Aggregating Reputation Feedback

Florent Garcin[1], Boi Faltings[1], and Radu Jurca[2]

[1] Ecole Polytechnique Fédérale de Lausanne (EPFL),
Artificial Intelligence Laboratory
{florent.garcin,boi.faltings}@epfl.ch
[2] Google Inc.
Switzerland
radu.jurca@gmail.com

Abstract. A fundamental task in reputation systems is to aggregate multiple feedback ratings into a single value that can be used to compare the reputation of different entities. Feedback is most commonly aggregated using the arithmetic mean. However, the mean is quite susceptible to outliers and biases, and thus may not be the most informative aggregate of the reports. We consider three criteria to assess the quality of an aggregator: the informativeness, the robustness and the strategyproofness, and analyze how different aggregators, in particular the mean, weighted mean, median and mode, perform with respect to these criteria. The results show that the arithmetic mean may not always be the best choice.

1 Introduction

Many sites on the world wide web offer people the possibility to share their experiences with products and services through reviews and ratings. This feedback helps people avoid bad choices, drives them towards more useful products, and brings more revenue to good producers. They are an important part of the user's decision making when buying goods or services.

We consider in particular reputation and rating systems for products and services, such as those operated by Amazon.com, Tripadvisor, and many other electronic commerce sites. These have the following characteristics:

- they collect ratings for individual well-identified products or services, and aggregate these ratings into a single score;
- the identity of raters does not have to be known;
- raters act to influence the score of the item they rate to make it as close as possible to the value they consider best. Note that this value might not reflect the true quality if the rater is not honest.

A common reflex for users of such sites is to order the choices according to their ratings, and only consider those that are at the top of such rankings. Such an order is usually obtained by aggregating individual feedback scores into a single value that establishes an ordering of the alternatives. The most common

way of aggregating ratings is by the arithmetic mean, but one can also consider aggregation using the weighted mean, the median and the mode.

In unbiased normal distributions, there is little difference between these aggregators. However, it is known that in reality, reviews are often biased [7]. Writing a review or even just leaving a rating requires effort, and since it is voluntary many of these ratings are left by people who have some ulterior motive or extreme opinion. One can thus observe that the distribution of ratings is far from the normal distribution one would expect from an unbiased population of raters. This means that the different ways of aggregating them can give very different results.

In this paper, we consider how the aggregation method influences the quality of the ranking. We evaluate quality using the following three criteria:

- informativeness, i.e. how likely is it that the ranking that a user finds at the time of making a choice will still be the ranking when the user is using the product or service;
- robustness, i.e. how easy is it for the ranking to be distorted by outliers or malicious reviews;
- strategyproofness, i.e. for a rater who wants the average ranking to be a certain value, is it best to report this value or manipulate the aggregation by reporting differently.

We compare four different ways of aggregating n numerical ratings $r_1, ..., r_n$, using different forms of averaging:

- the *mean* is the arithmetic mean $\overline{r_a} = \frac{1}{n}\sum_{i=1}^{n} r_i$.
- the *weighted mean* is the same as the arithmetic mean but with weights: $\overline{r_w} = \frac{\sum_{i=1}^{n} w(i)r_i}{\sum_{i=1}^{n} w(i)}$, where $w(i)$ is the weight function.
- the *median* is the smallest value $\overline{r_d}$ such that half of the values are $\geq \overline{r_d}$ and half of the values are $\leq \overline{r_d}$, i.e. there exist $X \subset \{r_1, .., r_n\}$ and $Y \subset \{r_1, .., r_n\}$ such that $(\forall r_i \in X) r_i \leq \overline{r_d}$ and $(\forall r_i \in Y) r_i \geq \overline{r_d}$ and $||X| - |Y|| \leq 1$.
- the *mode* is the smallest value $\overline{r_o}$ which occurs most frequently as a rating, i.e. for any $r' \neq \overline{r_o}$, $|\{r_i | r_i = \overline{r_o}\}| \geq |\{r_i | r_i = r'\}|$.

Both the median and the mode require a tie-breaking rule. When two values are possible, we select the smallest one. Moreover, it happens that two items have the same aggregated value. In that case, we use the number of reviews as a tie-breaking rule to make the final ranking.

We examine the reviews given on an actual review site and observe that the four notions of average differ significantly. In particular, the mode and median tend to be more robust to outliers and biased reviews than the mean and the weighted mean, and thus may be more informative for a user.

In this paper, we first present an analysis of the four different notions with respect to their robustness, and show that they have very different properties. We then analyze their behavior on data taken from an actual review web site, and show that they lead to very different rankings and also very different behavior of the rankings over time. In particular, our results suggest that the mean may

not be the most informative way of aggregating ratings since the ranking shown to a user is often very unstable.

2 Related Work

There are many reputation mechanisms that follow the model we assume in this paper. They can differ significantly in the way they aggregate and display reputation information to the users. Some mechanisms accumulate all reports into a reputation score that may potentially grow forever. eBay[3] and RentACoder[4] are two commercial sites where part of the reputation information is given by scores that reflect the total number of positive or negative interactions reported for an agent.

Amazon[5] or the popular movie review database IMDB[6] rank products by the arithmetic mean of ratings. They also publish histograms of the ratings[7] but the richer information is more difficult to find through the normal user interface, and is not used in any way for ranking the alternatives.

Tripadvisor[8] ranks hotels from cities around the world. The hotels are sorted by "popularity", defined here as the arithmetic mean of ratings. The reviews for a given hotel are ordered from the most recent to the oldest. Only 10 reviews are listed per page and contain information about the author (date, username, location of the reviewer) and the overall rating with a textual comment. The user has to click on the review to see more details.

Other mechanisms use discrete values for reputation information and define clear rules describing how sets of feedback are mapped to reputation values. The popular IT news site Slashdot[9] uses *karma* levels (i.e., *terrible, bad, neutral, positive, good, and excellent*) that characterize the quality of the news submissions posted by a user so far. Likewise, eBay sellers also have labels (e.g., *power seller*) that they can gain by meeting certain conditions.

The robustness of the reputation mechanism has also been an important concern of the research community. [5] discuss the risks associated with cheap *online pseudonyms* (i.e., users can easily create several online identities) and conclude that in any reputation mechanism newcomers must start with the lowest possible reputation. This property is later used by [3] to design moral hazard reputation mechanisms that are robust to identity changes.

[2] describes general techniques for making online feedback mechanisms immune to manipulation. A theoretical study of opinion manipulation is presented in [4], with the striking conclusion that manipulation can both increase and

[3] www.ebay.com
[4] www.rentacoder.com
[5] www.amazon.com
[6] www.imdb.com
[7] In addition, IMDB correlates demographics information with the histogram of scores.
[8] www.tripadvisor.com
[9] www.slashdot.org

decrease the information value of online forums. Other works addressing the robustness of the reputation information are [12] and [1].

[10] and [8] discuss general mechanisms for making reputation mechanisms incentive compatible. The idea is to reward the agents for reporting feedback such that the expected reward is maximized by being honest. [9] extends this idea to mechanisms that are also collusion resistant.

Our work differs from the above results in several important ways. First, we are looking at typical review forums where the social network of a user is unknown, and most users submit only one review. Second, we are looking at *single value* aggregators of reputation information, that can be easily understood and used by normal users to rank alternatives. Finally, we consider actual reviews and study how different information aggregators affect key properties like robustness, informativeness and strategyproofness.

3 Empirical Study

We consider feedbacks from a popular travel site that collects reviews of hotels from users around the world. The reviews contain a textual comment with a title, an overall rating and numerical ratings from 1 (lowest) to 5 (highest) for different features such as cleanliness, service, location, etc. The site provides ranking of hotels according to their location. Like most of the reputation sites, it aggregates reviews into a single value for each hotel and, based on that value, sorts hotels in ascending order. It uses a simple arithmetic mean on the overall ratings to recommend hotels.

We selected four cities for this study: Boston, Las Vegas, New York and Sydney. For each city, we took the first 100 hotels that have the highest number of reviews. Table 1 shows for each city the number of reviews and the distribution of hotels with respect to the star-rating provided by the website. Hotels that do not have a star-rating are classified as 'NA'. All data were collected by crawling the website in July 2007.

Table 1. A summary of the data set.

City	# Reviews	# of Hotels with NA, 1, 2, 3, 4 & 5 stars
Boston	5537	17, 2, 4, 23, 15, 5
Las Vegas	28017	19, 8, 18, 31, 17, 7
New York	29123	16, 9, 12, 35, 24, 4
Sydney	3629	41, 0, 1, 29, 19, 10

4 Robustness

In this section, we present an analysis of four aggregators, namely the mean, the weighted mean, the median and the mode, inspired by the robust statistics

theory. Robust statistics aim at analyzing and suggesting estimators that are unaffected by small deviations from the model assumptions. Interested readers may refer to [6] [14] for additional informations.

For this analysis, we quantify how robust aggregators are against outliers and malicious reports. In order to assess the quality of each aggregator, we define the *breakdown point* as a measure of this robustness. The breakdown point is the proportion of manipulated ratings required to make the aggregator return an arbitrary value.

Definition 1. *Let* $\{r_1, ..., r_{n-l}, r'_1, ..., r'_l\}$ *be a sample of* n *reviews where* r'_i *are outliers. The* finite-sample breakdown point ϵ *of an aggregator* $\overline{r}$ *is the smallest proportion* $\frac{l}{n}$ *for which the set* $\{r'_1, ..., r'_l\}$ *will cause* $\overline{r}$ *to be unbounded.*

Definition 2. *The* breakdown point ϵ^* *is the limit of the finite-sample breakdown point as* n *goes to infinity.*

This definition provides a tool to measure the robustness of every estimator. The higher the breakdown point, the more robust the estimator is. However, the breakdown point cannot exceed 0.5 because if more than half of the ratings are outliers, it is not possible anymore to distinguish the underlying distribution of the outliers. We will see in the next sections that two aggregators achieve this upper bound.

4.1 Mean and Weighted Mean

Let $\{r_1, ..., r_{n-l}, r'_1, ..., r'_l\}$ be the set of ratings for a given hotel h. r'_i are the outliers. We define the mean by

$$\overline{r_a} = \frac{1}{n}(\sum_{i=1}^{n-l} r_i + \sum_{i=1}^{l} r'_i) \tag{1}$$

One outlier is enough to change the value of the mean. Thus, the finite-sample breakdown point of the mean is $\epsilon = \frac{1}{n}$. The breakdown point is $\epsilon^* = \lim_{n\to\infty} \frac{1}{n} = 0$. The mean is extremely sensitive to outliers.

In this study, the ratings are bounded and we need more than one outlier to significantly alter the mean and thus the ranking. For that reason, we would like to quantify how many outliers are required to change the ranking of hotel h_j from position j to position i. Consider a hotel h_i with n ratings of mean $\overline{r_i}$. We add k outliers with mean $\overline{r}'$. How many outliers are needed to have a new mean lower or equal to the mean $\overline{r_j}$ of another hotel h_j? That is

$$\frac{n\overline{r_i} + k\overline{r}'}{n+k} \leq \overline{r_j} \tag{2}$$

After reordering, we get

$$k \geq \frac{n(\overline{r_i} - \overline{r_j})}{\overline{r_j} - \overline{r}'} \tag{3}$$

For instance, if we want to set the mean to 4 of a hotel with $n = 100$ ratings and mean $\overline{r} = 4.5$ by only adding lowest ratings of '1', then we need $k \geq 17$ outliers.

The weighted mean is similar to the simple mean except that ratings have assigned weights and some contribute more than others. Let $\{r'_1, ..., r'_l, r_1, ..., r_{n-l}\}$ be the set of ratings sorted from the most recent to the oldest for a given hotel h. The weighted mean is

$$\overline{r_w} = \frac{\sum_{i=1}^{l} w(i) r'_i + \sum_{i=1}^{n-l} w(i+l) r_i}{\sum_{i=1}^{n} w(i)} \tag{4}$$

The weights do not change the breakdown point and remains the same as the simple mean. Note that the mean is a special case of the weighted mean where the weights are all equal to 1. Obviously, the number of outliers needed to change the ranking of a hotel is upper bounded by Equation 3 and depends on the weight function $w(i)$.

4.2 Median

The median is the rating $\overline{r_d}$ separating the lower half from the upper half of a set of ratings. Let $\{r_1, ..., r_{n-m}, r'_1, ..., r'_m\}$ be the set of ratings sorted in ascending order for a given hotel h. r'_i are the outliers. If n is odd, that is $n = 2l + 1$, the median is located at $(l + 1)/n$. Recall that if n is even, i.e. $n = 2l$, we take the value at l.

To find the breakdown point, we determine the proportion of outliers required to change the value of the median. The finite-sample breakdown point is given by

$$\epsilon = \begin{cases} \frac{l+1}{n} = \frac{1}{2} + \frac{1}{2n}, & n = 2l + 1 \\ \frac{l}{n} = \frac{1}{2}, & n = 2l \end{cases} \tag{5}$$

Therefore, the breakdown point is $\epsilon^* = \lim_{n\to\infty} \epsilon = \frac{1}{2}$. The median is thus a robust aggregation function because it involves only the location and not the value of the ratings. To find the number of outliers required to change the ranking of a given hotel h with n ratings, we add k outliers to the ratings of h. In the worst case scenario, k should be at least equal to $n+1$. The first outlier determines the value of the median and thus the rank. For instance, if we want to change the median of a hotel that has 100 ratings, we need at least 101 malicious ratings and the new median is given by the first malicious ratings we introduce.

4.3 Mode

The mode, denoted $\overline{r_o}$, is another aggregation function and is equal to the rating that occurs the most frequently. That is, for any $r' \neq \overline{r_o}$,

$$|\{r_i | r_i = \overline{r_o}\}| \geq |\{r_i | r_i = r'\}| \tag{6}$$

Let m and l be the number of identical ratings r_1 and r_2 respectively, $m \neq l$. Obviously, if the mode is the rating r_1, then $m > l$. Therefore, $m \geq l + 1$. Thus, the finite-sample breakdown point of the mode is equal to $((n/2) + 1)/n$ for $n = m + l$ ratings. It follows that the breakdown point $\epsilon^* = \frac{1}{2}$. From the same reasoning, we need $k = n + 1$ outliers of the same value to change the mode. For instance, if a hotel as a mode $\overline{r_o} = 4$ with $n = 100$ ratings (of '4'), $k = 101$ outliers are required to change that mode.

5 Empirical Results

It is well-known that distributions of reports are far from normal due to reporting biases [7]. Aggregators such as the mean, median and mode have relatively the same value for normal distributions. However, they should have a significant difference for non-normal distributions. To support this hypothesis, we conducted the following experiment. For each of the four cities considered in our study, we computed a full ranking of the hotels according to each of the four aggregators explained in Section 1. Then, for every pair of aggregators we measured the *distance* between the corresponding orderings of hotels within a city. To measure the distance between the two rankings we chose the average absolute difference between the position of the same hotel in the two rankings.

For the weighted mean, we use Equation 7 as the weight function that is directly inspired by the logistic function applied in regression models. With this function (see Fig. 1), recent ratings have a high weight and the weight decreases while the rating is getting older. We use the following logit model for the relevance and thus the weight of a rating as a function of its order:

$$w(i) = \frac{0.9}{1 + e^{\beta(i-\mu)}} + 0.1 \tag{7}$$

Such logit models are commonly believed to be good models for probabilities that vary over time or space.

The results are presented in Table 2. For example, the rank of a hotel in Boston varies on the average with 7.7 positions (up or down) when the ranking is done according to the median instead of the mean. Likewise, the rank of a hotel in New York varies with an average of 16.9 positions (up or down) when the ranking considers the mode instead of the mean.

The average difference of ranks triggered by different aggregators is quite high: 8 to 17 ranks[10]. Considering that most feedback websites display only the first 5 or 10 "best" items, the results of Table 2 show that different aggregators can completely change the list of candidates suggested to the users. It therefore becomes important to better understand the properties of each aggregator.

[10] The only exception is the tuple *mean - weighted mean*. The two aggregators are conceptually very close, therefore the rankings span by them are also similar.

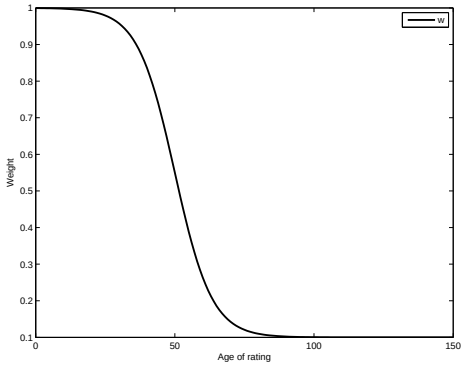

Fig. 1. The weight function: recent ratings (low index) get a highest weight. Note that the age is not related to the time but to the most recent rating.

Table 2. Average difference of ranking for the three aggregator functions.

	Boston	Las Vegas	New York	Sydney	*average*
mean - median	7.788	13.480	11.480	9.100	10.462
mean - mode	9.939	15.100	16.980	11.420	13.360
mean - weighted mean	2.394	2.760	5.480	1.340	2.993
median - mode	10.182	16.140	16.860	10.340	13.380
median - weighted mean	8.333	13.460	12.600	9.600	10.998
mode - weighted mean	10.848	15.740	17.940	11.700	14.057

5.1 Informativeness

In a reputation system, the goal of the aggregator is to reflect the user's reviews into one value. One assumption of aggregator is that users have reported their true experience. However, it is often not the case. For instance, the ratings are often part of discussion threads where past reviews influence future reports by creating prior expectations [13]. Therefore we can ask how an aggregator will continue to correctly reflect users' opinion. In Table 3, we look at the stability of each aggregator by counting the number of rankings that deviate by more than two ranks from the final ranking. The median is the most stable aggregator with two cities. However, the weighted mean seems more stable on average. The median follows closely. Then the mode and the mean come after.

As an example, Figure 2 provides the evolution of ranking and rating by the incoming reviews for a New York hotel. If we look at the mean aggregator only, when the rating decreases, the hotel loses ranks. However, around the 120th review, the rating increases and thus the hotel is going up in the global ranking. Although the median and the mode have a fixed value for the rating, the rank oscillates a little bit for the first reviews to stabilize very quickly. We observe such behavior for most of the hotels in our database.

Table 3. Average number of ranking that deviate from the final ranking with more than 2 ranks. In bold, the lowest value. The significance levels are computed with a one-way analysis of variance.

	Boston	Las Vegas	New York	Sydney
Weighted mean	29.606	**154.640**	96.660	23.450
Mean	45.833	227.120	156.930	28.460
Median	**23.652**	189.770	91.870	**12.760**
Mode	29.758	254.170	**73.550**	17.330
p-value	0.000	0.006	0.000	0.000

5.2 Robustness

Finally, we look at the robustness of each aggregator by taking the number of outliers required to alter the ranking of a given hotel. For each hotel, we inject outliers with the highest possible ratings, i.e. 5, until the rank changes. Table 4 summarizes the results for each city. Two reviews are enough to change the rank when the aggregator is the weighted mean, around 5 for the mean while the median and mode need 20 and 15 outliers respectively. The mean and weighted mean can be changed with a very low number of additional ratings. However, the mode, and in particular the median require a relatively large number of additional ratings, and are thus more difficult to manipulate.

Table 4. Average number of outliers (with highest ratings '5') required to alter the ranking. In bold, the highest value. The significance levels are computed with a one-way analysis of variance.

	Boston	Las Vegas	New York	Sydney
Weighted mean	1.922	2.153	2.155	1.464
Mean	3.328	5.102	8.041	1.948
Median	**10.297**	**40.602**	**22.639**	3.639
Mode	9.047	23.867	22.309	**3.691**
p-value	0.000	0.000	0.000	0.000

6 Strategyproofness

Besides influencing the conclusions that are drawn from a given set of ratings, the way that ratings are aggregated can also have an influence on the reports that users will submit. In this section, we consider to what degree users have an incentive to report a rating that differs from their true perception in order to manipulate the ranking.

We make the assumption that a rater has a single most preferred score that she would like to see as the aggregated score of the item being rated. For an honest rater, this value should be the true perception of quality. We furthermore

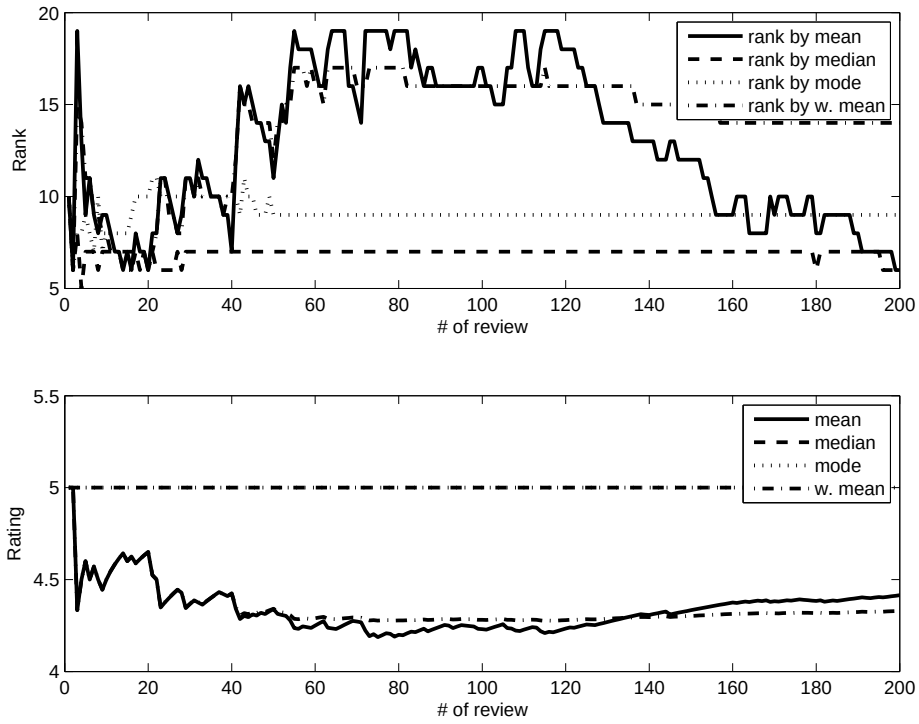

Fig. 2. A New York hotel

assume that when it is not possible to make the aggregated score take this most preferred score, the rater would like to bring it as close as possible to it. In the language of decision theory, this means that raters have a *single-peaked* preference profile: their preference for different ratings has a single peak at their most preferred score and drops monotonically to both sides of it.

Now consider the rating that such a user should report to best achieve its objective. In a *strategyproof* reputation system, a rater can expect the best possible outcome by reporting her most preferred score.[11] However, this is not always the case. For example, if a product currently has 5 reports with an arithmetic mean of 4, and the rater would like to see a score of 3, then it would be best off to report 1 and drive the mean to 3.5 rather than 3 and obtain a mean of 3.833. We believe that such manipulation strategies are the source of much of the reporting bias we can observe in practical reputation sites, and we conjecture that much more useful information could be obtained if the systems were indeed strategyproof.

Definition 3. *An aggregation function is* strategyproof *(or truthful) if there is no incentive for any of the reviewers to lie about or hide their private valuation.*

[11] Note that this does not have to be the true quality.

6.1 Mean and Weighted Mean

The mean and the weighted mean are not strategyproof. Consider that the reviewers are sorted in order of their private opinions. Let $\bar{r}$ be the mean (or weighted mean). Any reviewer a_j with a private opinion below the mean has the incentive to submit an exaggerated negative review in order to push the mean downwards. Likewise, a rater with a private opinion higher than the mean has the incentive to submit an exaggerated positive review.

6.2 Median

Moulin proves that, when preferences are single-peaked along the real line, the median is the only strategyproof preference aggregation scheme [11]. Assume that the reviewers $\{a_1, ..., a_n\}$ are sorted increasingly according to their private opinion of a hotel A. Let r_i be the private opinion of the reviewer a_i, so $r_{i+1} \geq r_i$. Let r^* denote the median rating, corresponding to reviewer a_{i^*}. Obviously, reviewer a_{i^*} should not deviate. If a reviewer a_j with $j < i^*$ misreports a lower value than r_j the median rating will not change. Misreporting a value higher than r_j, on the other hand, can only increase the median, and therefore make the public reputation of the hotel even further from a_j's private opinion. The same argument applies for any reviewer a_j with $j > i^*$. As long as the tie-breaking is independent of the reviews, then the same argument holds even if there is an even number of raters in the system.

In addition, Moulin ([11]) also shows that aggregation through the median is Pareto optimal and anonymous.

6.3 Mode

The mode is not strategyproof. Assume that two reviewers have the same private opinion r_1 and three reviewers have same private opinion $r_2 > r_1$. Let a_j be a reviewer whose private opinion is $r_j < r_1$. If the reviewer misreports and submits a review with the value r_1 she has successfully modified the public reputation of the hotel from r_2 to r_1, which is a better outcome for a_j.

7 Conclusion

We considered different ways of aggregating ratings, in particular the mean, weighted mean, median and mode. All review sites that we are aware of aggregate ratings by taking their mean, and if ratings were unbiased and normally distributed the different notions would not differ much. However, in actual review sites there are many biases, and the three methods give very different results. In hindsight, we find it surprising that other ways of forming averages have not been considered.

We considered three criteria: informativeness as reflected by the degree to which the ranking fluctuates over time, robustness as reflected by the number

of reports necessary to change the aggregate, and strategyproofness as reflected by the incentive to file truthful reports to move the average as close to them as possible.

On all three criteria, the mean seems to be the worst way of aggregating rankings: it changes the most frequently, it is the least robust, and it is not strategyproof. While the weighted mean is in general more informative, the median is significantly more robust. Finally, only the median is strategyproof. Strategyproofness may greatly increase the quality of rating information that is collected, provided that raters actually understand it. This would be an interesting subject for a user study.

We thus conclude that for using reputation sites to help users in their choices, aggregation through the median or mode are likely to be better choices than the mean. However, we recognize that if the purpose of the reputation system is to encourage good quality, i.e. to deal with the moral hazard problem, it may actually be desirable for raters to be able to move the ranking easily. The two aspects should be weighed by the designer of a reputation system.

Acknowledgments

The authors would like to thank David Parkes for pointing us to [11] and the anonymous reviewers for their contructive comments.

References

1. S. Buchegger and J.-Y. Le Boudec. The Effect of Rumour Spreading in Reputation Systems for Mobile Ad-hoc Networks. In *WiOpt '03: Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, Sophia-Antipolis, France, 2003.
2. C. Dellarocas. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *EC '00: Proceedings of the 2nd ACM conference on Electronic commerce*, pages 150–157, New York, NY, USA, 2000. ACM.
3. C. Dellarocas. Reputation Mechanism Design in Online Trading Environments with Pure Moral Hazard. *Information Systems Research*, 16(2):209–230, 2005.
4. C. Dellarocas. Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms. *Management Science*, 52(10):1577–1593, 2006.
5. E. Friedman and P. Resnick. The Social Cost of Cheap Pseudonyms. *Journal of Economics and Management Strategy*, 10(2):173–199, 2001.
6. F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, 1986.
7. N. Hu, P. Pavlou, and J. Zhang. Can Online Reviews Reveal a Product's True Quality? In *Proceedings of ACM Conference on Electronic Commerce (EC 06)*, 2006.
8. R. Jurca and B. Faltings. Minimum Payments that Reward Honest Reputation Feedback. In *Proceedings of the ACM Conference on Electronic Commerce (EC'06)*, pages 190–199, Ann Arbor, Michigan, USA, June 11–15 2006.
9. R. Jurca and B. Faltings. Collusion Resistant, Incentive Compatible Feedback Payments. In *Proceedings of the ACM Conference on Electronic Commerce (EC'07)*, pages 200–209, San Diego, USA, June 11–15 2007.

10. N. Miller, P. Resnick, and R. Zeckhauser. Eliciting Informative Feedback: The Peer-Prediction Method. *Management Science*, 51:1359 –1373, 2005.
11. H. Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35:437–455, 1980.
12. M. Srivatsa, L. Xiong, and L. Liu. TrustGuard: Countering Vulnerabilities in Reputation Management for Decentralized Networks. In *Proceedings of the World Wide Web Conference*, Japan, 2005.
13. A. Talwar, R. Jurca, and B. Faltings. Understanding user behavior in online feedback reporting. In *EC '07: Proceedings of the 8th ACM conference on Electronic commerce*, pages 134–142, New York, NY, USA, 2007. ACM.
14. R. R. Wilcox. *Introduction to Robust Estimation and Hypothesis Testing*. Elsevier Academic Press, 2nd edition edition, 2005.