Mathematical Modeling and Analysis of Product Rating with Partial Information

HONG XIE and JOHN C. S. LUI, The Chinese University of Hong Kong

Many Web services like Amazon, Epinions, and TripAdvisor provide historical product ratings so that users can evaluate the quality of products. Product ratings are important because 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 assigns 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 that one can make a reliable evaluation of its quality? How may users' misbehavior, such as cheating in product rating, affect the evaluation result? To answer these questions, we present a probabilistic model to capture various important factors (e.g., rating aggregation rules, rating behavior) that may influence the product quality assessment under the partial information setting. We derive the minimum number of ratings needed to produce a reliable indicator on the quality of a product. We extend our model to accommodate users' misbehavior in product rating. We derive the maximum fraction of misbehaving users that a rating aggregation rule can tolerate and the minimum number of ratings needed to compensate. We carry out experiments using both synthetic and real-world data (from Amazon and TripAdvisor). We not only validate our model but also show that the "average rating rule" produces more reliable and robust product quality assessments than the "majority rating rule" and the "median rating rule" in aggregating product ratings. Last, we perform experiments on two movie rating datasets (from Flixster and Netflix) to demonstrate how to apply our framework to improve the applications of recommender systems.

Categories and Subject Descriptors: H.3.5 [**Information Storage and Retrieval**]: Online Information Services

General Terms: Reliability, Human Factors, Experimentation

Additional Key Words and Phrases: Product rating, minimum number of ratings, misbehavior, bias, true quality, rating aggregation rule

ACM Reference Format:
Hong Xie and John C. S. Lui. 2015. Mathematical modeling and analysis of product rating with partial information. ACM Trans. Knowl. Discov. Data 9, 4, Article 26 (June 2015), 33 pages.
DOI: http://dx.doi.org/10.1145/2700386

1. INTRODUCTION

In many Web services, users can contribute their opinions or ideas in the form of ratings or reviews. For example, we see ratings or reviews in content-sharing Web sites (e.g., Flickr and YouTube), online recommendation systems (e.g., Amazon and MovieLens), product review Web sites (e.g., TripAdvisor and Epinions), and online e-commerce systems (e.g., eBay). With these ratings and product reviews, one can

The work of John C. S. Lui is supported in part by the HK GRF grant 415112.
Authors' addresses: H. Xie and J. C. S. Lui, Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, The People's Republic of China; emails: {hxie, cslui}@cse.cuhk.edu.hk.

DOI: http://dx.doi.org/10.1145/2700386

perform information search or make purchase decisions by taking advantage of the opinions of other users (also known as wisdom of the crowd). In addition, the service provider can infer users' preferences from these ratings to make better recommendations. Such Web-based online rating systems operate as follows. There are a number of items (e.g., products, pictures, videos). Each user provides ratings or reviews to a *subset of items*. The ratings of an item are available to the public.

Online rating systems are classified into two typical categories according to their application domains. The first category interprets ratings as public quality assessments on products [Lauw et al. 2008, 2012; Traupman and Wilensky 2006], where users act as *reviewers* and assess the quality of products in the form of ratings. For example, users assign ratings to reflect the quality of products or the reputation level of sellers in eBay. Other typical examples include Epinions, TripAdvisor, and Wikipedia. The main interest is in identifying the intrinsic quality of products from their historical ratings. The intrinsic quality of products can help a service provider make better promotions of products to increase the sales. In addition, users will be more likely to purchase a product that fulfills their expectation with the benefit of the intrinsic quality in comparing products, leading to more users being encouraged to participate (i.e., assign ratings or purchase products). The second category of online rating systems interprets ratings as users' preference information. For example, users express ratings to show their preferences or tastes in movies on IMDB. More examples include Flixster, Netflix, and RateBeer. Such systems are called *recommendation systems* [Adomavicius and Tuzhilin 2005; Resnick and Varian 1997], where personalized quality of a product is of essential importance instead of the intrinsic quality. The main objective of such systems is to make personalized recommendations [Herlocker et al. 2004; Resnick et al. 1994; Resnick and Varian 1997].

This article focuses on the first category of online rating systems—that is, ratings are interpreted as product quality assessments. Identifying the intrinsic quality of products is critical to the revenue of such online rating systems. However, we only have *partial information* on the product rating: each user only expresses ratings to a *small subset* of products. This partial information makes it challenging to assess the quality of products reliably. Hence, it is important to understand the *accuracy* and *effectiveness* of such online rating systems in assessing product quality. In particular, we seek to explore a number of fundamental questions: What is the minimum number of ratings a product needs to have a reliable reflection on its quality? How may users' misbehavior or inherent biases in rating affect the accuracy? The answers to these questions provide important building blocks to improve the applications of online rating systems. First, a service provider can design some efficient incentive mechanisms to incentivize users to contribute ratings to those products that have insufficient number of ratings (i.e., smaller than the minimum number of ratings) to improve the accuracy of product quality assessment. Second, a service provider can put more weights on those products that have a sufficient number of ratings (i.e., larger than the minimum number of ratings) to make more reliable product promotions. Third, a service provider can gain important insights in designing or deploying misbehavior detection algorithms. For example, if a very small fraction of misbehaving users can distort the product quality assessment, then a service provider needs to design (or deploy) a detection algorithm with a very high true positive value. However, if a small fraction of misbehaving users can be tolerated by compensating a small number of ratings, then a service provider may relax the restriction on the true positive value to attain better design trade-offs. Fourth, users can put more attention on those products that have a sufficient number of ratings to increase the possibility that they purchase a product fulfilling their expectation. Little attention has been paid to these fundamental questions. This work fills this void. Our contributions are as follows:

—We propose a probabilistic model to capture various important elements (e.g., rating aggregation rules, rating behavior) of an online rating system. Our model is general enough to represent both honest and misbehaving users in product rating.
—We derive the minimum number of ratings needed to produce a reliable product quality assessment under the honest rating and the misbehavior setting. We derive maximum fraction of misbehaving users that a rating aggregation rule can tolerate.
—We propose an inference algorithm to infer parameters of our model from *partial information*, say available ratings, to address the applicability of our framework.
—We perform experiments using both synthetic data and real-world data (from Amazon and TripAdvisor) to validate our model and examine various factors that may affect the accuracy of product quality assessment. We show that the *average score rule* is more robust and reliable than the *majority rule* and the *median rule* in evaluating product quality. We find that around 100 ratings are sufficient to reflect the true quality of an item on TripAdvisor (or Amazon).
—We perform experiments on two movie rating datasets (from Flixster and Netflix) to demonstrate how to apply our framework to recommender systems.

The rest of the article is organized as follows. In Section 2, we present the system model. In Section 3, we derive the minimum number of ratings needed under honest or misbehaving settings. In Section 4, we present an inference algorithm to infer model parameters. In Section 5, we present experimental results using synthetic data. In Section 6, we present experimental results using real-world datasets. In Section 7, we demonstrate how to apply our framework recommender systems. Related work is given in Section 8, and Section 9 concludes and discusses future directions.

2. SYSTEM MODEL

We consider an online rating system composed of a finite set of N products denoted by $P_1, \ldots, P_N$ and M users denoted by $U_1, \ldots, U_M$. Each user only expresses ratings to a *subset* of products on an m-level cardinal metric $\{1, \ldots, m\}$. For example, a three-level cardinal metric can be $\{1 = \text{“poor,”}\ 2 = \text{“good,”}\ 3 = \text{“excellent”}\}$. A larger rating implies higher quality. Users independently express ratings to products. Product P_i has $n_i \leq M$ ratings. Let $n_{i,k}$ denote the number of ratings for P_i that are of rating level k. Let $\mathbf{r}_i = \{r_{i,1}, \ldots, r_{i,M}\}$ denote a set of M ratings for P_i, where $r_{i,j} \in \{1, \ldots, m\}$ if U_j rates P_i; otherwise, $r_{i,j} = 0$. We emphasize that $r_{i,j} = 0$ implies a missing rating. We treat the available ratings as *partial information*. One most important application of product rating is in assessing the quality of products. This work aims to explore how various factors can influence the assessment accuracy.

2.1. Rating Aggregation Rules

The quality of a product is assessed by performing a rating aggregation on its historical ratings. We consider the following three commonly used rating aggregating rules.

—*Majority rule (MR)*: MR assesses the quality of a product via the *majority* of its historical ratings. Let $\widehat{\ell}_i$ denote the *evaluated label* of P_i produced by performing the *majority rule* on its historical ratings. Formally, we have $\widehat{\ell}_i = \arg\max_k\{n_{i,k}\}$. Let $\ell_i \in \{1, \ldots, m\}$ represent the label that reflects the true quality of P_i under the majority rule. We emphasize that users do not have any prior knowledge on ℓ_i. How many ratings do we need to have a strong guarantee that $\widehat{\ell}_i$ reveals the true label ℓ_i?
—*Median rule (MDR)*: MDR assesses the quality of a product via the *median* of its historical ratings. Let $\widehat{l}_i$ denote the median rating of P_i produced by performing the *median rule* on its historical ratings. Formally, we have $\widehat{l}_i = \arg\min_k\{|\{r_{i,j}|1 \leq r_{i,j} \leq k\}|/n_i > \frac{1}{2}\}$. Let $l_i \in \{1, \ldots, m\}$ represent the median rating that reflects the true quality

of P_i under the median rule. Users do not have any prior knowledge on l_i. How many ratings do we need to have a strong guarantee that $\widehat{l_i}$ reveals l_i?

—*Average score rule (ASR)*: ASR assesses the quality of a product the *average* of its historical ratings. Let $\widehat{\gamma_i} = \sum_j r_{i,j}/n_i$ denote the average rating of P_i produced by performing the *average score rule* on its historical ratings. Let $\gamma_i \in [1, m]$ denote the average rating that reflects the true quality of P_i under the average score rule. Users do not have any prior knowledge on γ_i. How many ratings do we need so that $\widehat{\gamma_i}$ accurately reflects γ_i?

Remark. First, thevariables ℓ_i, l_i, γ_i are treated as latent variables. Second, in practice, ℓ_i, l_i, γ_i may not be the same, because they represent three different ways in defining the true quality of products—that is, by opinions from the majority, by opinions that stand in the median, or by averaging the whole opinions.

2.2. Model for Rating Behavior

A user needs to assess the quality of a product to express a rating. We consider two of the most important factors that affect this assessment: a user's (1) intrinsic quality of products and (2) expertise level (or knowledge level).

We describe the rating behavior via a random variable that one can vary its mean or variance to reflect the preceding factors. A large mean implies a product having a high intrinsic quality, and a small variance implies that a user has a high expertise level in assessing the quality of a product. More concretely, we describe the rating behavior that results in $r_{i,j}$ using the following probability mass function (pmf) of $r_{i,j}$:

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

where $\rho_{ij,k} \geq 0$ and $\sum_{k=1}^{m} \rho_{ij,k} = 1$. We emphasize that different users may have different rating distributions over the same product. The collective rating behavior of the whole user population over product P_i reflects the public opinion on P_i. To model it, let $(\theta_1, \ldots, \theta_m)$ denote one instance of the probability distribution for a rating, where $0 \leq \theta_i \leq 1$ and $\sum_{i=1}^{m} \theta_i = 1$. We express the space over all possible probability distributions for a rating as $\mathcal{S} = \{(\theta_1, \ldots, \theta_m) | \sum_{i=1}^{m} \theta_i = 1, \theta_i \geq 0, \forall i\}$. We assume that there is an underlying distribution, say $\mathcal{D}(P_i)$, over the space $\mathcal{S}$ that defines the collective rating behavior to product P_i. In this study, $\mathcal{D}(P_i)$ is a Dirichlet distribution $Dirichlet(\boldsymbol{\alpha}_i)$ with density function

$$p(\theta_1, \ldots, \theta_m) = \frac{\Gamma(\sum_{k=1}^{m} \alpha_{i,k})}{\prod_{k=1}^{m} \Gamma(\alpha_{i,k})} \prod_{k=1}^{m} \theta_k^{\alpha_{i,k}-1}, \tag{2}$$

where $\boldsymbol{\alpha}_i = (\alpha_{i,1}, \ldots, \alpha_{i,m})$, $\sum_{k=1}^{m} \alpha_{i,k} = 1$, and $\alpha_{i,k} > 0, \forall k$. Note that we require $\sum_{k=1}^{m} \alpha_{i,k} = 1$ mainly for the purpose of simplifying notations. It does not lose any generality, because one will see later that our results canincorporate the case $\sum_{k=1}^{m} \alpha_{i,k} \neq 1$ by substituting $\alpha_{i,1}, \ldots, \alpha_{i,m}$ with $\frac{\alpha_{i,1}}{\sum_{k=1}^{m} \alpha_{i,k}}, \ldots, \frac{\alpha_{i,m}}{\sum_{k=1}^{m} \alpha_{i,k}}$, respectively.

Remark. We treat each observed rating as a random sample produced by the generative process that generates a user by drawing a multinomial distribution $\boldsymbol{\rho}_{ij} = (\rho_{ij,1}, \ldots, \rho_{ij,m})$ from $Dir(\boldsymbol{\alpha}_i)$ and generates a rating from this multinomial distribution. We have two reasons for choosing the *Dirichlet* distribution. First, $Dir(\boldsymbol{\alpha}_i)$ contains m parameters, and by varying them we can model any collective rating behavior of the whole user population to P_i. Second, $Dir(\boldsymbol{\alpha}_i)$ is a quite natural and reasonable distribution over the space $\mathcal{S}$ [Bishop 2006].

2.3. Model for Misbehavior

Users' misbehavior in ratings widely exist in many Web services, such as TripAdvisor, Amazon, and eBay. Misbehaving users in online rating systems have been reported extensively over the past decade. For example, it has been reported [ABC7 News 2012] that some users are paid to post five-star Google feedbacks. In addition, users of Amazon are paid to promote (five-star rating) or pan (one-star rating) a product [The New York Times 2004]. More reports on misbehaving users can be found in other sources [BBC News 2013; TheWallStreetJournal 2009]. Furthermore, some companies, such as PayPerPost [2015], provide professional fake review (rating)-writing services. Their services include promoting or badmouthing a product, or even assigning random ratings to attack an online rating system. Jindal and Liu [2007] published a dataset from Amazon, and they detected reviews (or ratings) assigned by misbehaving users. Their results indicate the existence of misbehavior in real data. To explore their impact, we consider the following typical cases of misbehavior:

—*Random misbehavior:* A random misbehavior implies that a user assigns a random rating to a product. Random misbehavior may arise in the following scenarios. A product rating Web site may launch some rating robots or hire some users to attack a competitor's Web site by assigning random ratings to distort this competitor's system. On the other hand, many product rating Web sites deploy incentive mechanisms that reward users' rating in the form of income share (e.g., Epinion), badges (e.g., TripAdvisor), and so forth. In such systems, it may happen that some users are impatient or do not want to spend a long time to evaluate the quality of an item, resulting in assignment of random ratings to earn the rewards. To illustrate, suppose that $r_{i,j}$ is assigned by a random misbehaving user. Formally, we have

$$\Pr[r_{i,j} = k|\boldsymbol{\rho}_{ij}] = \frac{1}{m}, \quad \forall k = 1, \ldots, m,$$

where $\boldsymbol{\rho}_{ij} = (\rho_{ij,1}, \ldots, \rho_{ij,m})$ is the true pmf of $r_{i,j}$. Let $f_r \in [0, 1]$ denote the fraction of random misbehaving users.

—*Biased misbehavior.* A biased misbehavior implies that a user is biased toward one particular rating. For example, 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. To illustrate, consider a biased misbehaving user U_j expressing a rating $r_{i,j}$ to P_i. Let $\ell'_j \in \{1, \ldots, m\}$ denote the rating toward which U_j is biased. Formally, we can model the rating under biased misbehavior as

$$\Pr[r_{i,j} = k|\boldsymbol{\rho}_{ij}] = \begin{cases} 1, & \text{if } k = \ell'_j, \\ 0, & \text{otherwise,} \end{cases}$$

where $\boldsymbol{\rho}_{ij} = (\rho_{ij,1}, \ldots, \rho_{ij,m})$ is the true pmf of $r_{i,j}$. Let $f_b \in [0, 1]$ represent the fraction of biased misbehaving users.

2.4. Model for Inherent User Biases

We use a probabilistic model to capture inherent user biases in rating. In many real-world scenarios, users' ratings may not accurately reflect their experiences on products, but rather they might be biased due to inherent user biases—that is, some critical users may express lower ratings, and some lenient users may express higher ratings. Based on inherent user biases, we categorize users into three types: *critical* means assigning lower ratings, *lenient* means assigning higher ratings, and *neutral* means assigning ratings accurately reflects users' experiences. Recall that in Section 2.2, we model the collective rating behavior of neutral user via the Dirichlet distribution

$Dir(\boldsymbol{\alpha}_i)$. We extend it to model the collective rating behavior of critical users via the Dirichlet $Dir(\boldsymbol{\alpha}'_i)$. One can vary the value of $\boldsymbol{\alpha}'_i$ to reflect the critical degree. For example, $\boldsymbol{\alpha}'_i \approx (1, 0, \ldots, 0)$ implies that critical users are most likely to assign the lowest rating of 1. Similarly, we model the collective rating behavior of lenient users via the Dirichlet distribution $Dir(\boldsymbol{\alpha}^*_i)$. One can vary the value of $\boldsymbol{\alpha}^*_i$ to reflect the degree of leniency. For example, $\boldsymbol{\alpha}^*_i \approx (0, \ldots, 0, 1)$ implies that lenient users are most likely to assign the highest rating of m.

3. THEORETICAL ANALYSIS

We derive the minimum number of ratings a given product needs so that its aggregate rating is statistically accurate to reflect its true quality under the honest rating and the misbehavior setting. We also derive the maximum fraction of misbehaving user that a rating aggregation rule can tolerate.

3.1. Majority Rule

We derive tight lower bounds on the number of ratings needed so that $\widehat{\ell}_i$ reveals the true label ℓ_i with high probability (i.e., $\Pr[\widehat{\ell}_i = \ell_i] \approx 1$). We also quantify the impact of misbehavior and inherent user biases on these lower bounds.

Analysis for honest rating. Let us begin our exploration by a simple case that all users rate products honestly and are neutral. In our analysis, we assume that the model parameters $\boldsymbol{\alpha}_i, \forall i$ are given. In Section 4, we will show how to infer these parameters from historical ratings. For the ease of presentation, let $\mathbf{r}^+_i = \{r^+_{i,1}, \ldots, r^+_{i,n_i}\}$ denote a set of all positive (or observed) ratings of P_i. We state the pmf of $r^+_{i,j}$ in the following lemma.

LEMMA 3.1. *The pmf of the rating $r^+_{i,j}$ can be expressed as $\Pr[r^+_{i,j} = k] = \alpha_{i,k}$, for all $k = 1, \ldots, m$, where $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$.*

PROOF. Please refer to the Appendix for the derivation. □

Remark. Collectively, P_i receives a rating k with probability $\alpha_{i,k}$. It implies that $\ell_i = \arg\max_k\{\alpha_{i,k}\}$ and $\widehat{\ell}_i$ converges to the true label ℓ_i as the number of ratings n_i goes to infinity. We next derive a practical lower bound on the number of ratings needed to guarantee $\widehat{\ell}_i = \ell_i$ with high probability.

Definition 3.2. Let $\widetilde{\alpha}_i \triangleq \max\{\alpha_{i,k} | k \neq \ell_i\}$ denote the second largest value among $\alpha_{i,1}, \ldots, \alpha_{i,m}$.

Definition 3.3. Let n'_i denote the minimum number of ratings needed to guarantee that the aggregate rating of P_i reflects its true quality with high confidence.

THEOREM 3.4 (HONEST RATING). *Suppose that users rate honestly and are neutral. If*

$$n_i \geq n'_i = (2(\alpha_{i,\ell_i} + \widetilde{\alpha}_i)(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^{-2} - 2 + 4(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^{-1}/3) \ln(m-1)\delta^{-1}, \tag{3}$$

then $\Pr[\widehat{\ell}_i = \ell_i] \geq 1 - \delta$.

PROOF. Please refer to the Appendix for the derivation. □

Remark. To increase the confidence $1 - \delta$ in revealing the true label, we need to increase the minimum number of ratings n'_i in a logarithmic rate. Observe that n'_i is proportional to $1/(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^2$. This implies that the smoother the model parameter $\boldsymbol{\alpha}_i$ curve, the larger the minimum number of ratings.

Table I presents some numerical examples on the minimum number of ratings. It depicts the level of rating metric m, the model parameter $\boldsymbol{\alpha}_i$, the success probability $1-\delta$, and the minimum number of ratings n'_i. An increase in $1-\delta$ from 0.7 to 0.9 results

Table I. Minimum Number of Ratings (MR, Honest Rating)

m	$\boldsymbol{\alpha}_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$	$1-\delta$	n_i'
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.7	57
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	66
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.9	81

Table II. Minimum Number of Ratings to Tolerate Random Misbehavior (MR)

m	$\boldsymbol{\alpha}_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$	$1-\delta$	f_r	$\overline{f}_r$	n_i'
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0	1	66
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.05	1	71
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.1	1	78

in an increase in the minimum number of ratings from 57 to 81. We next show that the bound derived in Theorem 3.4 is tight.

THEOREM 3.5 (TIGHTNESS OF BOUND). *Suppose that all users rate honestly and are neutral. Assume that* $\widetilde{\alpha}_i \geq 100\alpha_{i,\ell_i}/101$. *There exists a positive constant* η *such that for any* $\delta \leq \eta$, *if* $n_i = O((\alpha_{i,\ell_i} + \widetilde{\alpha}_i)(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^{-2} \ln \delta^{-1})$ *ratings, then* $\Pr[\widehat{\ell}_i \neq \ell_i] \geq \Omega(\delta)$.

PROOF. Please refer to the Appendix for the derivation. □

Remark. It implies that the lower bound derived in Theorem 3.4 is asymptotically tight, because it is asymptotically equal to $2(\alpha_{i,\ell_i} + \widetilde{\alpha}_i)(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^{-2} \ln \delta^{-1}$.

Analysis of rating under misbehavior. We now explore the impact of misbehavior in ratings. We first explore the impact of random misbehavior. We quantify its impact on the minimum number of ratings in the following theorem.

Definition 3.6. Let $\overline{f}_r$ denote the maximum fraction of random misbehaving users that a rating aggregation rule can tolerate.

THEOREM 3.7 (RANDOM MISBEHAVIOR). *Suppose that the fraction of random misbehaving users satisfies* $f_r < \overline{f}_r = 1$ *and other users rate honestly and are neutral. If* n_i *satisfies* $n_i \geq n_i' = ((4f_r/m + 2(1-f_r)(\alpha_{i,\ell_i} + \widetilde{\alpha}_i))(1-f_r)^{-2}(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^{-2} + \frac{4}{3}(1-f_r)^{-1}(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^{-1} - 2) \ln \frac{m-1}{\delta}$, *then* $\Pr[\widehat{\ell}_i = \ell_i] \geq 1-\delta$.

PROOF. Please refer to the Appendix for the derivation. □

Remark. If the fraction of random misbehaving users is less than 1, the majority rule can always tolerate it by compensating a large enough number of ratings. This implies that random misbehavior does not distort the product quality assessment.

Table II presents some numerical results on the minimum number of ratings. It depicts m, $\boldsymbol{\alpha}_i$, $1-\delta$, n_i', the fraction of random misbehaving users f_r, and $\overline{f}_r$, respectively. An increase in f_r from 0 to 0.1 results in an increase in n_i' from 66 to 78.

Discussion. In our analysis, f_r is set to be a pregiven value; however, in practice, we may not have any a priori knowledge on f_r. We will see later that the fraction of biased misbehaving users is set to be a pregiven value as well. We state their reasonability as follows. We investigate whether a rating aggregation rule can tolerate a small fraction of misbehaving users and how many ratings we need to compensate to tolerate this misbehavior. The answers to these questions may give us important insights on detect or defend misbehavior. For example, if a very small fraction of misbehaving users can distort the product quality assessment, then a service provider needs to design (or deploy) a detection algorithm with a very high true positive value. However, if a small

Table III. Minimum Number of Ratings to Tolerate Biased Misbehavior (MR)

m	$\boldsymbol{\alpha}_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$	$1-\delta$	ℓ'	$\overline{f}_b$	f_b	n_i'
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	5	0.211	0	66
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	5	0.211	0.05	112
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	5	0.211	0.1	237

fraction of misbehaving users can be tolerated by compensating a small number of ratings, then a service provider may relax the restriction on the true positive value to attain better design trade-offs. Furthermore, we derive analytical expressions for the maximum fraction of misbehaving users that a rating aggregation rule can tolerate. As illustrated in Section 6.4, we apply this result to analyze real-world data. Through the analysis result, we obtain important insights on the fraction of misbehaving users that a rating aggregation rule can tolerate in practice.

We now explore the impact of biased misbehavior. In the following analysis, we assume that all biased misbehaving users bias toward the same rating ℓ' (i.e., $\ell'_j = \ell', \forall j$).

Definition 3.8. Let $\overline{f}_b$ denote the maximum fraction of biased misbehaving users that a rating aggregation rule can tolerate.

THEOREM 3.9 (BIASED MISBEHAVIOR). *Suppose that biased misbehaving users bias against the ground-truth label—that is, $\ell' \neq \ell_i$. Assume that the fraction of biased misbehaving users satisfies*

$$f_b < \overline{f}_b = (\alpha_{i,\ell_i} - \alpha_{i,\ell'})/(1 + \alpha_{i,\ell_i} - \alpha_{i,\ell'}), \tag{4}$$

and other users rate honestly and are neutral. If n_i satisfies $n_i \geq n_i' = (2(\alpha_{i,\ell_i} + \max\{f_b/(1-f_b)+\alpha_{i,\ell'}, \widetilde{\alpha}_i\})(1-f_b)^{-1}(\alpha_{i,\ell_i} - \max\{f_b/(1-f_b)+\alpha_{i,\ell'}, \widetilde{\alpha}_i\})^{-2} + \frac{4}{3}(1-f_b)^{-1}(\alpha_{i,\ell_i} - \max\{f_b/(1-f_b)+\alpha_{i,\ell'}, \widetilde{\alpha}_i\})^{-1} - 2)\ln\frac{m-1}{\delta}$, then $\Pr[\widehat{\ell}_i = \ell_i] \geq 1-\delta$. If f_b does not satisfy Inequality (4), then it is impossible to extract the true label with high probability no matter how many ratings we have.

PROOF. Please refer to the Appendix for the derivation. □

Remark. The upper bound derived in Inequality (4) is proportional to $\alpha_{i,\ell_i} - \alpha_{i,\ell'}$. It implies that the smoother the model parameter $\boldsymbol{\alpha}_i$ curve, the smaller the fraction of biased misbehaving users that the majority rule can tolerate.

Table III presents some numerical examples on $\overline{f}_b$ and the minimum number of ratings needed to tolerate biased misbehavior. One can observe that the majority rule can tolerate a fraction of at most 0.211 biased misbehaving users. An increase in the fraction of biased misbehaving users from 0 to 0.1 results in an increase in the minimum number of ratings from 66 to 237—a significant increase. Tolerating biased misbehavior requires more ratings than random misbehavior (please refer to Table II).

Analysis of rating under inherent user biases. Now we explore the impact of inherent user biases. One can observe that the preceding analysis of misbehavior can easily be extended to quantify the impact of inherent user biases. We omit the analysis results here to avoid redundancies and save some spaces.

3.2. Median Rule

We now analyze the median rule. We follow the same analysis flow as the majority rule.

Analysis for honest rating. We assume that all users rate products honestly and are neutral. Applying Lemma 3.1, we have $l_i = \arg\min_k\{\sum_{\kappa=1}^{k} \alpha_{i,\kappa} > \frac{1}{2}\}$, and $\widehat{l}_i$ converges to

Table IV. Minimum Number of Ratings (MDR, Honest Rating)

m	$\alpha_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$	$1-\delta$	n_i'
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.7	38
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	47
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.9	60

l_i as the number of ratings n_i goes to infinity. We next derive a practical lower bound on the number of ratings needed to guarantee $\widehat{l_i} = l_i$ with high probability.

Definition 3.10. Let $F_i(k) \triangleq \Pr[r_{i,j}^+ \le k] = \sum_{k'=1}^{k} \alpha_{i,k'}$, where $k = 1, \ldots, m$, denote the probability that product P_i receives a rating smaller or equal to k.

THEOREM 3.11 (HONEST RATING). *Suppose that all users rate honestly and are neutral. If n_i satisfies $n_i \ge n_i' = (3(\alpha_{i,l_i} - |2F_i(l_i) - \alpha_{i,l_i} - 1|)^{-2} - 3 + 2(\alpha_{i,l_i} - |2F_i(l_i) - \alpha_{i,l_i} - 1|)^{-1})\frac{2}{3}\ln\frac{2}{\delta}$, then $\Pr[\widehat{l_i} = l_i] \ge 1 - \delta$.*

PROOF. Please refer to the Appendix for the derivation. □

Remark. To increase the confidence $1 - \delta$ in revealing the true label l_i, we need to increase the minimum number of ratings n_i' in a logarithmic rate. The minimum number of ratings is proportional to $(\alpha_{i,l_i} - |2F_i(l_i) - \alpha_{i,l_i} - 1|)^{-2} = (\min\{F_i(l_i) - 0.5, 0.5 - F_i(l_i - 1)\})^{-2}$. This implies that the smoother the model parameter $\boldsymbol{\alpha}_i$ curve, the larger the minimum number of ratings n_i'. The smoothness of the model parameter $\boldsymbol{\alpha}_i$ curve is the key factor in influencing n_i'.

Table IV presents some numerical results on the minimum number of ratings. An increase in the success probability $(1-\delta)$ from 0.7 to 0.9 results in an increase in the minimum number of ratings n_i' from 38 to 60. We next show this bound is tight.

THEOREM 3.12 (TIGHTNESS OF BOUND). *Suppose that all users rate honestly and are neutral. There exists three positive constants η_1, $\eta_2 > 0.5$ and $\eta_3 < 0.5$ such that for any $\delta \in [0, \eta_1]$, $F_i(l_i) \in [0.5, \eta_2]$, and $F_i(l_i - 1) \in [\eta_3, 0.5]$. If $n_i = O((\alpha_{i,l_i} - |2F_i(l_i) - \alpha_{i,l_i} - 1|)^{-2} \ln \delta^{-1})$, then $\Pr[\widehat{l_i} \ne l_i] \ge \Omega(\delta)$.*

PROOF. Please refer to the Appendix for the derivation. □

Remark. It states that the lower bound derived in Theorem 3.11 is asymptotically tight because it is asymptotically equal to $2(\alpha_{i,l_i} - |2F_i(l_i) - \alpha_{i,l_i} - 1|)^{-2} \ln \delta^{-1}$.

Analysis of rating under misbehavior. We first explore the impact of random misbehavior on the minimum number of ratings. The following theorem derives the maximum fraction of random misbehavior that the median rule can tolerate and the minimum number of ratings needed to compensate in tolerating a given fraction of random misbehaving users.

THEOREM 3.13 (RANDOM MISBEHAVIOR). *Suppose that the fraction of random misbehaving users satisfies*

$$f_r \le \overline{f}_r = \min\left\{\left(\frac{F_i(l_i) - 0.5}{F_i(l_i) - l_i/m}\right)^{\mathbf{I}_{i1}}, \left(\frac{0.5 - F_i(l_i - 1)}{(l_i - 1)/m - F_i(l_i - 1)}\right)^{\mathbf{I}_{i2}}\right\}, \tag{5}$$

and other users rate honestly and are neutral, where $\mathbf{I}_{i1} = 1$ if $l_i < mF_i(l_i)$, otherwise $\mathbf{I}_{i1} = 0$, and $\mathbf{I}_{i2} = 1$ if $l_i > mF_i(l_i - 1) + 1$, otherwise $\mathbf{I}_{i2} = 0$. If n_i satisfies $n_i \ge n_i' = (3(f_r/m + (1 - f_r)\alpha_{i,l_i} - |(2l_i - 1)f_r/m + (1 - f_r)(2F_i(l_i) - \alpha_{i,l_i}) - 1|)^{-2} - 3 + 2(f_r/m + (1 - f_r)\alpha_{i,l_i} - |(2l_i - 1)f_r/m + (1 - f_r)(2F_i(l_i) - \alpha_{i,l_i}) - 1|)^{-1})\frac{2}{3}\ln\frac{2}{\delta}$, then $\Pr[\widehat{l_i} = l_i] \ge 1 - \delta$. If f_r

Table V. Minimum Number of Ratings to Tolerate Random Misbehavior (MDR)

m	$\alpha_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$	$1-\delta$	$\overline{f}_r$	f_r	n_i'
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.625	0	47
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.625	0.05	55
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.625	0.1	66

Table VI. Minimum Number of Ratings to Tolerate Biased Misbehavior (MDR)

m	$\alpha_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$	$1-\delta$	ℓ'	$\overline{f}_b$	f_b	n_i'
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	1	0.25	0	47
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	1	0.25	0.05	72
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	1	0.25	0.1	126

does not satisfy Inequality (5), then it is impossible to extract the true label l_i with high probability no matter how many ratings we have.

PROOF. This proof is similar to that of Theorem 3.7. □

Remark. The upper bound derived in Inequality (5) is proportional to $\min\{F_i(l_i) - 0.5, 0.5 - F_i(l_i - 1)\}$. It implies that the smoother the model parameter $\boldsymbol{\alpha}_i$ curve, the smaller the fraction of random misbehaving users that the median rule can tolerate.

Table V presents some numerical results on $\overline{f}_r$ and n_i'. When $\boldsymbol{\alpha}_i = (2/30, 3/30, 5/30, 14/30, 6/30)$, the median rule can tolerate a maximum fraction of $\overline{f}_r = 0.625$ random misbehaving users. An increase in f_r from 0 to 0.1 results in an increase in n_i' from 47 to 66—a slight increase.

We next derive the maximum fraction of biased misbehavior that the median rule can tolerate and the minimum number of ratings needed to compensate in tolerating a given fraction of biased misbehaving users.

THEOREM 3.14 (BIASED MISBEHAVIOR). *Suppose that biased misbehaving users bias against the ground-truth label—that is, $\ell' \neq l_i$. Assume that the fraction of biased misbehaving users satisfies*

$$f_b \leq \overline{f}_b = \begin{cases} 1 - 1/(2F_i(l_i)), & \ell' > l_i \\ 1 - 1/(2 - 2F_i(l_i - 1)), & \ell' < l_i \end{cases}, \tag{6}$$

and other users rate honestly and are neutral. If n_i satisfies $n_i \geq n_i' = (3((1 - f_b)\alpha_{i,l_i} - |2f_b\mathbf{I}_{\{\ell' < l_i\}} + (1 - f_b)(2F_i(l_i) - \alpha_{i,l_i}) - 1|)^{-2} - 3 + 2((1 - f_b)\alpha_{i,l_i} - |2f_b\mathbf{I}_{\{\ell' < l_i\}} + (1 - f_b)(2F_i(l_i) - \alpha_{i,l_i}) - 1|)^{-1})\frac{2}{3}\ln\frac{2}{\delta}$, then $\Pr[\widehat{l_i} = l_i] \geq 1 - \delta$. If f_b does not satisfy Inequality (6), then it is impossible to extract the true label l_i with high probability no matter how many ratings we have.

PROOF. This proof is similar to that of Theorem 3.7. □

Remark. The upper bound derived in Inequality (6) is proportional to $\min\{F_i(l_i) - 0.5, 0.5 - F_i(l_i - 1)\}$. It implies that the smoother the model parameter $\boldsymbol{\alpha}_i$ curve, the smaller the fraction of biased misbehaving users that the median rule can tolerate.

Table VI presents some numerical examples on $\overline{f}_b$ and n_i'. When $\ell' = 1$ and $\boldsymbol{\alpha}_i = (2/30, 3/30, 5/30, 14/30, 6/30)$, the median rule can tolerate a maximum fraction of $\overline{f}_b = 0.25$ biased misbehaving users. An increase in f_b from 0 to 0.1 results in an increase in n_i' from 47 to 126—a significant increase.

Table VII. Minimum Number of Ratings (ASR, Honest Rating)

m	$\boldsymbol{\alpha}_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$	$1-\delta$	E_r	n_i'
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.7	0.5	32
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.5	39
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.9	0.5	50

3.3. Average Score Rule

We now analyze the average score rule following the same flow with majority rule.

Analysis for honest rating. We assume that all users rate products honestly and are neutral. Applying Lemma 3.1, we have $\gamma_i = \sum_k k\alpha_{i,k}$, and $\widehat{\gamma}_i$ converges to γ_i as the number of ratings n_i goes to infinity. We next derive a practical lower bound on the number of ratings needed so that $\widehat{\gamma}_i$ accurately reflects γ_i with high probability.

Definition 3.15. Let E_r denote the maximum acceptable estimation error between $\widehat{\gamma}_i$ and γ_i. In other words, we accept $\widehat{\gamma}_i$ only when $|\widehat{\gamma}_i - \gamma_i| \leq E_r$.

THEOREM 3.16 (HONEST RATING). *Suppose that all users rate honestly and are neutral. If the number of ratings n_i satisfies $n_i \geq n_i' = 2((\sum_k k^2\alpha_{i,k} - \gamma_i^2)E_r^{-2} + \frac{m}{3E_r})\ln\frac{2}{\delta}$, then* $\Pr[|\widehat{\gamma}_i - \gamma_i| \leq E_r] \geq 1-\delta$.

PROOF. Please refer to the Appendix for the derivation. □

Remark. With a large enough number of ratings, $\widehat{\gamma}_i$ can reflect γ_i with an arbitrarily small error. The minimum number of ratings n_i' is approximately equal to $2(\sum_k k^2\alpha_{i,k} - \gamma_i^2)E_r^{-2}\ln 2\delta^{-1}$. It implies that the estimation error E_r and the smoothness of the model parameter $\boldsymbol{\alpha}_i$ curve are key factors that influence the minimum number of ratings.

Table VII presents some numerical results on the minimum number of ratings n_i'. It depicts m, $\boldsymbol{\alpha}_i$, $1-\delta$, n_i' and E_r, respectively. An increase in the success probability $1-\delta$ from 0.7 to 0.9 results in an increase in the minimum number of ratings n_i' from 32 to 50. It is interesting to observe that using the average score rule, we need 39 ratings ($1-\delta = 0.8$), whereas one needs 66 ratings using the majority rule (please refer to Table I) and needs 47 ratings using the median rule (please refer to Table IV). Namely, the average score rule requires fewer ratings than the majority rule and the median rule. We next show that the bound derived in Theorem 3.16 is tight.

THEOREM 3.17 (TIGHTNESS OF BOUND). *Suppose that all users rate honestly and are neutral. Assume that $\sum_k k^2\alpha_{i,k} - \gamma_i^2 \geq m^2/100$. There exist two positive constants η_1, η_2 such that for any $E_r \in [0, \eta_1]$ and any $\delta \in [0, \eta_2]$, if $n_i = O((\sum_k k^2\alpha_{i,k} - \gamma_i^2)E_r^{-2}\ln\delta^{-1})$, then* $\Pr[|\widehat{\gamma}_i - \gamma_i| \geq E_r] \geq \Omega(\delta)$.

PROOF. Please refer to the Appendix for the derivation. □

Remark. It implies that the lower bound derived in Theorem 3.16 is asymptotically tight because it is asymptotically equal to $2(\sum_k k^2\alpha_{i,k} - \gamma_i^2)E_r^{-2}\ln 2\delta^{-1}$.

Analysis of rating under misbehavior. We first explore the impact of random misbehavior on the minimum number of ratings. The following theorem derives the maximum fraction of random misbehavior that the average score rule can tolerate and the minimum number of ratings needed to compensate in tolerating a given fraction of random misbehaving users.

THEOREM 3.18 (RANDOM MISBEHAVIOR). *Suppose that the fraction of random misbehaving users satisfies*

$$f_r < \overline{f}_r = E_r/|\gamma_i - (m+1)/2|, \tag{7}$$

Table VIII. Minimum Number of Ratings to Tolerate Random Misbehavior (ASR)

m	$\boldsymbol{\alpha}_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$	$1-\delta$	E_r	f_r	$\overline{f}_r$	n_i'
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.5	0	0.789	39
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.5	0.05	0.789	44
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.5	0.1	0.789	51

Table IX. Minimum Number of Ratings to Tolerate Biased Misbehavior (ASR)

m	$\boldsymbol{\alpha}_i = (\alpha_{i,1}, \alpha_{i,2}, \alpha_{i,3}, \alpha_{i,4}, \alpha_{i,5})$	$1-\delta$	E_r	ℓ'	$\overline{f}_b$	f_b	n_i'
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.5	5	0.366	0	39
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.5	5	0.366	0.05	49
5	(2/30, 3/30, 5/30, 14/30, 6/30)	0.8	0.5	5	0.366	0.1	66

and other users rate honestly and are neutral. If n_i satisfies $n_i \geq n_i' = (\sum_k (k - \frac{m+1}{2} f_r - (1-f_r)\gamma_i)^2 (f_r/m + (1-f_r)\alpha_{i,k})(E_r - |\gamma_i - \frac{m+1}{2}| f_r)^{-2} + \frac{1}{3} m/(E_r - |\gamma_i - \frac{m+1}{2}| f_r)) 2 \ln \frac{2}{\delta}$, then $\Pr[|\widehat{\gamma}_i - \gamma_i| \leq E_r] \geq 1 - \delta$. If f_r does not satisfy Inequality (7), then it is impossible to accurately reflect γ_i with high probability no matter how many ratings we have.

PROOF. This proof is similar to that of Theorem 3.7. □

Remark. An increase in the maximum acceptable estimation error E_r leads to an increase in the maximum fraction of random misbehaving users that the average score rule can tolerate. As γ_i gets close to $(m+1)/2$, the fraction $\overline{f}_r$ becomes large. This is because the random misbehavior shifts the mean of ratings toward $(m+1)/2$.

Table VIII presents some numerical results on $\overline{f}_r$ and n_i'. When $\boldsymbol{\alpha}_i = (2/30, 3/30, 5/30, 14/30, 6/30)$, the average rule can tolerate a maximum fraction of $\overline{f}_r = 0.789$ random misbehaving users. An increase in f_r from 0 to 0.1 results in an increase in n_i' from 39 to 51—a slight increase.

We next derive the maximum fraction of biased misbehavior that the average score rule can tolerate and the minimum number of ratings needed to compensate in tolerating a given fraction of biased misbehaving users.

THEOREM 3.19 (BIASED MISBEHAVIOR). *Suppose that biased misbehaving users bias toward ℓ'. Assume that the fraction of biased misbehaving users f_b satisfies*

$$f_b < \overline{f}_b = E_r / |\gamma_i - \ell'|, \tag{8}$$

and other users rate honestly and are neutral. If n_i satisfies $n_i \geq n_i' = (\frac{1}{3} m/(E_r - |\gamma_i - \ell'| f_b) + \sum_k (k - \ell' f_b - (1-f_b)\gamma_i)^2 (f_b \mathbf{I}_{\{k=\ell'\}} + (1-f_b)\alpha_{i,k})(E_r - |\gamma_i - \ell'| f_b)^{-2}) 2 \ln \frac{2}{\delta}$, then $\Pr[|\widehat{\gamma}_i - \gamma_i| \leq E_r] \geq 1 - \delta$. If f_b does not satisfy Inequality (8), then it is impossible to accurately reflect γ_i with high probability no matter how many ratings we have.

PROOF. This proof is similar to that of Theorem 3.7. □

Remark. An increase in the maximum acceptable estimation error E_r leads to an increase in the maximum fraction of biased misbehaving users (i.e., $\overline{f}_b$) that the average score rule can tolerate. As the true quality γ_i gets close to ℓ', the fraction $\overline{f}_b$ becomes large. This is because the biased misbehavior shifts the mean of ratings toward ℓ'.

Table IX presents some numerical examples on $\overline{f}_b$ and n_i'. When $\boldsymbol{\alpha}_i = (2/30, 3/30, 5/30, 14/30, 6/30)$, the average rule can tolerate a maximum fraction of $\overline{f}_b = 0.366$ biased misbehaving users. An increase in f_b from 0 to 0.1 results in an increase in n_i' from 39 to 66—a slight increase.

3.4. Extensions to More Than One Type of Misbehavior

Our analysis thus far allows at most one type of misbehaving users. For completeness, we generalize our analysis to incorporate additional types.

We first explore a competition among biased misbehaving users—that is, a company may hire some users to promote its products (promoters), whereas its competitors may hire some users to bad-mouth its products (bad-mouthers). Promoters assign the highest rating m, and f_m denotes the fraction of them. Bad-mouthers assign the lowest rating of 1, and f_1 denotes the fraction of them. Which of them will win out? What is the condition to guarantee that one of them wins with a high probability? The following theorem answers these questions.

THEOREM 3.20. *Suppose that a fraction of f_m users bias toward m, a fraction of f_1 users bias toward 1, and other users rate honestly and are neutral. Suppose that the majority rule is adopted to aggregate ratings. These promoters (bias toward m) can win out with high probability only if f_m and f_1 satisfy $f_m > \max\{(\alpha_{i,1} - \alpha_{i,m} + (1 - \alpha_{i,1} + \alpha_{i,m})f_1)/(1 + \alpha_{i,1} - \alpha_{i,m}), (\alpha_{i,\ell_i} - \alpha_{i,m} + \alpha_{i,m}f_1)/(1 - \alpha_{i,m})\}$.*

PROOF. This proof is similar to that of Theorem 3.9. □

Remark. Theorem 3.20 states a condition that promoters can win with a high probability. When we increase the fraction of bad-mouthers, we need to increase the fraction of promoters so that they can win out with a high probability. It is easy to extend the preceding results to the median rule and the average score rule. For the sake of brevity, we omit them in our article.

We now explore the case that both random misbehaving users and biased misbehaving users exist in the system. We seek to explore whether these two types of misbehavior can jointly contribute to rating—that is, reveal the true quality. We assume that biased misbehaving users bias toward ℓ' and bias against the ground truth—that is, $\ell' \neq \ell_i, \ell' \neq l_i$, and $\ell' \neq \gamma_i$. In the following theorem, we derive necessary conditions that f_r, f_b must satisfy so that random misbehavior and biased misbehavior jointly contribute in revealing the true quality.

THEOREM 3.21. *Consider a fraction of f_r random misbehaving users, a fraction of f_b biased misbehaving users, and other users rate honestly and neutral. For the majority rule, it is impossible that the majority rating of the ratings assigned by misbehaving users converges to ℓ_i. For the median rule, if $(l_i/m - 1/2)f_r > (1/2 - \mathbf{I}_{\{\ell' < l_i\}})f_b > ((l_i - 1)/m - 1/2)f_r$, then the median of the ratings assigned by misbehaving users converges to l_i. For the average score rule, if $f_r/f_b = (\gamma_i - \ell')/((m+1)/2 - \gamma_i)$, the mean of the ratings assigned by misbehaving users converges to γ_i.*

PROOF. Please refer to the Appendix for the derivation. □

Remark. It implies that random misbehavior and biased misbehavior may jointly contribute to the ratings by properly mixing them. It suggests a possible approach to employ some random misbehaving users against biased misbehaving users.

Discussion. There are many interesting cases to further explore. For example, some misbehaving users assign random ratings, some promote products, and others bad-mouth products. Due to space limitation, it is impossible for us to explore all of them in detail. We believe that our framework is general enough to the key ideas, and it can be easily extended to explore other cases of misbehavior.

4. INFERRING MODEL PARAMETERS

In this section, we present a maximum likelihood algorithm to *infer* $\boldsymbol{\alpha}_i$ from historical ratings of P_i. With this inference algorithm, we can apply our framework to analyze

real data and improve the applications of online rating systems in Web services (e.g., eBay, TripAdvisor).

Recall that $\mathbf{r}_i^+ = \{r_{i,1}^+, \ldots, r_{i,n_i}^+\}$ represents a set of all observed ratings of product P_i. We seek to infer $\boldsymbol{\alpha}_i$ from $\mathbf{r}_i^+$. Applying Lemma 3.1, we express the likelihood of the parameter $\boldsymbol{\alpha}_i$ given a set of observed ratings $\mathbf{r}_i^+$ as

$$\mathcal{L}(\boldsymbol{\alpha}_i) = \Pr[\mathbf{r}_i^+ | \boldsymbol{\alpha}_i] = \prod_{j=1}^{n_i} \Pr[r_{i,j}^+ | \boldsymbol{\alpha}_i] = \prod_{k=1}^{m} (\alpha_{i,k})^{n_{i,k}}.$$

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

$$L(\boldsymbol{\alpha}_i) = \log \mathcal{L}(\boldsymbol{\alpha}_i) = \sum_{k=1}^{m} n_{i,k} \log \alpha_{i,k} = \sum_{k=1}^{m-1} n_{i,k} \log \alpha_{i,k} + n_{i,m} \log \left(1 - \sum_{k=1}^{m-1} \alpha_{i,k}\right).$$

By maximizing $L(\boldsymbol{\alpha}_i)$, we obtain the maximum likelihood estimation of $\boldsymbol{\alpha}_i$ as follows:

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

Based on this result, we outline the inference algorithm in Algorithm 1.

ALGORITHM 1: Algorithm for inferring $\boldsymbol{\alpha}_i$

Input: A set of ratings of product P_i: $\mathbf{r}_i^+ = \{r_{i,1}^+, \ldots, r_{i,n_i}^+\}$
Output: $\widehat{\boldsymbol{\alpha}}_i$
$n_{i,k} = |\{r_{i,j}^+ | r_{i,j}^+ \in \mathbf{r}_i^+, r_{i,j}^+ = k\}|, k = 1, \ldots, m$;
for *k = 1 to m* **do**
    $\widehat{\alpha}_{i,k} = n_{i,k}/n_i$
end

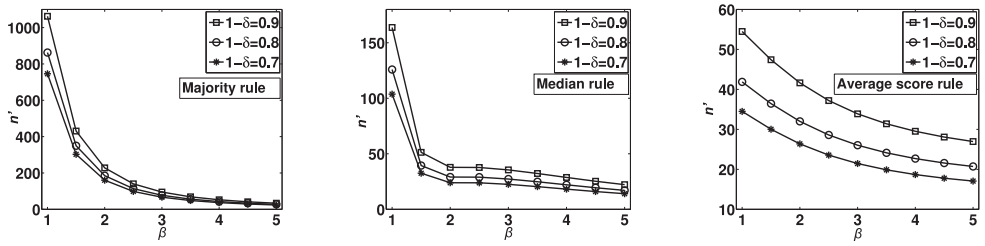
Remark. The running time of this algorithm is $\Theta(|\mathbf{r}_i^+|) = \Theta(n_i)$, or the running time is *linear* to the number of ratings for P_i.

5. EXPERIMENTS ON SYNTHETIC DATA

We carry out experiments on a synthetic dataset to examine various factors that influence the accuracy of product quality assessment. We synthesize a rating dataset that captures important elements of real-world online rating systems. We show that the average score rule requires fewer ratings to reveal the true quality than the majority rule and the median rule under the honest rating and the misbehavior setting.

5.1. Synthetic Dataset

We synthesize a rating dataset that captures important elements of real-world online rating systems and elicits key factors that influence the efficiency and effectiveness of product quality assessment. We first consider the case in which all users rate honestly and are neutral. To be consistent with real-world online rating systems, we consider a five-level cardinal rating metric, say $m = 5$. The smoothness of the model parameter $\boldsymbol{\alpha}_i$ curve is one key factor that influences the minimum number ratings as shown in Section 3. We seek to synthesize $\boldsymbol{\alpha}_i$ by varying its smoothness from low representing that the ratings of P_i have a strong concentration to high representing that the ratings of P_i have a large variation. We formally synthesize $\boldsymbol{\alpha}_i$ by



(a) Majority rule (b) Median rule (c) Average score rule

Fig. 1. Impact of model parameter $\boldsymbol{\alpha}_i$ with success probability $1-\delta$ (honest rating).

$\boldsymbol{\alpha}_i = (1, 2^\beta, 3^\beta, 5^\beta, 4^\beta)/\sum_{k=1}^{5} k^\beta$, where $\beta \in [0, \infty)$ controls smoothness of the model parameter $\boldsymbol{\alpha}_i$ curve. The smaller the value of β, the smoother the curve of $\boldsymbol{\alpha}_i$. For example, $\beta = 0$ implies $\boldsymbol{\alpha}_i \xrightarrow{\beta=0} (\frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5}, \frac{1}{5})$ and β going to infinity implies $\boldsymbol{\alpha}_i \xrightarrow{\beta\to 0} (0, 0, 0, 1, 0)$. In practice, $\boldsymbol{\alpha}_i$ rarely attains these two extremal points, but rather its curve is more likely to attain a medium level of smoothness. To capture this characteristic, we vary β from 1 ($\boldsymbol{\alpha}_i \approx (0.07, 0.13, 0.2, 0.33, 0.27)$) to 5 ($\boldsymbol{\alpha}_i \approx (0.0007, 0.0073, 0.055, 0.706, 0.231)$) in our experiments. We now incorporate misbehavior in rating. Suppose that there is a fraction of f_r random misbehavior, and other users rate honestly and are neutral. We generate a rating by $\boldsymbol{\alpha}_i$ with probability $1 - f_r$, and with probability f_r we draw a rating randomly from $\{1, \ldots, m\}$. This idea can be easily extended to incorporate biased misbehavior or inherent user biases.

5.2. Honest Rating

We explore the impact of the model parameter $\boldsymbol{\alpha}_i$, the success probability $1-\delta$, and the maximum acceptable error E_r on the minimum number of ratings when all users rate honestly and are neutral.

We first explore the impact of $1-\delta$ and $\boldsymbol{\alpha}_i$ on the minimum number of ratings. We vary $1-\delta$ from 0.7 to 0.9 and vary β from 1 to 5. For the average score rule, we set a maximum acceptable error $E_r = 0.5$. We set this maximum acceptable error mainly for the purpose of fair comparison among the majority rule, the median rule, and the average score rule. When the majority rule (or the median rule) is applied to aggregate product ratings, the possible outcome is one of $1, 2, \ldots, m$. The possible outcome can be any value in $[1, m]$ if the average score rule is applied. To fairly compare these three rules, we need to set $2E_r$ to be 1 for the average score rule (i.e., $E_r = 0.5$). Furthermore, $E_r = 0.5$ makes sense in most practical applications. Let n' denote the minimum number of ratings. The numerical results of n' corresponding to the majority rule, the median rule, and the average score rule are shown in Figure 1, where the horizontal axis represents the value of β (i.e., the smoothness of $\boldsymbol{\alpha}_i$). An increase in β results in a decrease in the minimum number of ratings. In other words, the smoother the curve of $\boldsymbol{\alpha}_i$, the larger the minimum number of ratings. Considering the majority rule, an increase in β from 1 to 2 leads to a decrease in the minimum number of ratings from around 1,000 to 200—a significant decrease. A further increase in β from 2 to 5 results in a slight decrease in the minimum number of ratings because the curve for the minimum number of ratings is flat. This implies that the minimum number of ratings is sensitive to the smoothness of $\boldsymbol{\alpha}_i$ when it is high but is invariant of it when it is low. This statement also holds for the median rule. For the average score rule, decreasing the smoothness of $\boldsymbol{\alpha}_i$ only decreases the minimum number of ratings slightly no matter whether the smoothnessis high or low. An increase in the success probability $1-\delta$ results in a slight increase in the minimum number of ratings because the curves corresponding to $1-\delta = 0.7$, 0.8 and 0.9 nearly overlap.

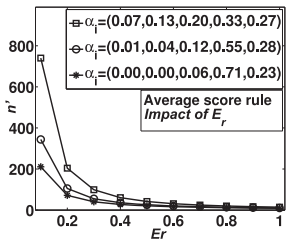


Fig. 2. Impact of maximum acceptable error E_r (honest rating).

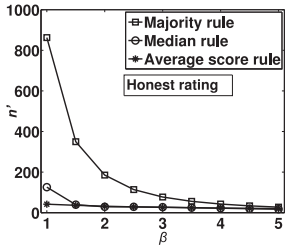


Fig. 3. Comparing the minimum number of ratings under the honest rating.

We now explore the impact of E_r on the minimum number of ratings. We vary E_r from 0.1 to 1 and fix $1-\delta = 0.8$. We set the default success probability as $1-\delta = 0.8$. We use this default success probability because it makes sense for many practical applications. Furthermore, the minimum number of ratings increases in a logarithmic rate to the success probability, and this default success probability is mainly used for comparison studies, such as comparing the efficiency of rating aggregation rules. We choose three representatives of $\boldsymbol{\alpha}_i$ corresponding to $\beta = 1, 3$, and 5 (representing high, medium, and low levels of smoothness, respectively). The numerical results of n' are shown in Figure 2. An increase in the maximum acceptable error E_r results in a decrease in the minimum number of ratings. When $E_r \leq 0.2$, decreasing E_r can decrease the minimum number of ratings remarkably. When $E_r \geq 0.2$, decreasing E_r decreases the minimum number of ratings slightly. The smoothness has a small impact on the minimum number of ratings, as these three curves nearly overlap together.

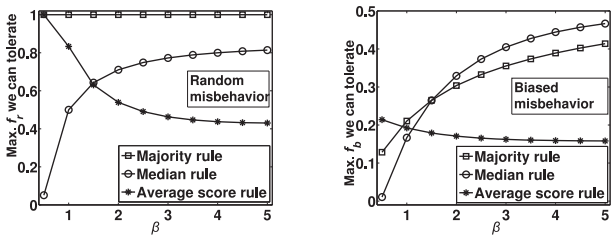
We compare the minimum number of ratings that each rating aggregation rule requires. Figure 3 presents the minimum number of ratings required by the majority rule, the median rule, and the average score rule, respectively. To achieve the same success probability, the average score rule requires the smallest number of ratings.When β is smaller than 1.5 (i.e., a high smoothness of $\boldsymbol{\alpha}_i$), the average score rule requires remarkably fewer ratings than the majority rule and the median rule.

Lessons learned. Assume that all users rate honestly and are neutral. To increase the success probability, we need to increase the minimum number of ratings slightly. A small increase in the smoothness of $\boldsymbol{\alpha}_i$ leads to a significant increase in the minimum number of ratings for the majority rule and the median rule and a slight increase on the minimum number of ratings for the average score rule. The average score rule requires fewer ratings than the majority rule and the median rule.

5.3. Impact of Misbehavior

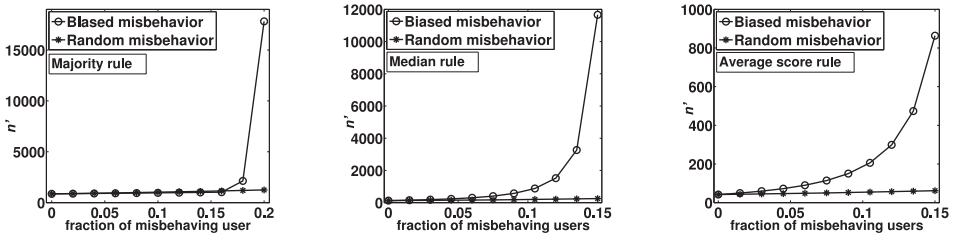
We explore the maximum fraction of misbehaving users that each rating aggregation rule can tolerate and the minimum number of ratings needed to compensate. We compare the robustness of the majority rule, the median rule, and the average score rule against misbehavior as well.

We first explore the maximum fraction of misbehaving users that each rating aggregation rule can tolerate. We set the success probability as $1-\delta = 0.8$ and the maximum acceptable error as $E_r = 0.5$. Figure 4 shows the maximum fraction of random misbehaving users $\overline{f}_r$ and the maximum fraction of random misbehaving users $\overline{f}_b$ that can be tolerated, where the horizontal axis represents β (or the smoothness of $\boldsymbol{\alpha}_i$). For the majority rule and the median rule, an increase in β results in an increase in $\overline{f}_r$ and $\overline{f}_b$. Namely, the lower the smoothness of $\boldsymbol{\alpha}_i$, the larger the fraction of misbehavior that



(a) Random misbehavior (b) Biased misbehavior

Fig. 4. The maximum fraction of misbehaving users that we can tolerate.

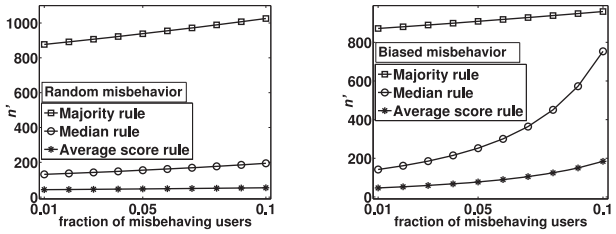


(a) Majority rule (b) Median rule (c) Average score rule rule

Fig. 5. Impact of random and biased misbehavior.

can be tolerated. The average score rule shows a contrary trend—that is, the lower the smoothness of $\boldsymbol{\alpha}_i$, the smaller the fraction of misbehavior that can be tolerated. The average score rule can tolerate at least a fraction of 0.4 random misbehaving users and around a fraction of 0.2 biased misbehaving users. When $\beta = 0.5$, the median rule can tolerate less than a fraction of 0.02 misbehaving (random or biased) users. This implies that when the smoothness of $\boldsymbol{\alpha}_i$ is high, the median rule is extremely vulnerable to misbehavior. The majority rule can always tolerate random misbehavior because $\overline{f}_r = 1$. When $\beta = 0.5$, the majority rule can only tolerate around a fraction of 0.1 biased misbehaving users.

We now explore the minimum number of ratings needed to compensate to tolerate misbehavior. We use the same experimental settings as earlier. For the ease of presentation, we choose one representative $\boldsymbol{\alpha}_i$ to study—that is, $\boldsymbol{\alpha}_i = (\frac{1}{15}, \frac{2}{15}, \frac{3}{15}, \frac{5}{15}, \frac{4}{15})$. The curve for this parameter is smoother than the rating distribution for most products in practice. Hence, this parameter gives an upper bound on the minimum number of ratings needed to compensate. We vary the fraction of misbehaving users (random misbehavior or biased misbehavior) from 0 to 0.2 (the majority rule) and 0 to 0.15 (the median rule and the average score rule). We choose thissetting mainly because the maximum fraction of misbehaviors that can be tolerated across different rating aggregation rules is different. Let n' denote the minimum number of ratings needed to compensate. Figure 5 presents the numerical results of n', where the horizontal axis represents the fraction of misbehaving users. An increase in the fraction of random misbehaving users results in a slight increase in the minimum number of ratings, because the curves for the minimum number of ratings are flat. When the fraction of biased misbehaving users is below 0.1, an increase in the fraction of biased misbehaving users results in a slight increase in the minimum number of ratings. When that fraction is above 0.1, a small increase in the fraction of biased misbehaving users can remarkably increase the minimum number of ratings. The majority rule, the median



(a) Random misbehavior (b) Biased misbehavior

Fig. 6. Comparing the minimum number of ratings under misbehavior.

Table X. Statistics for Three Rating Datasets

	TripAdvisor	Amazon
Number of Items	11,540	32,888
Total Number of Ratings	3,114,876	5,066,070
Maximum/Minimum on Number of Ratings	9930/1	24,195/1
Mean/Median on Number of Ratings	270/179	154/47
Rating Metric: $\{1, \ldots, m\}$	$1, \ldots, 5$	$1, \ldots, 5$

rule, and the average score rule require more ratings to tolerate biased misbehavior than random misbehavior.

We compare the minimum number of ratings that each rating aggregation rule requires in tolerating misbehavior. We vary the fraction of misbehaving (random or biased) users from 0.001 to 0.1. Figure 6 shows the minimum number of ratings needed to compensate in tolerating random and biased misbehavior, where the horizontal axis represents the fraction of misbehaving users. The average score rule is shown to be more robust against misbehavior (random misbehavior or biased misbehavior), because to tolerate the same fraction of misbehaving users, it requires fewer ratings than the majority rule and the median rule.

Lessons learned. A certain fraction of random misbehavior or biased misbehavior can be tolerated by compensating a number of ratings. An increase the smoothness of $\boldsymbol{\alpha}_i$ results in a significant increase in the maximum fraction of misbehaving users that the majority rule and the median rule can tolerate and results in a slight in decrease in the maximum fraction of misbehaving users that the average score rule can tolerate. Tolerating biased misbehavior requires a lot more ratings than tolerating random misbehavior. The average score rule has more robustness than the majority rule and the median rule, as it requires fewer ratings than the majority rule and the median rule under the honest rating and the misbehavior setting.

6. EXPERIMENTS ON REAL DATA

We present experimental results on two *large* datasets from TripAdvisor and Amazon. We validate our model and apply the minimum number of ratings to explore the reliability of rating aggregation rules, the reliability of ratings, and the rating sufficiency for TripAdvisor and Amazon. Last, we explore the maximum fraction of misbehaving (random or biased) users that an item can tolerate and the minimum number needed to compensate.

6.1. Datasets

We crawled ratings from TripAdvisor and Amazon, which are shown in Table X.

Table XI. Number of Selected Items for Model Validation

	TripAdvisor	Amazon
Number of Selected Items	2,368	2,396
Total Number of Ratings	2,030,716	2,857,823

TripAdvisor. TripAdvisor is a popular travel Web site that assists customers in booking hotels, restaurants, and so forth and sharing opinions (or experiences) in the form of ratings (or reviews) of the businesses. We crawled historical ratings of 11,540 hotels.

Amazon. Amazon is a popular E-commerce system that assists customers in product adoption and sharing opinions (or knowledge) on products in the form of ratings (or reviews). We crawled the historical ratings of 32,888 products.

6.2. Model Validation

We validate our model by showing that if an item meets the requirement on the minimum number of ratings, the assessment of its quality is indeed reliable; otherwise, this assessment is unreliable.

We select a subset of items from our datasets to valid our model. More precisely, we select items that have at least 400 ratings. We justify this selecting criterion as follows. We seek to apply Algorithm 1 to estimate the parameter $\boldsymbol{\alpha}_i$ accurately for each selected item. The larger the number of ratings for an item, the higher the estimation accuracy. However, a high selection criterion, such as selecting those with thousands of ratings, results in only a small number of items being selected. Examining our datasets, we set a selecting criterion of 400 to attain a good trade-off between the estimating accuracy and the sufficiency of our validation datasets. Table XI shows the number of selected items and the corresponding total number of ratings. In total, we select $2,368\,(\text{TripAdvisor}) + 2,396\,(\text{Amazon}) = 4,764$ items out. For the ease of presentation, we denote these 4,764 items by $P'_1, \ldots, P'_{4,764}$ and use $\mathbf{r}'_i$ to denote a set of all historical ratings of P'_i.

We now extract the minimum number of ratings and the true quality of each selected item. We apply Algorithm 1 on the historical ratings of $P'_i, i \in \{1, \ldots, 4,764\}$, to estimate $\boldsymbol{\alpha}_i$. We denote this estimation by $\widehat{\boldsymbol{\alpha}}_i$. We treat $\widehat{\boldsymbol{\alpha}}_i$ as the true value of $\boldsymbol{\alpha}_i$ and we extract the true quality of P'_i from it as follows: $\ell_i = \arg\max_k\{\widehat{\alpha}_{i,k}\}$(majority rule), $l_i = \arg\min\{k|\sum_{j=1}^{k}\widehat{\alpha}_{i,j} > \frac{1}{2}\}$(median rule), $\gamma_i = \sum_k k\widehat{\alpha}_{i,k}$(average score rule). We apply the bounds derived in Theorems 3.4, 3.11, and 3.16 to $\widehat{\boldsymbol{\alpha}}_i$ to extract minimum number of ratings that P'_i needs—that is, $n'_i = 2((\sum_k k^2\widehat{\alpha}_{i,k} - (\sum_k k\widehat{\alpha}_{i,k})^2)/E_r^2 + \frac{m}{3E_r})\ln\frac{2}{\delta}$ (average score rule)—setting the success probability to $1-\delta = 0.8$ and $E_r = 0.5$.

We now present the design of our model validation algorithm. We seek to quantify the reliability of a product quality assessment when the number of ratings of a product exceeds and is smaller than the minimum number of ratings, respectively. Ratings in our datasets are associated with time stamps. Consider a rating time stamp of P'_i. We say that it meets the minimum requirement if the number of ratings of P'_i (up to this time stamp) exceeds the minimum number of ratings; otherwise, we say that it does not meet the minimum requirement. For each rating time stamp of P'_i, we apply a rating aggregation rule to aggregate the historical ratings of P'_i (up to this time stamp). If this aggregate rating reveals the true quality, we say that this time stamp is reliable; otherwise, we say that it is unreliable. Let N^i_{test} and $\overline{N}^i_{test}$ denote the number of rating time stamps of P'_i that meet and do not meet the minimum requirement, respectively. Let $N^i_{reliable}$ and $\overline{N}^i_{reliable}$ denote the number of reliable time stamps of P'_i that meet and

Table XII. Fraction of Reliable Evaluations When Items Meet (or Do Not Meet) the Requirement on Minimum Number of Ratings

Aggregation Rule	N_{test}	$N_{reliable}$	$f_{reliable}$	f_{test}	$\overline{N}_{test}$	$\overline{N}_{reliable}$	$\overline{f}_{reliable}$
TripAdvisor ($N_{total} = 2,030,716$)							
Average score	1,957,727	1,918,157	97.98%	94.41%	72,989	48,620	66.61%
Majority	1,307,524	1,299,767	99.41%	64.39%	723,192	560,181	77.46%
Median	1,502,809	1,456,133	96.89%	74.00%	527,907	364,815	69.11%
Amazon ($N_{total} = 2,857,823$)							
Average score	2,763,490	2,693,606	97.47%	96.70%	94,333	63,075	66.91%
Majority	2,601,660	2,574,363	98.95%	91.04%	256,163	192,826	75.27%
Median	2,212,826	2,128,099	96.17%	77.43%	644,997	431,596	66.92%

do not the minimum requirement, respectively. Denote $f_{reliable} = \sum_i N^i_{reliable} / \sum_i N^i_{test}$, which quantifies the reliability of a product quality assessment when the number of ratings exceeds the minimum number of ratings. Denote $\overline{f}_{reliable} = \sum_i \overline{N}^i_{reliable} / \sum_i \overline{N}^i_{test}$, which quantifies the reliability of a product quality assessment when the number of ratings is smaller than the minimum number of ratings. We compute these statistics in Algorithm 2, where we fix the rating aggregation rule to be the average score rule for brevity.

ALGORITHM 2: Algorithm for model validation

Input: A set of products $\{P'_{i_1}, \ldots, P'_{i_n}\}$, average score rule, δ, E_r

Output: N_{test}, $N_{reliable}$, $f_{reliable}$, $\overline{N}_{test}$, $\overline{N}_{reliable}$, $\overline{f}_{reliable}$

$N_{test} \leftarrow 0, N_{reliable} \leftarrow 0, \overline{N}_{test} \leftarrow 0, \overline{N}_{reliable} \leftarrow 0$;

for $\kappa \in \{i_1, \ldots, i_n\}$ **do**

  $N^\kappa_{test} \leftarrow 0, N^\kappa_{reliable} \leftarrow 0$;

  $\widehat{\boldsymbol{\alpha}}_\kappa \leftarrow$perform Algorithm 1 on $\mathbf{r}'_\kappa$; /* estimate model parameter for item P'_κ */

  $\gamma_\kappa = \sum_j j\widehat{\alpha}_{\kappa,j}$; /* true quality */

  $n'_\kappa \leftarrow 2((\sum_k k^2\widehat{\alpha}_{i,k} - (\sum_k k\widehat{\alpha}_{i,k})^2)/E_r^2 + \frac{m}{3E_r})\ln\frac{2}{\delta}$; /* compute the minimum number of ratings,via substituting $\boldsymbol{\alpha}_\kappa$with $\widehat{\boldsymbol{\alpha}}_\kappa$in the bound derive in Theorem 3.16. */

  sort the ratings of $\mathbf{r}'_\kappa$based on time stamps; /* $r'_{\kappa,1}$ denotes the earliest rating */

  for $j = 1$ *to* $|\mathbf{r}'_\kappa|$ **do** /* rating reliability test */

    If $j \geq n'_\kappa$, update $N^\kappa_{test} \leftarrow N^\kappa_{test} + 1$; otherwise, $\overline{N}^\kappa_{test} \leftarrow \overline{N}^\kappa_{test} + 1$;

    $\widehat{\gamma}_\kappa \leftarrow$ the average rating of $\{r'_{\kappa,1}, \ldots, r'_{\kappa,j}\}$; /* evaluate quality */

    if $|\widehat{\gamma}_\kappa - \gamma_\kappa| < E_r$ **then** /* the evaluated quality equals to true quality */

      If $j \geq n'_\kappa$, update $N^\kappa_{reliable} \leftarrow N^\kappa_{reliable} + 1$, otherwise $\overline{N}^\kappa_{reliable} \leftarrow \overline{N}^\kappa_{reliable} + 1$;

    end

  end

end

$N_{reliable} \leftarrow \sum_{\kappa\in\{i_1,\ldots,i_n\}} N^\kappa_{reliable}$, $N_{test} \leftarrow \sum_{\kappa\in\{i_1,\ldots,i_n\}} N^\kappa_{test}$, $f_{reliable} = N_{reliable}/N_{test}$;

$\overline{N}_{reliable} \leftarrow \sum_{\kappa\in\{i_1,\ldots,i_n\}} \overline{N}^\kappa_{reliable}$, $\overline{N}_{test} \leftarrow \sum_{\kappa\in\{i_1,\ldots,i_n\}} \overline{N}^\kappa_{test}$, $\overline{f}_{reliable} = \overline{N}_{reliable}/\overline{N}_{test}$;

We apply Algorithm 2 on the rating datasets described in Table XI. Table XII presents the experimental results of $f_{reliable}$ and $\overline{f}_{reliable}$, where N_{total} denotes the total number of rating time stamps used in our experiments and $f_{test} = N_{test}/N_{total}$ denotes the fraction of rating time stamps that meets the minimum requirement. Consider the average score rule and TripAdvisor; we have $f_{reliable} = 97.98\%$ and $\overline{f}_{reliable} = 66.61\%$. Changing the average score rule to the majority rule (or the median rule) varies $f_{reliable}$ and $\overline{f}_{reliable}$ slightly. It implies that if a hotel on TripAdvisor meets the requirement on the minimum

Table XIII. Overall Statistics of Minimum Number of Ratings across Items

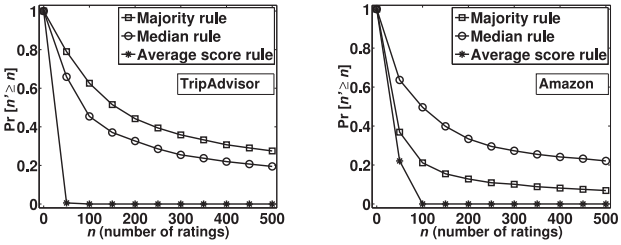
	Median on n'	Mean on n'	Minimum on n'	Maximum on n'
TripAdvisor				
Majority Rule	161	15,204	6	9,089,015
Median Rule	84	7,625	4	2,490,080
Average Score Rule	32	32	18	58
Amazon				
Majority Rule	35	3,476	5	4,511,992
Median Rule	97	15,982	4	5,585,767
Average Score Rule	40	41	18	80

number of ratings, then the assessment of its quality is indeed reliable; otherwise, this assessment is unreliable. This statement also holds for items on Amazon. Hence, our model is correct and captures important elements of real-world online rating systems. Consider TripAdvisor and the average score rule; we have $f_{test} = 94.41\%$—that is, 94.41% of the rating time stamps meet the requirement on the minimum number of ratings. Changing the dataset to Amazon results in $f_{test} = 96.70\%$. Namely, the minimum number of ratings is practical and applicable. The average score rule is more robust than the majority rule and the median rule, as it has more rating time stamps that satisfy the minimum requirement. Take TripAdvisor as an example; the number of rating time stamps that meet the minimum requirement for the average score rule is around 1.5 times ($0.9441 / 0.9439 \approx 1.5$) thenumber for the majority rule and 1.3 times ($0.9441 / 0.7400 \approx 1.3$) the number for the median rule. We explore why this is the case in the next section.

6.3. Applications of Minimum Number of Ratings

We apply the minimum number of ratings to explore the reliability of rating aggregation rules, the reliability of ratings, and the rating sufficiency for TripAdvisor and Amazon, respectively.

Reliability of rating aggregation rules. We explore the minimum number of ratings across items and show that the average score is more reliable (i.e., requires fewer ratings) than the majority rule and the median rule. We consider the items described in Table XI and the same settings as in Section 6.2. We first explore overall statistics of the minimum number of ratings across items (i.e., mean, median, maximum, and minimum, which are shown in Table XIII), where n' denotes the minimum number of ratings that an item needs. Consider a maximum of n'; the average score rule requires 58 (TripAdvisor) and 80 (Amazon) ratings, and the majority rule and the median rule require millions of ratings, respectively. Namely, in the worst case, the average score rule requires extremely fewer ratings than the majority rule and the median rule. This statement also holds in the average case—that is, comparing these three rules using the mean of n'. Examining the minimum of n', these three rules require fewer than 60 ratings. The median of n' is around 100. In other words, 100 ratings are sufficient for half of the items. Consider the median of n'; the average score rule requires fewer ratings than the majority rule and the median rule, except for the Amazon dataset, where the average score requires fewer ratings than the median rule but 5 more ratings than the majority rule. We now explore the complementary cumulative distribution function of minimum number of ratings across items. Let $\Pr[n' \geq n]$ denote the fraction of items that require a minimum number of ratings larger or equal to n. Figure 7 shows the numerical results of $\Pr[n' \geq n]$. The complementary cumulative distribution function curve corresponding to average score rule lies under the curve corresponding to the majority rule and the median rule. This implies that the average score rule



(a) TripAdvisor (b) Amazon

Fig. 7. Distribution of minimum number of ratings across items.

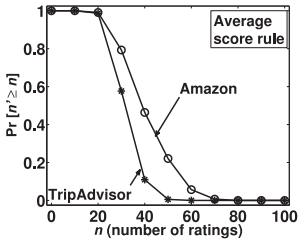


Fig. 8. Comparing reliability of rating on TripAdvisor and Amazon.

Table XIV. Fraction of Items Satisfying the Requirement of Minimum Number of Ratings

	Minimum Number of ratings	Number of Iems	N_s	f_s
TripAdvisor	32	11,540	9,033	78.28%
Amazon	40	32,888	17,914	54.47%

requires fewer ratings than the majority rule and the median rule. In summary, the average score rule is more reliable than the majority rule and the median rule.

Rating reliability. We explore the following question: Between TripAdvisor and Amazon, whose ratings are more reliable? We answer this question by examining the complementary cumulative distribution function of minimum number of ratings across items. Here we fix the rating aggregation rule to be the average score rule. Again, $\Pr[n' \geq n]$ denotes the fraction of items that require a minimum number of ratings larger or equal to n. Figure 8 presents the numerical results of $\Pr[n' \geq n]$. The complementary cumulative distribution function curve corresponding to TripAdvisor lie in the bottom. This implies that TripAdvisor requires fewer ratings than Amazon. In other words, the ratings on TripAdvisor are more reliable than those on Amazon.

Rating sufficiency. We examinethe following question: What is the fraction of items on TripAdvisor (or Amazon) that have a sufficient number of ratings? Assume that the items described in Table X are representative samples from TripAdvisor and Amazon. We fix the rating aggregation rule to be the average score rule. From Table XIII, we obtain that the medians of the minimum number of ratings across items are 32 (TripAdvisor) and 40 (Amazon). We use them as the condition for testing rating sufficiency—for example, we say that a product on Amazon has a sufficient number of ratings if it has no fewer than 40 ratings. We perform this test on all items described in Table X. Let f_s denote the fraction of items that have a sufficient number of ratings. The numerical results f_s are shown in Table XIV, where N_s denotes the number of items that have a sufficient number of ratings. Only 78.28% hotels on TripAdvisor and 54.47% products on Amazon have a sufficient number of ratings. TripAdvisor has a higher rating sufficiency than Amazon.

Lessons learned and tips. The average score rule requires fewer ratings than the majority rule and the median rule. Assuming that we apply the average score rule to aggregate ratings, around 70 ratings can guarantee a reliable assessment on the quality of an item. Ratings on TripAdvisor are more reliable than those on Amazon. TripAdvisor has a higher rating sufficiency than Amazon.

6.4. Analysis of Rating under Misbehavior

We explore the maximum fraction of misbehaving (random or biased) users that an item can tolerate and the minimum number needed to compensate.

Table XV. Maximum Fraction of Random Misbehaving Users $\overline{f}_r$ That We Can Tolerate

	Mean on $\overline{f}_r$	Median on $\overline{f}_r$	Max on $\overline{f}_r$	Min on $\overline{f}_r$
TripAdvisor	0.5145	0.4409	1	0.2599
Amazon	0.4727	0.3914	1	0.2565

Table XVI. Minimum Number of Ratings to Tolerate Random Misbehavior

	Mean on n'	Median on n'	Max on n'	Min on n'
TripAdvisor	55	55	74	35
Amazon	69	69	100	46

Table XVII. Maximum Fraction of Biased Misbehaving Users $\overline{f}_b$ That We Can Tolerate

	Mean on $\overline{f}_b$	Median on $\overline{f}_b$	Max on $\overline{f}_b$	Min on $\overline{f}_b$
TripAdvisor	0.1655	0.1595	0.5619	0.1247
Amazon	0.1636	0.1526	1	0.1266

Random misbehavior. We inject some ratings by random misbehaving users into the rating dataset described in Table XI. We explore the maximum fraction of random misbehaving users that a rating aggregation rule can tolerate first. We pick the average score rule to study for brevity. Recall that $\overline{f}_r$ denotes the maximum fraction of random misbehaving users that the average score rule can tolerate. Table XV presents the mean, median, maximum, and minimum of $\overline{f}_r$ across items. Consider TripAdvisor; the minimum and maximum of $\overline{f}_r$ are 0.2599 and 1, respectively. In other words, items on TripAdvisor can tolerate a fraction of at least 0.25 and at most 1 random misbehaving users. This statement also holds for Amazon. Most items on TripAdvisor (or Amazon) can tolerate a fraction of 0.4 random misbehaving users, as the mean and the median of $\overline{f}_r$ are around 0.4. We now inject a fraction of 0.1 random misbehaving ratings into the dataset described in Table XI and explore the minimum number of ratings needed to tolerate them, denoted by n'. Table XVI shows the mean, median, maximum, and minimum of n' across items. Items on TripAdvisor can tolerate the random misbehavior by compensating at most 74 ratings, as the maximum of n' is 74. Similarly, items on Amazon require compensating at most 100 ratings.

Biased misbehavior. We inject some ratings by biased misbehaving users (bias toward 1) into the rating dataset described in Table XI. We explore the maximum fraction of biased misbehaving users that a rating aggregation rule can tolerate first (i.e., $\overline{f}_b$). Again, we pick the average score rule to study. Table XVII shows the mean, median, maximum, and minimum of $\overline{f}_b$ across items. Hotels on TripAdvisor can tolerate a fraction of at least 0.1274 and at most 0.5619 biased misbehaving users, respectively. Most hotels on TripAdvisor can tolerate a fraction of around 0.16 biased misbehaving users, as the mean and median of $\overline{f}_b$ are around 0.16. We now inject a fraction of 0.1 biased misbehaving ratings into the rating dataset described in Table XI and explore the minimum number of ratings needed to tolerate them denoted by n'. Table XVIII shows the mean, median, maximum, and minimum of n' across items. Hotels on TripAdvisor require compensating at most 666 ratings, as the maximum of n' is 666. Similarly, products on Amazon require compensating at most 697 ratings. Most items on TripAdvisor (Amazon) require compensating around 268 (370) ratings, as the corresponding medians of n' is 268 (370).

Lessons learned. Suppose that we adopt the average score ruleto aggregate ratings. Most items on TripAdvisor (or Amazon) can tolerate a fraction of around 0.4/0.15 random/biased misbehaving users, and they require compensating around 70/300 ratings to tolerate a fraction of 0.1 rrandom/biased misbehaving users.

Table XVIII. Minimum Number of Ratings to Tolerate Biased Misbehavior

	Mean on n'	Median on n'	Max on n'	Min on n'
TripAdvisor	285	268	666	53
Amazon	370	370	697	43

Table XIX. Statistics for Two Movie Rating Datasets

	Netflix	Flixster
Number of Movies	17,770	48,794
Number of Users	480,189	147,612
Total Number of Ratings	100,480,507	8,196,077
Mean Number of Ratings per Movie	5,655	168
Mean Number of Ratings per User	209	56
Rating Metric: $\{1, \ldots, m\}$	$1, \ldots, 5$	$1, 1.5, \ldots, 5$

7. APPLICATIONS TO RECOMMENDATION SYSTEMS

We perform experiments on two movie rating datasets from Netflix and Flixster to demonstrate how to apply our framework to recommendation systems.

7.1. Dataset

Table XIX shows the overall statistics of our dataset from Netflix and Flixster. Netflix is a popular Web site that provides on-demand Internet streaming media. It makes personalized video recommendations based on ratings and reviews by users. We use the version of Netflix movie rating dataset provided by the Netflix prize competition [Netflix 2009]. Flixster is a popular Web site that allows users to discover new movies and learn about movies. It maintains a personalized movie recommendation system based on ratings (or reviews) by its users and the social relationship between users.We use the movie rating dataset released by Jamali and Ester [2010].

7.2. Applications to Personalized Recommendation Systems

We apply our framework to examine the minimum number of ratings that a user needs to express to reflect his bias or leniency accurately.

Personalized recommendation systems (also known as recommender systems) [Adomavicius and Tuzhilin 2005; Resnick and Varian 1997] recommend items to a user taking into account the preference of that user. Such recommendation tasks rely on rating prediction—that is, predicting a user's potential ratings to items. The average of the historical ratings by a user (average rating of a user) acts as a normalizing factor [Adomavicius and Tuzhilin 2005] in most rating prediction algorithms, because it reflects the inherent bias or leniency of a user. Furthermore, due to different rating habits of users, many recommendation algorithms employ rating normalization [Jin and Si 2004] techniques, which convert different users rating to a normalized scale applying the average rating of a user. Hence, it is important to estimate the average rating of a user accurately. However, we only have partial information on these ratings—that is, each user only expresses ratings to a small subset of products, which makes it difficult to accurately estimate the average rating of a user.

We now apply our framework to determine the minimum number of ratings that a user needs to express to guarantee an accurate estimation on his average rating. To apply our framework, we only need to interchange the role between the users and the products. Similar to Section 6, we select the users who express at least 400 ratings to study. In our dataset, 74,509 users from Netflix and 5,306 users from Flixster satisfy this selection criterion. We then compute the minimum of ratings for each selected user using the same method as that described in Section 6. Let n' denote the minimum

Table XX. Statistics of the Minimum Number of Ratings That a User Needs to Express

	Mean on n'	Median on n'	Minimum on n'	Maximum on n'
Netflix	32	32	16	87
Flixster	84	79	28	305

number of ratings that a user needs to express. Table XX shows the mean, median, maximum, and minimum of n' across users. Netflix users need to express at most 87 ratings, as the maximum of n' is 87. Most of them need to express 32 ratings, as the mean and the median of n' are both 32. Flixster users need to express at most 305 ratings, and most of them need to express 80 ratings.

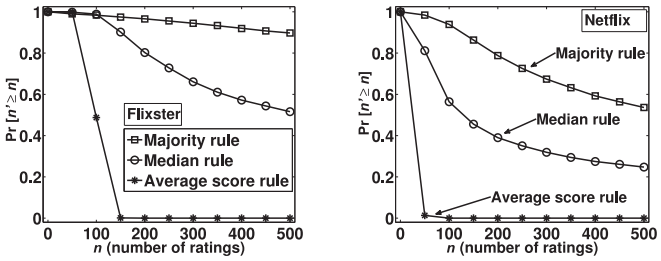
Lessons learned and implications. To reflect a user's bias or leniency accurately via the average rating, a Netflix user needs to express around 30 ratings and a Flixster user needs to express around 80 ratings.

7.3. Applications to Group Recommendation Systems

We apply our framework to examine rating sufficiency conditions for group recommendation systems.

Group recommendation systems [Boratto and Carta 2011] were introduced to deal with the contexts that users operate in *groups*. Such recommendation tasks rely on eliciting collective preferences of the user group. However, we only have partial preference information (i.e., each user only expresses ratings to a *subset* of products), which makes it difficult to make an accurate recommendation. Generally, group recommendation algorithms are heuristics, and their accuracy relies heavily on whether items receive sufficient ratings. More concretely, when items have sufficient ratings, some very simple recommendation algorithms can make accurate recommendations, but if not, even the most sophisticated recommendation algorithm can make poor recommendations. Hence, determining the rating sufficiency condition is important. However, this is a challenging task. Our objective is to demonstrate how to apply our framework to address these challenges. We explore the following questions: What is the minimum number of ratings that a product needs so that the system can accurately elicit the collective preference of a user group? Considering the average score rule, the majority rule, and the median rule, which one is more efficient for a group recommendation task?

We apply our framework to explore the minimum number of ratings that a movie needs. We treat all Netflix users (or Flixster users) as a whole group, and the system seek to recommend movies to this group. Similar to Section 6, we select the movies with at least 400 ratings to study. In our dataset, 2,807 movies from Flixster and 10,135 movies from Netflix satisfy this selection criterion. We then compute the minimum of ratings for each selected movie using the same method as that described in Section 6. Let n' denote the minimum number of ratings that a movie needs. Let $\Pr[n' \geq n]$ denote the fraction of movies with a minimum number of ratings larger than or equal to n. Figure 9 shows the numerical results of $\Pr[n' \geq n]$ corresponding to Flixster and Netflix. The complementary cumulative distribution function curve corresponding to the average score rule lies in the bottom. In other words, the average score rule requires fewer ratings than the majority rule and median rule. The majority rule requires a lot more ratings than the median rule and average score rule. Hence, the average score rule is more efficient than the majority rule and the median rule for group recommendation tasks. Table XXI presents the mean, median, maximum, and minimum of n' corresponding to the average score rule. Netflix movies need at most 64 ratings, as the maximum of n' is 64. Most Netflix movies need 36 ratings, as the



(a) Flixster (b) Netflix

Fig. 9. Distribution of minimum number of ratings across movies.

Table XXI. Statistics of the Minimum Number of Ratings That a Movie Needs to Receive

	Mean on n'	Median on n'	Maximum on n'	Minimum on n'
Netflix	37	36	23	64
Flixster	101	100	52	157

mean and the median of n' are around 36. Similarly, Flixster movies need at most 157 ratings, and most of them need a 100 rating.

Lessons learned and tips. The average score rule requires fewer ratings than the majority rule and median rule for group product recommendation tasks. Suppose that we adopt the average score rule; around 36 ratings are sufficient for Netflix movies and 100 ratings are sufficient for Flixster movies.

8. RELATED WORK

A variety of works studied adoption maximization for a given product. These works are orthogonal to ours in that increasing the number of adoptions may lead to an increase in the number of users who express ratings to a product. Namely, these works increase the rating sufficiency indirectly. Our work aims to establish rating sufficiency conditions. There are two typical approaches to maximize product adoptions. One approach maximizes the influence of a product across a social network [Bhagat et al. 2012; Goyal et al. 2010; Kempe et al. 2003; Yang and Leskovec 2010]. The works using this approach assumed an adoption (or influence) propagation mechanism that an adoption by a user triggers the same adoption by his neighbors (or friends) with some probabilities. Their objective is to determine a small subset of initial adoptions that can maximize the expected total number of adoptions. Another approach predicts product adoptions via modeling and analyzing social correlations [Chua et al. 2011, 2013]. This works using this approach predicted individual adoptions using a user–item adoption network and a user–user social network.

Our work is related to works on score normalization. Score normalization normalizes ratings assigned by different users to the same scale by some criteria. It is motivated by assumption that users have different rating habits in many applications, such as recommender systems. For example, some lenient users assign higher ratings, whereas some critical users assign lower ratings. There are two representative types of score normalization approaches. One is the deterministic approach, which converts a rating to a specific score. For example, Resnick et al. [1994] developed a z-score normalization–based method, Sarwar et al. [2000] proposed a dimensionality reduction–based method, Lemire [2005] proposed an L_p normalization method, and Traupman and Wilensky [2006] proposed a factor analysis–based method. Another one is the probabilistic approach, which converts a user rating to a probability value [Fernández et al. 2006; Jin

et al. 2003; Jin and Si 2004]. This approach requires a larger number of ratings by a user to guarantee a good estimation on this probability value.

Rating aggregation rules have been studied extensively. A class of rating aggregation rules that incorporate human factors such as bias, leniency, and controversy were developed in Lauw et al. [2008, 2012]. Reputation-aware aggregation rules [Chen and Singh 2001; Riggs and Wilensky 2001] compute the average rating of products weighted on the reputation score of users. Other nice rating aggregation rules were developed in Jin and Si [2004], Jin et al. [2003], and Traupman and Wilensky [2006]. Our work differs from them in that we study the three most widely used rating aggregation rules—the majority rule, median rule, and average score rule—and explore under which condition they can produce reliable product quality evaluation.

Several works have investigated fraud detection.For example, a review of spam detection in online reviewing systems was explored in Jindal and Liu [2007, 2008]. Fraud detection in trading communities was studied in Dellarocas [2000] and Zhang and Cohen [2006]. Fraud detection in recommendation systems was investigated in Lam and Riedl [2004] and Mobasher et al. [2006]. The robustness of collaborative filtering algorithms was studied in Van Roy and Yan [2010]. Our work provides a general probabilistic model and analysis of misbehaving users in ratings systems and establishes the condition under which a rating system may fail to reflect the true quality of products.

Online rating systems are widely deployed in recommendation systems [Adomavicius and Tuzhilin 2005; Resnick and Varian 1997]. Recommendation systems were introduced following the seminal work on collaborative filtering [Hill et al. 1995; Resnick et al. 1994]. In general, recommendation systems interpret ratings as preferences of users and try to make personalized recommendations taking into account their preferences. Researchers have investigated various algorithmic and complexity issues in designing recommendation systems [Adomavicius and Tuzhilin 2005; Herlocker et al. 2004; Resnick and Varian 1997]. Our work differs from them in that we treat ratings as product quality assessment and explore the condition under which we can reveal the true quality with high probability.

9. CONCLUSION AND FUTURE WORK

This article presented a general model and analysis of product rating with partial information. We explored a probabilistic model to capture various important factors of an online rating system. We derived the minimum number of ratings needed to reveal the quality of a product. We extended our model to accommodate users' misbehavior in product rating. We derived the maximum fraction of misbehaving users that a rating aggregation rule can tolerate and the minimum number of ratings needed to compensate. We performed experiments using both synthetic and real-world data (from Amazon and TripAdvisor). We validated our model and showed that the average score rule produces more reliable and robust product quality assessments than the majority rule and the median rule. We found that around 100 ratings are sufficient to reflect the true quality of an item on TripAdvisor and Amazon. Ratings on TripAdvisor are more reliable than those found on Amazon. TripAdvisor has a higher rating sufficiency than Amazon. Finally, we performed experiments on two movie rating datasets (from Flixster and Netflix) to demonstrate that our framework also applies to recommender systems.

We believe that there are a number of future directions for further studies. A user's rating might be influenced by other users' opinions over a product, especially friends' opinions. Extending our model to capture these dependencies makes this research problem more realistic, and one may obtain more interesting results. Specifically, we plan to explore the following questions. How are users' opinions influenced by others? Do most users heavily rely on others' opinions? Can we reveal the true quality?

APPENDIX

THEOREM 1 ([MATOUSŠEK AND VONDRÁK 2001]). *Let $X_1, X_2, \ldots, X_n$ denote n independent random variables each attaining values in $[0, 1]$. Let $X = X_1 + X_2 + \cdots + X_n$ and let $\sigma^2 = Var[X] = \sum_{i=1}^{n} Var[X_i]$. Then for any $t \geq 0$, the following conditions hold* $\Pr[X \geq E[X] + t] < \exp(-\frac{t^2}{2(\sigma^2 + t/3)})$, $\Pr[X \leq E[X] - t] < \exp(-\frac{t^2}{2(\sigma^2 + t/3)})$.

THEOREM 2 ([MATOUSŠEK AND VONDRÁK 2001; FELLER 1943]). *Let X be a sum of independent random variables each attaining values in $[0, 1]$ and let $\sigma = \sqrt{Var[X]} \geq 200$. For all $t \in [0, \frac{\sigma^2}{100}]$, $\Pr[X \geq E[X] + t] \geq c \exp(-\frac{t^2}{3\sigma^2})$ holds for a suitable constant $c > 0$.*

COROLLARY 3 (ANTICONCENTRATION). *Let X be a sum of independent random variables each attaining values in [0, 1] and let $\sigma = \sqrt{Var[X]} \geq 200$. For all $t \in [0, \frac{\sigma^2}{100}]$, $\Pr[X \leq E[X] - t] \geq c \exp(-\frac{t^2}{3\sigma^2})$ holds for a suitable constant $c > 0$.*

A.1. Proof of Lemma 3.1

Let $\boldsymbol{\rho}' = (\rho'_{ij,1}, \ldots, \rho'_{ij,m})$ denote the rating distribution representing the rating behavior that results in $r^+_{i,j}$. Recall that the $\boldsymbol{\rho}'$ is a sample randomly drawn from $Dir(\boldsymbol{\alpha}_i)$; given $\boldsymbol{\rho}'$, the conditional pmf of $r^+_{i,j}$ is specified by $\Pr[r^+_{i,j} = k|\boldsymbol{\rho}'] = \rho'_{ij,k}$. It follows that $\Pr[r^+_{i,j} = k] = \int \Pr[\boldsymbol{\rho}'] \Pr[r^+_{i,j} = k|\boldsymbol{\rho}']d\boldsymbol{\rho}' = \int \Pr[\boldsymbol{\rho}']\rho'_{ij,k}d\boldsymbol{\rho}' = \alpha_{i,k}$, where the last step follows a basic property of Dirichlet distributions [Bishop 2006]. □

A.2. Proof of Theorem 3.4

We prove this theorem by applying Theorem 1 to show that $\Pr[\widehat{\ell_i} \neq \ell_i] \leq \delta$. By some basic probability arguments, we have

$$\Pr\left[\widehat{\ell_i} \neq \ell_i\right] = \Pr\left[\bigcup\nolimits_{k \neq \ell_i} \left\{\widehat{\ell_i} = k\right\}\right] \leq \sum_{k \neq \ell_i} \Pr\left[\widehat{\ell_i} = k\right] \leq \sum_{k \neq \ell_i} \Pr[n_{i,k} \geq n_{i,\ell_i}].$$

Let R^k_{ij}, where $j = 1, \ldots, n_i$ and $k = 1, \ldots, m$ denote a set of random variables with

$$R^k_{ij} = \begin{cases} 1, & \text{with probability } \Pr[r^+_{i,j} = k], \\ 0, & \text{with probability } \Pr[r^+_{i,j} = \ell_i], \\ 1/2, & \text{otherwise,} \end{cases}$$

where the pmf of $r^+_{i,j}$ is derived in Lemma 3.1. Let $R^k_i = \sum_j R^k_{ij}$. Observe that $R^k_i = n_{i,k} + (n_i - n_{i,k} - n_{i,\ell_i})/2$. It follows that $n_{i,k} \geq n_{i,\ell_i}$ if and only if R^k_i is larger than or equal to $n_i/2$, or mathematically,

$$R^k_i \geq \frac{n_i}{2} \Leftrightarrow n_{i,k} + \frac{n_i - n_{i,k} - n_{i,\ell_i}}{2} \geq \frac{n_i}{2} \Leftrightarrow n_{i,k} \geq n_{i,\ell_i}.$$

Thus, we have $\Pr[n_{i,k} \geq n_{i,\ell_i}] = \Pr[R^k_i \geq n_i/2]$. We finish this proof by showing that $\Pr[R^k_i \geq n_i/2] \leq \delta/(m-1), \forall k \neq \ell_i$ because this inequality implies that $\Pr[\widehat{\ell_i} \neq \ell_i] \leq \sum_{k \neq \ell_i} \Pr[n_{i,k} > n_{i,\ell_i}] = \sum_{k \neq \ell_i} \Pr[R^k_i \geq n_i/2] \leq \delta$. We express the expectation and variance of R^k_i as $E[R^k_i] = n_i/2 - (\alpha_{i,\ell_i} - \alpha_{i,k})n_i/2$, $\text{Var}[R^k_i] = (\alpha_{i,\ell_i} + \alpha_{i,k} - (\alpha_{i,\ell_i} - \alpha_{i,k})^2)n_i/4$. Let

$t = (\alpha_{i,\ell_i} - \alpha_{i,k})n_i/2$. Then by applying Theorem 1, we have

$$\begin{aligned}\Pr\left[R_i^k \geq \frac{n_i}{2}\right] &= \Pr\left[R_i^k \geq E\left[R_i^k\right] + t\right] \leq \exp\left(-\frac{t^2}{2(\mathrm{Var}\left[R_i^k\right] + t/3)}\right) \\ &= \exp\left(-n_i \cdot \left(\frac{2(\alpha_{i,\ell_i} + \alpha_{i,k})}{(\alpha_{i,\ell_i} - \alpha_{i,k})^2} - 2 + \frac{4}{3(\alpha_{i,\ell_i} - \alpha_{i,k})}\right)^{-1}\right) \\ &\leq \exp\left(-n_i \cdot \left(\frac{2(\alpha_{i,\ell_i} + \widetilde{\alpha}_i)}{(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^2} - 2 + \frac{4}{3(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)}\right)^{-1}\right) \leq \frac{\delta}{m-1}.\end{aligned}$$

This proof is then complete. □

A.3. Proof of Theorem 3.5

We prove this theorem by applying Theorem 2 to show that $\Pr[\widehat{\ell} \neq \ell_i] \geq \Omega(\delta)$. Let $\widetilde{\ell}_i \in \arg_k\{\alpha_{i,k} = \widetilde{\alpha}_i\}$ denote the index that $\alpha_{i,\widetilde{\ell}_i}$ attains the value $\widetilde{\alpha}_i$. Let $R_i = \sum_j R_{i,j}^{\widetilde{\ell}_i}$, where $R_{i,j}^{\widetilde{\ell}_i}$ is defined in the proof of Theorem 3.4. We express the expectation and variance of R_i as $E[R_i] = n_i/2 - (\alpha_{i,\ell_i} - \widetilde{\alpha}_i)n_i/2$, $\mathrm{Var}[R_i] = (\alpha_{i,\ell_i} + \widetilde{\alpha}_i - (\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^2)n_i/4$. We derive an upper bound of $\alpha_{i,\ell_i} - \widetilde{\alpha}_i$ as

$$\frac{100}{101}\alpha_{i,\ell_i} \leq \widetilde{\alpha}_i \Leftrightarrow \alpha_{i,\ell_i} - \widetilde{\alpha}_i \leq \frac{1}{201}(\alpha_{i,\ell_i} + \widetilde{\alpha}_i). \tag{9}$$

With this upper bound, we derive a lower bound of $\alpha_{i,\ell_i} + \widetilde{\alpha}_i - (\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^2$ as $\alpha_{i,\ell_i} + \widetilde{\alpha}_i - (\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^2 \geq \alpha_{i,\ell_i} + \widetilde{\alpha}_i - (\frac{\alpha_{i,\ell_i} + \widetilde{\alpha}_i}{201})^2 \geq \frac{40400}{201^2}(\alpha_{i,\ell_i} + \widetilde{\alpha}_i) \geq 1.99\alpha_{i,\ell_i}$. It follows that $\mathrm{Var}[R_i] \geq 1.99\alpha_{i,\ell_i}n_i \geq \frac{1.99n_i}{m}$.

We first consider the case in which $\mathrm{Var}[R_i] \leq \frac{1.99}{m}\ln\frac{1}{\delta}$. We seek to show that $\Pr[\widehat{\ell} \neq \ell_i] \geq \delta$. Observe that $n_{i,\ell_i} = 0$ implies that the $\widehat{\ell}_i \neq \ell_i$. This fact gives a lower bound of $\Pr[\widehat{\ell} \neq \ell_i] \geq (1-\alpha_{i,\ell_i})^{n_i}$. Since $\frac{100}{101}\alpha_{i,\ell_i} \leq \widetilde{\alpha}_i$ and $\alpha_{i,\ell_i} + \widetilde{\alpha}_i \leq 1$, then an upper bound of α_{i,ℓ_i} is given by $\alpha_{i,\ell_i} \leq \frac{101}{201}$. Since $\mathrm{Var}[R_i] \geq \frac{1.99n_i}{m}$ and $\mathrm{Var}[R_i] \leq \frac{1.99}{m}\ln\frac{1}{\delta}$, then an upper bound of n_i is given by $n_i \leq \ln\frac{1}{\delta}$. It follows that $\Pr[\widehat{\ell} \neq \ell_i] \geq (1-\alpha_{i,\ell_i})^{n_i} \geq (1-\frac{101}{201})^{\ln\frac{1}{\delta}} = \delta^{\ln\frac{201}{100}} \geq \delta$.

We now consider the case that $\mathrm{Var}[R_i] > \frac{1.99}{m}\ln\frac{1}{\delta}$. We apply Theorem 2 to finish this proof. With some basic probability arguments, we have $\Pr[\widehat{\ell}_i \neq \ell_i] \geq \Pr[n_{i,\widetilde{\ell}_i} > n_{i,\ell_i}] = \Pr[R_i > n_i/2]$. With a similar derivation as that of Theorem 3.4, we have $\Pr[R_i > n_i/2] = \Pr[R_i > E[R_i] + (\alpha_{i,\ell_i} - \widetilde{\alpha}_i)n_i/2]$. Let us choose $t = n_i(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)/2 + \epsilon$, where $0 < \epsilon < 10^{-10}$. Observe that $E[R_i] + t = n_i/2 + \epsilon > n_i/2$, and thus $\Pr[R_i > n_i/2] \geq \Pr[R_i \geq E[R_i] + t]$. We now check whether the conditions specified in Theorem 2 are satisfied. We first show that $\sqrt{\mathrm{Var}[R_i]} \geq 200$ holds. Choosing $\eta = e^{-20101m}$, then for any $\delta \leq \eta$, it holds that $\mathrm{Var}[R_i] \geq \frac{1.99}{m}\ln\frac{1}{\delta} \geq \frac{1.99}{m}20101m \geq 40,000$. We now show that $t \in [0, \mathrm{Var}[R_i]/100]$ holds. By applying Inequality (9), we have

$$\begin{aligned}\frac{t}{\mathrm{Var}[R_i]} &= \frac{2(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)}{\alpha_{i,\ell_i} + \widetilde{\alpha}_i - (\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^2} + \frac{\epsilon}{\mathrm{Var}[R_i]} \leq 2\left(\frac{\alpha_{i,\ell_i} + \widetilde{\alpha}_i}{\alpha_{i,\ell_i} - \widetilde{\alpha}_i} - (\alpha_{i,\ell_i} - \widetilde{\alpha}_i)\right)^{-1} + 10^{-14} \\ &\leq 2\left(201 - (\alpha_{i,\ell_i} + \widetilde{\alpha}_i)/201\right)^{-1} + 10^{-14} \leq 1/100.\end{aligned}$$

Choosing $n_i \leq \alpha_{i,\ell_i}(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^{-2} \ln \frac{1}{\delta}$ and applying Theorem 2, we have

$$\begin{aligned}\Pr[R_i > n_i/2] &\geq \Pr[R_i \geq E[R_i] + t] \geq c \exp\big(-t^2/(3\mathrm{Var}[R_i])\big) \\ &= c \exp\left(-\frac{1}{3}\frac{n_i(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^2}{\alpha_{i,\ell_i} + \widetilde{\alpha}_i - (\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^2}\right) \geq c \exp\left(-\frac{1}{3}\frac{n_i(\alpha_{i,\ell_i} - \widetilde{\alpha}_i)^2}{1.99\alpha_{i,\ell_i}}\right) \geq c\delta.\end{aligned}$$

This proof is then complete. □

A.4. Proof of Theorem 3.7

We prove this theorem by extending Theorem 3.4. We first present a probabilistic interpretation of the rating process under a fraction of f_r random misbehaving users: with probability f_r, a random misbehaving user is generated to rate P_i, and with probability $1 - f_r$, an honest user is generated to rate P_i. Then we have

$$\Pr[P_i \text{ receives a rating } k] = f_r/m + (1 - f_r)\alpha_{i,k}, \quad \text{for } k = 1, \ldots, m.$$

It follows that $f_r/m + (1 - f_r)\alpha_{i,\ell_i}$ and $f_r/m + (1 - f_r)\widetilde{\alpha}_i$ are the largest and second largest among $\Pr[P_i \text{ receives a rating } 1], \ldots, \Pr[P_i \text{ receives a rating } m]$, respectively. Plugging the two terms into Theorem 3.4, we finish this proof. □

A.5. Proof of Theorem 3.9

Similar to the proof of Theorem 3.7, we present a probabilistic interpretation of the rating process under a fraction of f_b biased misbehaving users: with probability f_b, a biased misbehaving user is generated to rate P_i, and with probability $1 - f_r$, an honest user is generated to rate P_i. Then we have

$$\Pr[P_i \text{ receives a rating } k] = \begin{cases} f_b + (1 - f_b)\alpha_{i,\ell'}, & \text{for } k = \ell' \\ (1 - f_b)\alpha_{i,k}, & \text{otherwise.} \end{cases}$$

The biased misbehavior can be tolerated if and only if the rating ℓ_i still has the largest probability mass—that is, $(1 - f_b)\alpha_{i,\ell_i} > f_b + (1 - f_b)\alpha_{i,\ell'}$. Namely, we need $f_b < (\alpha_{i,\ell_i} - \alpha_{i,\ell'})/(1 + \alpha_{i,\ell_i} - \alpha_{i,\ell'})$. Thus, $\overline{f}_b = (\alpha_{i,\ell_i} - \alpha_{i,\ell'})/(1 + \alpha_{i,\ell_i} - \alpha_{i,\ell'})$. Observe that $(1 - f_b)\alpha_{i,\ell_i}$ and $\max\{f_b + (1 - f_b)\alpha_{i,\ell'}, (1 - f_b)\widetilde{\alpha}_i\}$ are the largest and second largest value among $\Pr[P_i \text{ receives a rating } 1], \ldots, \Pr[P_i \text{ receives a rating } m]$, respectively. Plugging the two terms into Theorem 3.4, we finish this proof. □

A.6. Proof of Theorem 3.11

We prove this theorem by applying Theorem 1 to show $\Pr[\widehat{l}_i \neq l_i] \leq \delta$. With some basic probability arguments, we have $\Pr[\widehat{l}_i \neq l_i] = \Pr[|\{r_{i,j} | r_{i,j} \leq l_i\}|/n_i < 1/2] + \Pr[|\{r_{i,j} | 1 \leq r_{i,j} < l_i\}|/n_i > 1/2]$. We show $\Pr[|\{r_{i,j} | r_{i,j} \leq l_i\}|/n_i < 1/2] \leq \delta/2$ first. Let $R_{i,j}$, denote $j \in \{1, \ldots, n_i\}$, and denote a set of independent random variables with $R_{i,j} = 1$ if $r^+_{i,j} \leq l_i$; otherwise, $R_{i,j} = 0$. Let $R_i = \sum_j R_{i,j}$. We set its expectation and variance as $E[R_i] = n_i F_i(l_i)$ and $\mathrm{Var}[R_i] = n_i F_i(l_i)(1 - F_i(l_i))$. Observe that $R_i = |\{r^+_{i,j} | r^+_{i,j} \leq l_i\}| = |\{r_{i,j} | r_{i,j} \leq l_i\}|$. Based on this fact, we have $\Pr[|\{r_{i,j} | r_{i,j} \leq l_i\}|/n_i < 1/2] = \Pr[R_i/n_i < 1/2] = \Pr[R_i - n_i F(l_i) < -n_i(F(l_i) - 1/2)] = \Pr[R_i - E[R_i] < -n_i(F(l_i) - 1/2)]$. Observe that $(\alpha_{i,l_i} - |2F_i(l_i) - \alpha_{i,l_i} - 1|)/2 = (F_i(l_i) - F_i(l_i - 1) - |F_i(l_i) + F_i(l_i - 1) - 1|)/2 = \min\{F_i(l_i) - 0.5, 0.5 - F_i(l_i - 1)\} \leq F_i(l_i) - 0.5$. By applying Theorem 1 and with a similar derivation as that of Theorem 3.4, we obtain that $\Pr[|\{r_{i,j} | r_{i,j} \leq l_i\}|/n_i < \frac{1}{2}] \leq \delta/2$. Similarly, we have $\Pr[|\{r_{i,j} | 1 \leq r_{i,j} < l_i\}|/n_i > \frac{1}{2}] \leq \delta/2$. Hence, $\Pr[\widehat{l}_i \neq l_i] \leq \delta$. □

A.7. Proof of Theorem 3.12

We prove this theorem by showing that $\Pr[\widehat{l}_i \neq l_i] \geq \Omega(\delta)$. We consider two cases.

Case 1. $F_i(l_i) - \frac{1}{2} \leq \frac{1}{2} - F_i(l_i - 1)$. It follows that $\alpha_{i,l_i} - |2F_i(l_i) - \alpha_{i,l_i} - 1| = 2F_i(l_i) - 1$. We adopt the notations $R_{i,j}$ and R_i defined in the proof of Theorem 3.11. Recall that $\text{Var}[R_i] = n_i F_i(l_i)(1 - F_i(l_i))$. Let us choose $\eta_2 = (\sqrt{10001} - 99)/2$. Recall that $F_i(l_i) \in [0.5, \eta_2]$. It follows that $\text{Var}[R_i] > 0.24n_i$. We first consider the case $\text{Var}[R_i] \leq 0.24 \ln \frac{1}{\delta}$. Observe that all ratings of P_i being larger than l_i implies that $\widehat{l_i} \neq l_i$. This fact gives $\Pr[\widehat{l_i} \neq l_i] \geq \Pr[r_{i,j}^+ > l_i, \forall j] = (F_i(l_i))^{n_i} \geq (1 - (\sqrt{10001} - 99)/2)^{n_i} \geq 1/e^{n_i} \geq \delta$, where the last step follows that $\text{Var}[R_i] > 0.24n_i$ and $\text{Var}[R_i] \leq 0.24 \ln \frac{1}{\delta}$, implying that $n_i \leq \ln \frac{1}{\delta}$. We now consider the case $\text{Var}[R_i] > 0.24 \ln \frac{1}{\delta}$. By applying Corollary 3, and with a similar derivation as that of Theorem 3.5, we conclude that $\Pr[\widehat{l_i} \neq l_i] = \Omega(\delta)$.

Case 2. $F_i(l_i) - \frac{1}{2} \geq \frac{1}{2} - F_i(l_i - 1)$. Let us choose $\eta_3 = (101 - \sqrt{10001})/2$. Recall that $F_i(l_i - 1) \in [0, \eta_3]$. Then, by applying Theorem 10.2 and with a similar derivation as that of the proof of Case 1, we obtain $\Pr[\widehat{l_i} \neq l_i] = \Omega(\delta)$. The proof of this theorem is then complete. □

A.8. Proof of Theorem 3.16

We prove this theorem by applying Theorem .1 to derive an upper bound of $\Pr[|\widehat{\gamma_i} - \gamma_i| > E_r]$. Let $X_{i,j} = r_{i,j}^+/m$, $j = 1, \ldots, n_i$ denote a a set of independent random variables. We express the pmf of $X_{i,j}$ as $\Pr[X_{i,j} = k/m] = \alpha_{i,k}$ for all $k = 1, \ldots, m$ and for all $j = 1, \ldots, n_i$. Note that $X_{i,j} \in [0, 1]$. Let $X_i = \sum_j X_{i,j}$. Observe that $\widehat{\gamma_i} = \sum_j r_{i,j}^+/n_i = mX_i/n_i$ and $\gamma_i = \sum_k k\alpha_{i,k} = mE[X_i]/n_i$. Then we have $\Pr[|\widehat{\gamma_i} - \gamma_i| > E_r] = \Pr[X_i > E[X_i] + \frac{n_i}{m}E_r] + \Pr[X_i < E[X_i] - \frac{n_i}{m}E_r]$. Then by applying Theorem 1 and with a similar derivation as that of Theorem 3.4, we conclude this theorem. □

A.9. Proof of Theorem 3.17

Let $X_i = \sum_j r_{i,j}^+/m$ denote the same random variable as specified in the proof of Theorem 3.16. We express the expectation and variance of X_i as $E[X_i] = \gamma_i n_i/m$ and $\text{Var}[X_i] = (\sum_k k^2\alpha_{i,k} - \gamma_i^2)n_i/m^2$. Note that $\sum_k k^2\alpha_{i,k} - \gamma_i^2 \geq \frac{m^2}{100}$, thus $\text{Var}[X_i] \geq n_i/100$.

We first explore the case of $\text{Var}[X_i] \leq \frac{1}{100}\log_{202}\frac{2}{\delta}$. We seek to show that $\Pr[|\widehat{\gamma_i} - \gamma_i| > E_r] \geq \delta$. We divide the rating into three disjoint groups based on their distance to γ_i—that is, $\mathcal{G}_1 = \{k|k < \gamma_i - E_r\}$, $\mathcal{G}_2 = \{k||k - \gamma_i| \leq E_r\}$ and $\mathcal{G}_3 = \{k|k > \gamma_i + E_r\}$. Observe that $\sum_k k^2\alpha_{i,k} - \gamma_i^2 = \sum_k (k - \gamma_i)^2\alpha_{i,k} \leq \sum_{k\in\mathcal{G}_2}\alpha_{i,k}E_r^2 + \sum_{k\in\mathcal{G}_1\cup\mathcal{G}_3}\alpha_{i,k}m^2$. Let us choose $\eta_1 = (\sum_k k^2\alpha_{i,k} - \gamma_i^2)/100m$. It is easy to prove that $\sum_k k^2\alpha_{i,k} - \gamma_i^2 \leq (m-1)^2/4$, then for any $E_r \in [0, \eta_1]$, we have $E_r \leq \frac{1}{100m}\cdot\frac{(m-1)^2}{4} \leq \frac{m}{400}$. Note that $\sum_k k^2\alpha_{i,k} - \gamma_i^2 \geq \frac{m^2}{100}$; it follows that $\frac{m^2}{100} \leq \sum_{k\in\mathcal{G}_2}\alpha_{i,k}(\frac{m}{400})^2 + \sum_{k\in\mathcal{G}_1\cup\mathcal{G}_3}\alpha_{i,k}m^2$. Observe that $\sum_{k\in\mathcal{G}_1\cup\mathcal{G}_3}\alpha_{i,k} = 1 - \sum_{k\in\mathcal{G}_2}\alpha_{i,k}$; we then have $\sum_{k\in\mathcal{G}_2}\alpha_{i,k} \leq \frac{m^2 - m^2/100}{m^2 - (m/400)^2} \leq \frac{100}{101}$. Note that all ratings of P_i being in $\mathcal{G}_1$ implies that $\widehat{\gamma_i} < \gamma_i - E_r$, and all of these ratings being in $\mathcal{G}_3$ implies that $\widehat{\gamma_i} > \gamma_i + E_r$. These facts give the following lower bound:

$$\begin{aligned}\Pr[|\widehat{\gamma_i} - \gamma_i| > E_r] &\geq \Pr\left[r_{i,j}^+ \in \mathcal{G}_1, \forall j\right] + \Pr\left[r_{i,j}^+ \in \mathcal{G}_3, \forall j\right] \\ &= \left(\sum_{k\in\mathcal{G}_1}\alpha_{i,k}\right)^{n_i} + \left(\sum_{k\in\mathcal{G}_3}\alpha_{i,k}\right)^{n_i} \geq 2\left(\sum_{k\in\mathcal{G}_1\cup\mathcal{G}_3}\alpha_{i,k}/2\right)^{n_i} \geq 2\cdot 202^{-n_i} \geq \delta,\end{aligned}$$

where the last step follows $\text{Var}[X_i] \geq \frac{n_i}{100}$ and $\text{Var}[X_i] \leq \frac{1}{100}\log_{202}\frac{2}{\delta}$, implying that $n_i \leq \log_{202}\frac{2}{\delta}$.

We now explore the case $\text{Var}[X_i] \geq \frac{1}{100}\log_{202}\frac{2}{\delta}$. By applying Theorem 2, and with a similar derivation as that of Theorem 3.5, we conclude this theorem. □

A.10. Proof of Lemma 3.21

With a similar derivation as that of Theorem 3.7, we have that for all $k = 1, \ldots, m$, it holds that $\Pr[P_i$ receives a rating $k|$assigned by misbehaving users$] = \frac{f_r}{f_r+f_b}\frac{1}{m} + \frac{\mathbf{I}_{\{k=\ell'\}}f_b}{f_r+f_b}$. It follows that the majority rating of the ratings assigned by misbehaving users converges to ℓ'. Since $\ell' \neq \ell_i$, we therefore conclude the result for the majority rule. Consider the median rule; the median of the ratings assigned by misbehaving users converges to l_i if and only if $\frac{l_i}{m}\frac{f_r}{f_r+f_b} + \frac{\mathbf{I}_{\{\ell'<l_i\}}f_b}{f_r+f_b} > \frac{1}{2}$ and $\frac{l_i-1}{m}\frac{f_r}{f_r+f_b} + \frac{\mathbf{I}_{\{\ell'<l_i\}}f_b}{f_r+f_b} < \frac{1}{2}$. Namely, we need $(\frac{l_i}{m} - \frac{1}{2})f_r > (\frac{1}{2} - \mathbf{I}_{\{\ell'<l_i\}})f_b > (\frac{l_i-1}{m} - \frac{1}{2})f_r$. We now consider the average score rule. The mean of the ratings assigned by misbehaving users is $\frac{f_r}{f_r+f_b}\frac{m+1}{2} + \frac{\ell' f_b}{f_r+f_b}$. Thus, to guarantee $\frac{f_r}{f_r+f_b}\frac{m+1}{2} + \frac{\ell' f_b}{f_r+f_b} = \gamma_i$, we only need $\frac{f_r}{f_b} = \frac{\gamma_i-\ell'}{(m+1)/2-\gamma_i}$.

REFERENCES

ABC7 News. 2012. Woman Paid to Post Five-Star Google Feedback. Retrieved April 16, 2015, from http://www.thedenverchannel.com/news/woman-paid-to-post-five-st ar-google-feedback.

Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17, 6, 734–749.

BBCNews. 2013. Samsung Probed in Taiwan over "Fake Web Reviews." Retrieved April 16, 2015, from http://www.bbc.com/news/technology-22166606.

Smriti Bhagat, Amit Goyal, and Laks V. S. Lakshmanan. 2012. Maximizing product adoption in social networks. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*. 603–612.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

Ludovico Boratto and Salvatore Carta. 2011. State-of-the-art in group recommendation and new approaches for automatic identification of groups. In *Information Retrieval and Mining in Distributed Environments*. Studies in Computational Intelligence, Vol. 324. Springer, 1–20.

Mao Chen and Jaswinder Pal Singh. 2001. Computing and using reputations for Internet ratings. In *Proceedings of the 3rd ACM Conference on Electronic Commerce*. 154–162.

Freddy C. T. Chua, Hady W. Lauw, and Ee-Peng Lim. 2013. Generative models for item adoptions using social correlation. *IEEE Transactions on Knowledge and Data Engineering* 25, 9, 2036–2048.

Freddy Chong Tat Chua, Hady Wirawan Lauw, and Ee-Peng Lim. 2011. Predicting item adoption using social correlation. In *Proceedings of the 2011 SIAM International Conference on Data Mining*. 367–378.

Chrysanthos Dellarocas. 2000. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proceedings of the 2nd ACM Conference on Electronic Commerce*. 150–157.

William Feller. 1943. Generalization of a probability limit theorem of Cramer. *Transactions of the American Mathematical Society* 54, 3, 361–372.

Miriam Fernández, David Vallet, and Pablo Castells. 2006. Probabilistic score normalization for rank aggregation. In *Proceedings of the 28th European Conference on Advances in Information Retrieval*. 553–556.

Amit Goyal, Francesco Bonchi, and Laks V. S. Lakshmanan. 2010. Learning influence probabilities in social networks. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. 241–250.

Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems* 22, 1, 5–53.

Will Hill, Larry Stead, Mark Rosenstein, and George Furnas. 1995. Recommending and evaluating choices in a virtual community of use. In *Proceedings of the 1st SIGCHI Conference on Human Factors in Computing Systems*. 194–201.

Mohsen Jamali and Martin Ester. 2010. A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the 4th ACM Conference on Recommender Systems*. 135–142.

Rong Jin and Luo Si. 2004. A study of methods for normalizing user ratings in collaborative filtering. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 568–569.

Rong Jin, Luo Si, ChengXiang Zhai, and Jamie Callan. 2003. Collaborative filtering with decoupled models for preferences and ratings. In *Proceedings of the 12th International Conference on Information and Knowledge Management*. 309–316.

Nitin Jindal and Bing Liu. 2007. Review spam detection. In *Proceedings of the 16th International Conference on World Wide Web*. 1189–1190.

Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. 219–230.

David Kempe, Jon Kleinberg, and Éva Tardos. 2003. Maximizing the spread of influence through a social network. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 137–146.

Shyong K. Lam and John Riedl. 2004. Shilling recommender systems for fun and profit. In *Proceedings of the 13th International Conference on World Wide Web*. 393–402.

Hady Wirawan Lauw, Ee-Peng Lim, and Ke Wang. 2008. Bias and controversy in evaluation systems. *IEEE Transactions on Knowledge and Data Engineering* 20, 11, 1490–1504.

Hady Wirawan Lauw, Ee-Peng Lim, and Ke Wang. 2012. Quality and leniency in online collaborative rating systems. *ACM Transactions on the Web* 6, 1, 4.

Daniel Lemire. 2005. Scale and translation invariant collaborative filtering systems. *Information Retrieval* 8, 1, 129–150.

Jirší Matoušek and Jan Vondrák. 2001. *The Probabilistic Method*. Charles University.

Bamshad Mobasher, Robin Burke, and J. J. Sandvig. 2006. Model-based collaborative filtering as a defense against profile injection attacks. In *Proceedings of the 21st National Conference on Artificial Intelligence*. 1388–1393.

Netflix. 2009. Netflix Prize. Retrieved April 16, 2015, from http://www.netflixprize.com.

PayPerPost. 2015. PayPerPost Home Page. Retrieve April 16, 2015, from https://payperpost.com.

Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 5th ACM Conference on Computer Supported Cooperative Work*. 175–186.

Paul Resnick and Hal R. Varian. 1997. Recommender systems. *Communications of the ACM* 40, 3, 56–58.

Tracy Riggs and Robert Wilensky. 2001. An algorithm for automated rating of reviewers. In *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries*. 381–387.

Badrul M. Sarwar, George Karypis, Joseph A. Konstan, and John T. Riedl. 2000. Application of dimensionality reduction in recommender system—a case study. In *Proceedings of the Web Mining for E-Commerce—Challenges and Opportunities Workshop*.

New York Times. 2004. Amazon Glitch Unmasks War of Reviewers. Retrieved April 16, 2015, from http://www.nytimes.com/2004/02/14/us/amazon-glitch-unmasks-war-of-reviewers.html.

Wall Street Journal 2009. A Fake Amazon Reviewer Confesses. Retrieved April 16, 2015, from http://blogs.wsj.com/wallet/2009/07/09/delonghis-strange-brew-tracking-down-fake-amazon-raves/.

Jonathan Traupman and Robert Wilensky. 2006. Collaborative quality filtering: Establishing consensus or recovering ground truth? In *Advances in Web Mining and Web Usage Analysis*. Springer, 73–86.

Benjamin Van Roy and Xiang Yan. 2010. Manipulation robustness of collaborative filtering. *Management Science* 56, 11, 1911–1929.

Jaewon Yang and Jure Leskovec. 2010. Modeling information diffusion in implicit networks. In *Proceedings of the 10th IEEE International Conference on Data Mining*. 599–608.

Jie Zhang and Robin Cohen. 2006. Trusting advice from other buyers in e-marketplaces: The problem of unfair ratings. In *Proceedings of the 8th International Conference on Electronic Commerce*. 225–234.

Received September 2013; revised July 2014; accepted November 2014