How Aggregators Influence Human Rater Behavior?

FLORENT GARCIN, Swiss Federal Institute of Technology, Lausanne
LIRONG XIA, Harvard University
BOI FALTINGS, Swiss Federal Institute of Technology, Lausanne

1. INTRODUCTION

Review websites play a major role in customers' decision making, superseeding traditional word-of-mouth. The importance of online reviews is such that product companies use them as proof of quality and do not hesitate to advertise it. For instance, Tripadvisor is issuing a "certificate of excellence" (Fig. 1) based on the overall rating score achieved by a hotel. These certificates are then displayed on hotel websites and entrance doors.

In general, review websites order products according to their ratings, and user only consider those that are at the top of such rankings. Such ordering is obtained by aggregating individual ratings into a single value. By submitting a rating, users influence the overall score of the product they rate, and thus the final decision of potential customers.

Ratings are most commonly aggregated using the arithmetic mean. However, the mean is susceptible to outliers and biases [Garcin et al. 2009a], and thus may not be the most appropriate aggregator because reviews are often biased [Hu et al. 2006; Jurca et al. 2010]. As such, any website that collects users feedbacks, reviews or ratings has an interest in designing a robust and strategyproof mechanism to aggregate users preferences and ratings.

Besides the arithmetic mean, there are two other well-known ways of aggregating numerical values: the *median* and the *mode*. For a reviewer who wants to make the aggregate as close to its own opinion, the aggregator has a big impact on the best strategy.

Fig. 1. Tripadvisor's certificate of excellence: this certificate has been issued to an hotel reaching 4.5 stars out of 5 in 2012.

When aggregation happens through the mean, exaggerating the rating to the smallest or largest possible value allows the strongest influence on the aggregate. However, when aggregating through the median, there is no better strategy than to report the desired rating. For the mode, one should report the rating that is closest to the desired aggregate but also has a chance of being the most frequent one. As strategies thus would depend on beliefs that are hard to characterize, we do not consider the mode in this study.

While the median can be expected to induce the most truthful rating behavior, it is not as likely as the mean to reflect the true quality of the product when the reviewers are truthful. This is justified as follows. Suppose that each personal experience is the true quality perturbed by Gaussian noise, and the reviewers are truthful. Then, the mean is the maximum-likelihood estimator while the median often gives a different value. Thus, it is interesting to consider the tradeoff between the quality increase achieved through more truthful rating, and the quality loss that results from the possibly suboptimal aggregation.

To better study this tradeoff, we further consider an additional class of aggregators called the *truncated mean* method that falls between the mean and median. The truncated mean method is parameterized by $\alpha \in [0, 50)$ and forms the mean of ratings after eliminating the $\alpha\%$ smallest and $\alpha\%$ largest reported ratings. The truncated mean with $\alpha = 0$ equals to the mean method, and the truncated mean as $\alpha \to 50$ approaches the median method.

The goal of our work is to study the influence of different aggregators on rating behavior and the accuracy of the final aggregate. It would be best if we could carry out a study using data from actual rating websites. However, nobody knows what the true quality of the rated items is, nor do we have access to the beliefs of the raters. Furthermore, current rating websites are polluted by spammers that have other motivations than obtaining accurate rankings. Our interest is to model the behavior of raters whose intention is to make the website reflect the true quality, either because they are altruistic or because the website provides incentives based on its success.

We therefore introduce a new way of studying behavior. We place people in a game situation that closely mirrors the rating scenario. In the game, we can carefully control the beliefs and motivations of users, and observe their reaction. More specifically, we construct a game where the goal is to estimate the value of a variable by aggregating the noisy perceptions reported by the players. Rather than being scored on individual reports, players get their scores according to the aggregate of their own report with others. By controlling the observations and the other reports that are supposedly present, we can place players in different situations and evaluate their behavior.

2. RELATED WORK

The literature on online ratings focuses mostly on the economic [Chen et al. 2004; Chevalier and Mayzlin 2006; Duan et al. 2008] and social [Dellarocas 2000; Jøsang et al. 2007; Adamic et al. 2011] dimensions. Some works look more closely at the underlying distribution of ratings and the potential biases [Hu et al. 2006; Hu et al. 2009; Jurca et al. 2010; Feng et al. 2012]. However, only a few researches explore how rating aggregation should be made [Garcin et al. 2009a; McGlohon et al. 2010; Leberknight et al. 2012].

Ratings are most commonly aggregated using the arithmetic mean. However, the mean is quite sensitive to outliers and biases and thus may not be the most informative aggregator. Garcin et al. [2009a; 2009b] show that other aggregating functions perform better with respect to different criteria such as the informativeness, robustness and strategyproofness. On all these criteria, they demonstrate that the mean seems to be

the worst way of aggregating ratings. The median is more robust, has the advantage of being strategyproof and improves recommendation accuracy.

McGlohon et al. [2010] study how to aggregate reviews of different scales and from different sources. They look at statistical and re-weighting methods for aggregating ratings. For the former, they use techniques such as the mean, the median and lower bounds on normal and bimodal confidence interval. For the later, the idea is to give more weight to useful ratings. Because it is impossible to know the ground truth about the true quality of a product, they evaluate the accuracy of the proposed methods based on pairwise ranking of items and sampling from the existing users' ratings. The accuracy is then computed on the number of correctly ranked pairs of items. They conclude that the proposed methods do not outperform the mean, and the median performs poorly because of multiple ties.

Leberknight et al. [2012] introduce a rating aggregation technique based on the rating volatility. It has the advantage to capture the temporal trend of a product or service, and to be more responsive than the mean.

Closely related to rating aggregation is the research on vote aggregation in social choice theory [Clemen and Winkler 1999; Ariely et al. 2000; Larrick and Soll 2006]. Clemen and Winkler discuss the combination of experts' probability distributions in risk analysis. They conclude that simple rules such as the mean are important because they are easy to use, have robust performance, and are easy to justify in public policy settings. Ariely et al. [2000] suggests to take averages for quantitative judgement because it is a powerful and robust way of reducing the judgement error. Larrick and Soll [2006] also go in this direction by showing that people in general have misconceptions about the average and it should be used to reduce judgement error.

Armstrong [2001] surveys the literature about combining forecasts, and concludes that averaging improves accuracy. He also suggests that the truncated mean might be helpful when aggregating forecasts from 5 or more sources. However, he does not give any evidence to support this claim. In this work, we will see that truncated mean methods play an important role when the number of ratings is large enough.

For all these works, the ground truth about the true quality of a product is never known, and thus it makes it difficult to know how an aggregator will behave when implemented on a real website.

3. METHODOLOGY

When we consider product reports, it is almost impossible to know what is the true valuation of the quality of the product made by one user. Indeed, there will be always some bias [Jurca et al. 2010], and we are not yet able to read the user's mind. We need a way to "inject" a value in the user's mind.

Specifically, we model the product quality as a continuous signal following a known probability distribution function with known parameters. The user *samples* the signal according to this probability distribution function such that she forms an opinion on the perceived signal. This opinion will be very close to the value we want to "inject" as a belief.

More formally, let q^* denote the true quality of an item i, and $f(\cdot|q^*)$ a probability distribution. We denote $q_u \sim f(\cdot|q^*)$ the user's u perceived value of the item, and r_u the report of user u. The user u can actively get samples s such that $s \sim g(\cdot|q_u)$, where g is another (possibly the same as f) probability distribution. The range of g should be close to q_u in order to avoid too much noise, hence we bound g such that $s \in [q_u - \epsilon, q_u + \epsilon]$, with small ϵ. The probability distribution g models the uncertainty of a user on *perceiving* the personalized quality of a product, whereas f models the noise in generating personalized qualities of an item.

We are interested in measuring the true opinions about a product. In a traditional user study, we would show existing products to a group of users, and gather their opinions. However, this methodology would not work because the user already has an opinion on the displayed product which is difficult to change.

Instead of studying the behavior of actual raters, we construct an interactive game. This technique of *gamification* [Deterding et al. 2011] allows us to conduct a user study and collect data from users while they are playing a game. This approach increases user engagement and data quality [Chiu et al. 2009].

3.1. Game description

To avoid any influence from previous beliefs, the game places the players in a completely unfamiliar situation that closely mirrors the characteristics of a product rating scenario, assuming that the rater is motivated to make the rating reflect reality. In this game (called the *fishing game*), the player is assumed to be hired by a fishing company, and her job is to estimate the concentration of fish in different regions of the lake. For a given region, the player u has to evaluate and report the number of fish r_u in this region. The player sees reports $R \setminus \{r_u\}$ submitted by other players. She can probe as much as she wants the different regions of the lake to obtain samples s of the number of fish. The probe has a cost C, but this cost is negligible such that it does not influence the strategy of the player. The reward R depends on how closely the aggregate value corresponds to the true value as observed by the company. The player starts with a given budget and the goal of the game is to gather as much reward as possible. As the game is simulated, we can place the player in situations where her own measurements are very similar or very different to the reports of other players, and observe her behavior.

We define the reward function as follows

$$Re(R|q_u, \mathcal{A}) = 1 - \frac{|q_u - \mathcal{A}(R)|}{b - a} \tag{1}$$

where R is the set of all reports including the player's report, $\mathcal{A}$ the aggregator and q_u the user's perceived quality of the item, or in this case the user's perceived number of fish. For simplicity, we bound the probability distributions f and g between a and b, which means $q^*, q_u \in [a, b]$. We justify these bounds by the fact that on most review websites, the reports are bounded (for instance, 5 stars or 10 stars). Note that since the cost C is negligible, the utility of the player is equal to the reward function. The reward is maximum when the user's perceived number of fish equals the aggregated value $q_u = \mathcal{A}(R)$.

In order to collect accurate data, we test different combinations of the initial parameters: the true amount of fish q^* and the aggregator $\mathcal{A}$. This defines a scenario as a tuple $S = (q^*, \mathcal{A})$. The rest of the parameters (i.e. q_u) are directly derived from the scenario, except for f, g, ϵ and the boundaries a, b which are fixed. For a given scenario, we collect 3 reports (one per region of the lake).

Most product reviews websites such as Amazon, eBay or Google Play store display the report distribution as well as the average report. Only a few websites such as IMDB or BeerAdvocate use a complex weighted average and explicitly explain to the user how product reports are aggregated into an overall score. We believe that the information about the aggregator available to the users plays a crucial role in the behavior of the users. To ensure that users understand how reports are aggregated, we make them pass a short examination before entering the actual game.

The process of the game is the following: the player

(1) sees a short description of the game with its goal,

(2) sees a description of the aggregator with an example,
(3) takes an exam to validate the understanding of the aggregator,
(4) if she passes, plays one scenario S

Each of these steps is displayed on a new panel (screen). If a player fails at step 3, she can take another exam. However, we track the success rate of each player to filter out those players who do not understand the aggregator, or who tend to play randomly. When playing one scenario (step 4), it is possible to access step 1 and 2 at any time.

The game consists of a sequence of scenarios, where we vary the parameters we are interested in. We present in the next section their domain.

3.2. Settings

Intuitively, product reports should have a normal distribution centred on the true product quality, but it is often not the case. Hu et al. observe that Amazon product reports follow a U-shape distribution [Hu et al. 2006] or a J-shape distribution [Hu et al. 2009], a normal distribution centered to the extreme report. The cause of these report distributions comes from purchasing bias and under-reporting bias. Users with extreme opinions are more likely to "brag or moan", justifying a U-shape distribution. Users tend to purchase products with higher valuation, making a J-shape distribution. We consider only normal distributions, and we leave the study of other types of distribution as future work. Hence, we set $[a, b] = [0, 10]$ and let

$$q^* \sim \mathcal{U}(0, 10) \tag{2}$$
$$q_u \sim f = \mathcal{N}(q^*, \sigma^2) \tag{3}$$
$$s \sim g = \mathcal{N}(q_u, 0.45) \tag{4}$$

In the above formulas $\mathcal{U}$ is the uniform distribution and $\mathcal{N}$ is the normal distribution. We choose σ^2 such that it varies with respect to the true quality of the item q^*. If q^* is located near the boundaries of the domain $[a, b] = [0, 10]$ then σ^2 is smaller, making the distribution f narrower. We believe products with an average true quality tend to have a larger variance than products which are on the extremes.

Reports are most commonly aggregated using the arithmetic mean. However, the mean is quite sensitive to outliers and biases and thus may not be the most appropriate aggregator. There exist aggregation functions such as the median with interesting properties [Garcin et al. 2009a; Garcin et al. 2009b]. We consider the following aggregators:

— the mean is defined as the average value $\text{Mean}(R) = \frac{1}{|R|} \sum_{r \in R} r$
— the median is the value in the middle between the lower and upper half of the reports when the reports are ordered by their magnitude. When the number of reports is even, we break the ties by taking the average of them. Formally, consider $R = (r_{(1)}, r_{(2)}, ..., r_{(n)})$ the reports sorted in ascending order, and let $n = 2m - 1$ if n is odd, and $n = 2m$ if n is even for some integer m. The median is the value such as:

$$\text{Median}(R) = \begin{cases} r_{(m)} & \text{if } n \text{ is odd,} \\ \frac{r_{(m)} + r_{(m+1)}}{2} & \text{if } n \text{ is even.} \end{cases} \tag{5}$$

— the α-truncated mean (with $\alpha \in [0, 50)$) drops the highest and lowest $k = \lfloor \alpha(n - 1)/100 \rfloor$ reports, and compute the mean of the remaining reports.

$$\text{Mean}_\alpha(R) = \frac{1}{n - 2k} \sum_{i=k+1}^{n-k} r_{(i)} \tag{6}$$

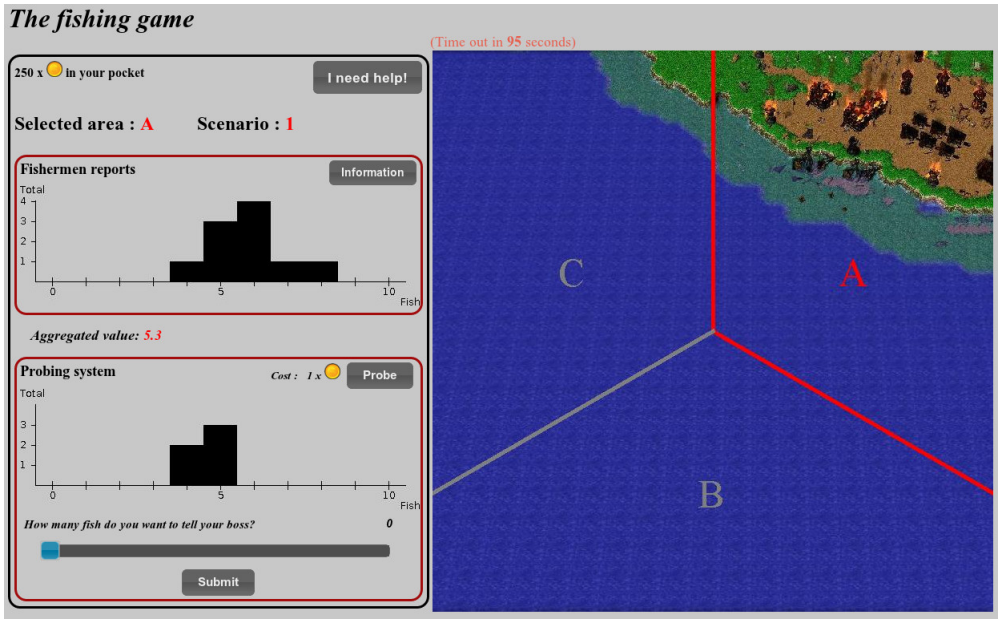


Fig. 2. Screenshot of the game: current reports made by other players (top left), sampling system (bottom left) and current region to estimate (right).

During the experiments, we will vary $\alpha \in \{10, 20, 30, 40\}$. Note that the mean and median are special cases of the truncated mean, respectively at $\alpha = 0$ and the median is the limit as $\alpha \to 50$ (at 50% no ratings would remain to take the average).

We measure two behaviors. The first one concerns the user and her reports. We want to see if users tend to report their true valuation or try to drive the aggregate value towards their true preference. Therefore, we will measure the percentage of *truthful* reports.

Definition 3.1 (*Truthfulness*). A report r_u from user u is *truthful* if

$$|q_u - r_u| \leq \epsilon \tag{7}$$

Truthfulness is defined within a margin ϵ which goes for the noise involved in the sampling made by the user.

Second, we look at the capability for an aggregator to reveal the true quality of the product. Consider a probability density function f symmetric on some unknown point θ that we want to estimate. In theory, mean, median and truncated mean are consistent estimators of θ because they converge to this value as the sample size gets bigger. However, in practice it might not be the case. Thus we need to look at another way to measure their performance. We define the *Mean Absolute Error* (MAE) as a measure of how spread out the estimator (the aggregated value) of a reports distribution is from the true value of a product.

Definition 3.2 (*Mean absolute error*). The *mean absolute error* e of an aggregator $\mathcal{A}(R)$ for a set of reports R, $n = |R|$ is given by

$$e(R|\mathcal{A}, q^*) = \frac{1}{n} \sum_{i=1}^{n} |q^* - \mathcal{A}(\{r_1, ..., r_i\})| \tag{8}$$

In addition, we measure the robustness of an aggregator to manipulations. We define a successful manipulation as a case where manipulators set the aggregate to a value of their choosing through malicious reports.

Definition 3.3. The *robustness* of an aggregator is a total order $\geq_r$ over aggregators. For any pair of aggregators $\mathcal{A}_1, \mathcal{A}_2$, we say that $\mathcal{A}_1 \geq_r \mathcal{A}_2$ iff in all scenarios where there is a manipulation through malicious reports that succeeds with aggregator $\mathcal{A}_1$, there is also a manipulation that succeeds with aggregator $\mathcal{A}_2$. We say that $\mathcal{A}_1 >_r \mathcal{A}_2$ iff $\mathcal{A}_1 \geq_r \mathcal{A}_2$ but not $\mathcal{A}_2 \geq_r \mathcal{A}_1$.

Thus, whenever $\alpha_1 > \alpha_2$, the α_1-truncated mean is more robust than the α_2-truncated mean, since it discards more reports and thus more reports of manipulators. This also shows that the mean is the least robust aggregator, while the median is the most robust aggregator.

Figure 2 is a screenshot of the game interface. On the right side, it shows which area of the lake is currently selected. On the left side, we display the distribution of reports from other players, and below the samples obtained from the probing system. In this particular case, the true number of fish was set to $q^* = 6$, while the true number of fish for the user was $q_u = 5$.

4. RESULTS

The first hypothesis is based on the observation that most people trust their own observations more than those of others, so that users would tend to want the aggregate to be as close as possible to their own beliefs. Thus, they would strategize less when the aggregator is more robust, leading to the following hypothesis:

HYPOTHESIS 1. *The number of true reports (truthfulness) increases as the degree of robustness increases.*

Besides truthfulness of the reports, another interesting question is how well the aggregate reflects the true value. The lowest mean absolute error would be reached with truthful reports and the mean as an aggregator. However, the mean does not elicit truthful reports and the median does not perform the best aggregation. Thus, we hypothesize:

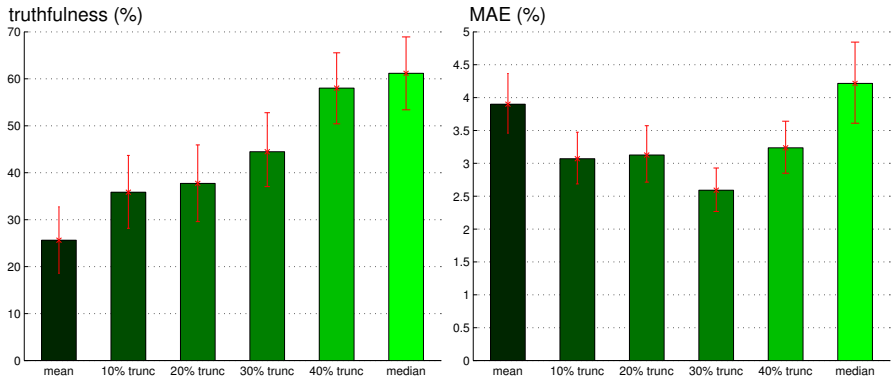
HYPOTHESIS 2. *There is a tradeoff between truthfulness and mean absolute error.*

We conduct two types of experiments. In the first type, we look at a static setting in which a product with artificial reports is presented to the users. In the game, we generate 10 artificial reports from the same distribution as q_u, and ask the users to play the game as described previously.

In the second type, we look at the dynamic process of rating a product. We conjecture that the behavior changes with the number of reports. A product with few reports is more versatile than one with a lot of reports. Hence a fake report would influence more the overall rating when the total number of reports is low. In addition, a malicious user might have greater incentives if she sees that it is easier to alter the aggregated value. So we will consider a dynamic setting where users rate a product in a sequence.

4.1. Static experiment

The static experiment is the following. For a given user, we select one aggregator and we generate the true number of fish $q^* \sim \mathcal{U}(0, 10)$. We draw 10 artificial reports from the distribution $\mathcal{N}(q^*, \sigma^2)$, plus one value which will be the user's true perceived value of the number of fish $q_u \sim \mathcal{N}(q^*, \sigma^2)$. This value q_u is the value we want to "inject" as a belief, and the user will be able to sample from. We do not take into account reports from users who do not sample.



(a) Truthfulness ($\epsilon = 1$) (b) Mean absolute error

Fig. 3. Truthfulness and mean absolute error for the static experiment.

We evaluate 6 aggregators: the mean, truncated mean at 10%, 20%, 30%, 40% and the median. Each user plays 3 scenarios per aggregator, for a total of 18 scenarios. Since there are 3 areas of the lake per scenario, it means she reports in total 54 ratings. 21 users from the Swiss Federal Institute of Technology (EPFL) community played the game. They are all scholars (Master students, PhD students, PostDoc and Professors) with a background in computer science. They played altogether for 7 hours and 16 minutes, which makes it 21 minutes per users, for an average of 1 minutes and 10 seconds per scenarios. We present averages with confidence intervals at 90% (bootstrap sampling, 10'000 samples).

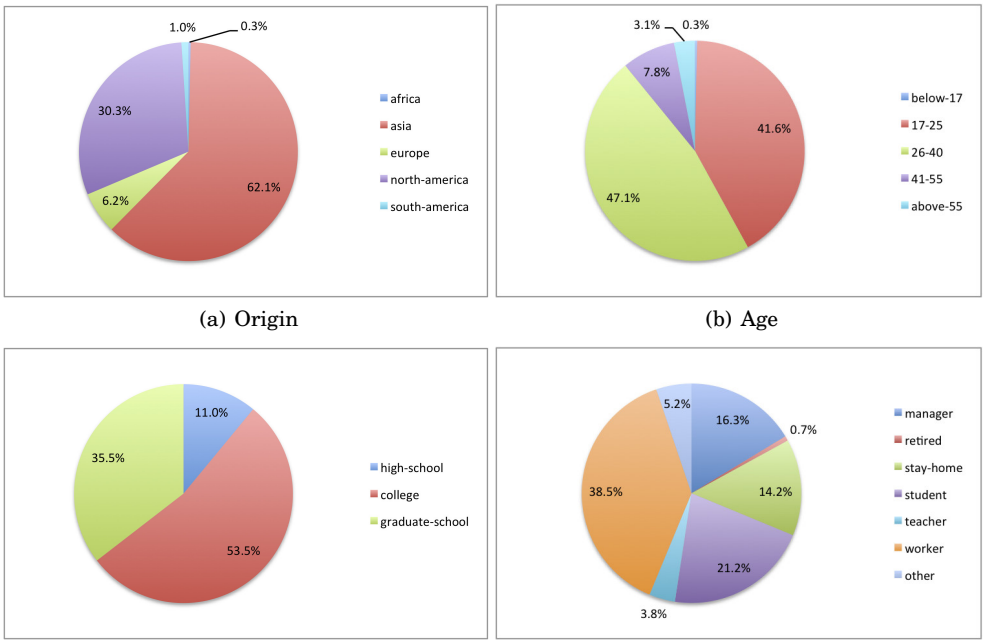
Figure 3(a) validates Hypothesis 1, and shows that indeed the truthfulness increases with the robustness. We fixed the threshold ϵ to 1. With the median which is the most robust aggregator, more than 60% of the reports are truthful, while with the mean we obtain only about 25% of truthful reports.

We might expect the truthfulness to be maximum at 100% with the median [Moulin 1980], but in practice it is not the case. This may actually be due to the fact that some participants do not fall prey to the fallacy of trusting their own observation, but also trust the reports of others. For normally distributed reports, the mean is the most accurate aggregator as it minimizes the mean square error. Users who trust the accuracy of other users' reports should be truthful with the mean, and non-truthful with the median because the median would not pick the best aggregated value.

Intuitively, the mean has the lowest MAE, and the median the highest. This holds under the assumption that the reports are truthful and that they come from the same underlying distribution. In practice, Figure 3(b) demonstrates that it is not the case, leading to Hypothesis 2. The mean and median have about the same MAE ($\sim$4%). However, the 30%-truncated mean dominates with a MAE near 2.5%. It is important to note that these numbers are not surprisingly low because the other reports are artificial and come from the same underlying distribution. We normalized the MAE by the maximum possible deviation (error) to the true value.

4.2. Dynamic experiment

In the static experiment, the reports attributed to other players were actually randomly generated so that our measurements of mean absolute error are not representative of a realistic scenario where players influence one another. Thus, we also ran a dynamic experiment using Amazon Mechanical Turk where a player saw actual reports from other players. 350 players played the game altogether for 3 days 18 hours



(a) Origin (b) Age

(c) Education (d) Occupation

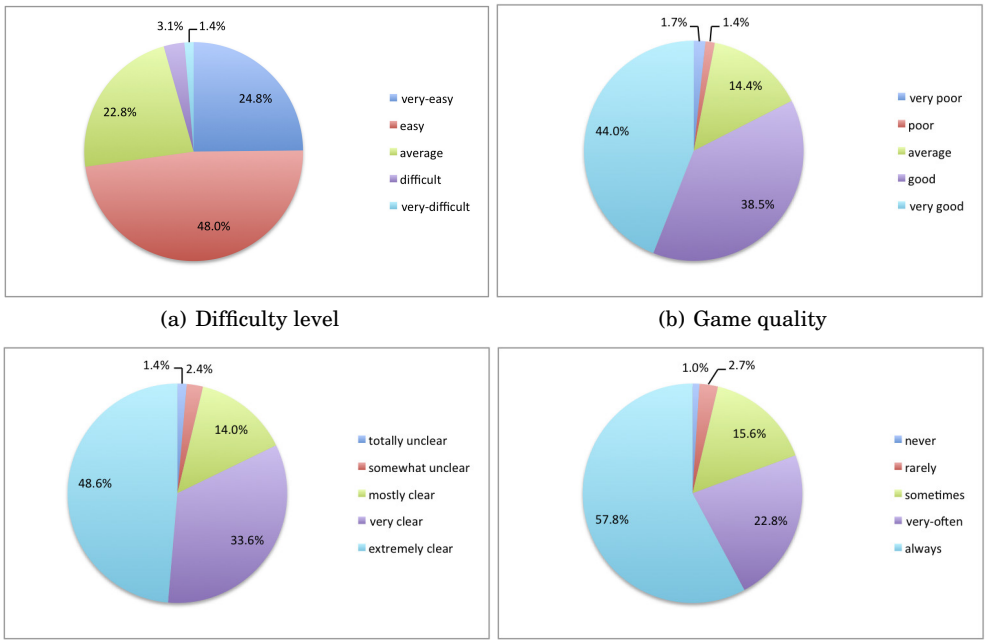
Fig. 4. Players' demographics

and 29 minutes, and gave 4765 reports (before filtering), for an average of 1 minutes and 8 seconds per report. For a given aggregator, each player played 5 scenarios, and gave one report per area for a total of 15 reports (5 scenarios with 3 areas). We do not take into account reports from players who do not sample. After this filtering, we had a total of 615 reports from different players in order to make a sequence of 41 reports per aggregator, scenario and area. We present averages with confidence intervals at 90% (bootstrap sampling, 10'000 samples).

Figure 4 shows the demographics of the players: their origin, age, education and current occupation. About two-thirds of the players are male (66.4%). These statistics about Amazon Mechanical Turk are not surprising and follow the trends observed by Ross et al. ([2010]).

Regarding the game itself (Fig. 5), players are satisfied. More than 93% of the players enjoyed playing this game. It was designed in such a way that it is not difficult to understand and play: more than 82% find the instruction clear, and about 73% find the game easy. The engagement of the players is such that they would like to see more tasks involving games in Amazon Mechanical Turk (Fig. 5(d)). It is important to note that players thought they were playing a game, and they did not realized they were involved in a user behaviour study.

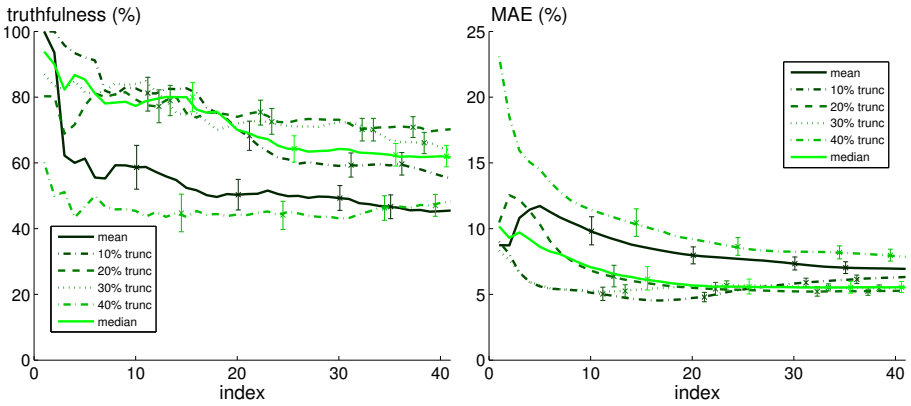
Figure 6 shows the average degree of truthfulness and the mean absolute error for different aggregators as the number of reports increases. At each time point, the player is shown the reports of previous players and contributes its own report. The curves show the evolution of both as reports are gathered incrementally, with players always shown the earlier reports. It thus mirrors closely the behavior of an online rating website. We normalized the MAE by the maximum possible deviation (error) to the true value.



(a) Difficulty level (b) Game quality

(c) Instruction clarity (d) Demand for similar tasks

Fig. 5. Game features



(a) Truthfulness ($\epsilon = 1$) (b) Mean absolute error

Fig. 6. Truthfulness and mean absolute error for the dynamic experiment.

The biggest differences in truthfulness and MAE exist when the number of reports is small. This can be expected as here the potential for manipulation is greatest. At the same time, it is also very crucial since the rating will determine the popularity of an item and thus the potential for even obtaining future ratings.

For both truthfulness and MAE, the mean is clearly not the best aggregator. The truthfulness drops dramatically when the number of reports is small. Players want the aggregate to be as close as possible to their own belief. However when the number of reports increases, some players tend to trust the reports of others (Fig. 6(a)). After

40 reports, we still see a difference between the mean and other aggregators. We fixed the threshold ϵ to 1.

Surprisingly, we observe that the 40%-truncated mean is the least truthful. We would expect it closer to the median. However, the truthfulness of this aggregator seems to increase after 35 reports. It is possible that we do not have enough reports to clearly see its behavior.

In general, the MAE is rather stable for all aggregators and the difference among them after 40 reports is small (Fig. 6(b)). The 10%- and 30%-truncated mean reach very fast a low MAE with few reports (less than 5). However, the MAE of the 10%-truncated mean increases as the number of reports increases.

Regarding Hypothesis 1, there is a difference between the most and least robust aggregators, i.e. the median and mean respectively. However, the ranking is not clear for the truncated mean methods. Perhaps the aggregators are less stable when the number of reports is small and they need more than 40 reports to show a proper behavior.

Hypothesis 2 is also valid in the dynamic setting. However this tradeoff exists mostly when the number of reports is small because the difference among aggregators is more important for both truthfulness and MAE. Note however that in a product ranking, even small differences in ratings can have a huge influence on the position in the ranking, so that the differences in MAE can still have a big impact.

Overall, the 20%- or 30%-truncated mean seem to be good candidates for aggregators. They both achieve high truthfulness and low MAE over small and large numbers of reports, consistent with the behavior in the static scenario.

5. CONCLUSION AND FUTURE WORK

We have used a new form of experiment based on a game to predict behavior of raters on product review websites. We have focused on the behavior of users who are well-intentioned and act so as to make the aggregate ranking reflect their belief of the true quality value, either through altruism or through incentives based on the success with other users. The fishing game closely mirrors the incentives of a product rating scenario, but in a completely artificial context where players' beliefs can be perfectly manipulated. Using two experiments, we have demonstrated that while the median aggregator elicits more truthful reports, its overall accuracy is not better than that of the mean since it does not perform the most accurate aggregation. We have shown that a different aggregator, the 30%-truncated mean, provides the best overall result.

The fact that the median elicits the most truthful reports can be explained by the fact that people tend to trust their own observations more than others'. Actually, a user who trusts others' reports should be most truthful when aggregation is via the mean, as this is the most accurate aggregator.

The usefulness of the results for constructing accurate rating websites is limited because users may not understand the functioning of aggregators such as the truncated mean. In our study with Amazon Mechanical Turk, about $1/3$ of the users did not pass the quiz that tested their understanding of the aggregator. However, we believe that our study provides an important insight that should be studied further in the context of operational websites, as users may become increasingly sophisticated.

As future work and with the help of this methodology, we plan to study different distributions of ratings such as the bimodal, explore new aggregators like the geometric mean, investigate the effect of discretization of the ratings (for instance, small (5 stars) vs large (10 stars or more) range to choose the rating). We also plan to extend the framework to multi-dimensional ratings (overall ratings and ratings per feature of a product).

REFERENCES

ADAMIC, L. A., LAUTERBACH, D., TENG, C.-Y., AND ACKERMAN, M. 2011. Rating friends without making enemies. In *Proceedings of the International Conference on Weblogs and Social Media*.

ARIELY, D., TUNG AU, W., BENDER, R., BUDESCU, D., DIETZ, C., GU, H., WALLSTEN, T., AND ZAUBERMAN, G. 2000. The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied 6,* 2, 130–147.

ARMSTRONG, J. S. 2001. Combining forecasts. *International series in operations research and management science*, 417–440.

CHEN, P.-Y., WU, S.-Y., AND YOON, J. 2004. The impact of online recommendations and consumer feedback on sales. In *Proceedings of the International Conference on Information Systems*. 711–724.

CHEVALIER, J. AND MAYZLIN, D. 2006. The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research 43,* 3, 345–354.

CHIU, M.-C., CHANG, S.-P., CHANG, Y.-C., CHU, H.-H., CHEN, C. C.-H., HSIAO, F.-H., AND KO, J.-C. 2009. Playful bottle: a mobile social persuasion system to motivate healthy water intake. In *Proceedings of the International Conference on Ubiquitous Computing*.

CLEMEN, R. AND WINKLER, R. 1999. Combining probability distributions from experts in risk analysis. *Risk Analysis 19*, 187–203.

DELLAROCAS, C. 2000. Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In *Proceedings of the ACM Conference on Electronic Commerce*. 150–157.

DETERDING, S., KHALED, R., NACKE, L. E., AND DIXON, D. 2011. Gamification, toward a definition. In *Proceedings of the CHI 2011 Gamification Workshop*.

DUAN, W., GU, B., AND WHINSTON, A. 2008. Do online reviews matter? an empirical investigation of panel data. *Decision Support Systems 45,* 4, 1007–1016.

FENG, S., XING, L., GOGAR, A., AND CHOI, Y. 2012. Distributional footprints of deceptive product reviews. In *Proceedings of the International Conference on Weblogs and Social Media*.

GARCIN, F., FALTINGS, B., AND JURCA, R. 2009a. Aggregating reputation feedback. In *Proceedings of the International Conference on Reputation: Theory and Technology - ICORE 09*.

GARCIN, F., FALTINGS, B., JURCA, R., AND JOSWIG, N. 2009b. Rating aggregation in collaborative filtering systems. In *Proceedings of the International Conference on Recommender Systems*.

HU, N., PAVLOU, P. A., AND ZHANG, J. 2006. Can online reviews reveal a product's true quality?: empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the ACM conference on Electronic Commerce*. 324–330.

HU, N., ZHANG, J., AND PAVLOU, P. A. 2009. Overcoming the j-shaped distribution of product reviews. *Communications of the ACM 52,* 10, 144–147.

JØSANG, A., ISMAIL, R., AND BOYD, C. 2007. A survey of trust and reputation systems for online service provision. *Decision Support Systems 43,* 2, 618–644.

JURCA, R., GARCIN, F., TALWAR, A., AND FALTINGS, B. 2010. Reporting incentives and biases in online review forums. *Transactions on the Web 4,* 2, 1–27.

LARRICK, R. AND SOLL, J. 2006. Intuitions about combining opinions: Misappreciation of the averaging principle. *Management science 52,* 1, 111–127.

LEBERKNIGHT, C., SEN, S., AND CHIANG, M. 2012. On the volatility of online ratings: An empirical study. In *Workshop on eBusiness*.

MCGLOHON, M., GLANCE, N., AND REITER, Z. 2010. Star quality: Aggregating reviews to rank products and merchants. In *Proceedings of the International Conference on Weblogs and Social Media*.

MOULIN, H. 1980. On strategy-proofness and single peakedness. *Public Choice 35*, 437–455.

ROSS, J., IRANI, L., SILBERMAN, M. S., ZALDIVAR, A., AND TOMLINSON, B. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the international conference on Human factors in computing systems*.