsilon^2} \tag{15}
\]

holds with probability at least \( 1 - \delta \). Observe that

\[
\left| \frac{m}{2} - \gamma_i \right| f_i - \left| \frac{\sum_j r_{i,j}}{n_i} - (\gamma_i + \frac{m}{2} f_i - \gamma_i f_i) \right| \leq \left| \frac{\sum_j r_{i,j}}{n_i} - \gamma_i \right| \\
\leq \frac{m}{2} - \gamma_i \left| f_i + \left| \frac{\sum_j r_{i,j}}{n_i} - (\gamma_i + \frac{m}{2} f_i - \gamma_i f_i) \right| ,
\]

by applying Inequality (15) to this inequality we can finish the proof of this theorem. \[\square\]