Reporting Incentives and Biases in Online Review Forums

RADU JURCA
Google Switzerland
and
FLORENT GARCIN, ARJUN TALWAR, and BOI FALTINGS
Artificial Intelligence Lab, Ecole Polytechnique Fédérale de Lausanne (EPFL)

Online reviews have become increasingly popular as a way to judge the quality of various products and services. However, recent work demonstrates that the absence of reporting incentives leads to a biased set of reviews that may not reflect the true quality. In this paper, we investigate underlying factors that influence users when reporting feedback. In particular, we study both reporting incentives and reporting biases observed in a widely used review forum, the Tripadvisor Web site. We consider three sources of information: first, the numerical ratings left by the user for different aspects of quality; second, the textual comment accompanying a review; third, the patterns in the time sequence of reports. We first show that groups of users who discuss a certain feature at length are more likely to agree in their ratings. Second, we show that users are more motivated to give feedback when they perceive a greater risk involved in a transaction. Third, a user's rating partly reflects the difference between true quality and prior expectation of quality, as inferred from previous reviews. We finally observe that because of these biases, when averaging review scores there are strong differences between the mean and the median. We speculate that the median may be a better way to summarize the ratings.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics

General Terms: Economics, Experimentation, Reliability

Additional Key Words and Phrases: Online reviews, reputation mechanisms

ACM Reference Format:

Jurca, R., Garcin, F., Talwar, A., and Faltings, B. 2010. Reporting incentives and biases in online review forums. ACM Trans. Web 4, 2, Article 5 (April 2010), 27 pages.
DOI = 10.1145/1734200.1734202 http://doi.acm.org/10.1145/1734200.1734202

The article extends Talwar et al. [2007].
This work was conducted while R. Jurca was at EPFL.

Authors' addresses: R. Jurca, Google Inc. Switzerland; email: radu.jurca@gmail.com; F. Garcin, A. Talwar, and B. Faltings, EPFL, I&C, LIA, Lausanne, Switzerland; email: florent.garcin@epfl.ch; arjun@math.stanford.edu; boi.faltings@epfl.ch.

DOI 10.1145/1734200.1734202 http://doi.acm.org/10.1145/1734200.1734202

1. INTRODUCTION

The spread of the Internet has made online feedback forums (or reputation mechanisms) an important channel for word-of-mouth regarding products, services, or other types of commercial interactions. Numerous empirical studies show that buyers seriously consider online feedback when making purchasing decisions, and are willing to pay *reputation premiums* for products or services that have a good reputation [Houser and Wooders 2006; Melnik and Alm 2002; Kalyanam and McIntyre 2001; Dellarocas et al. 2006].

Recent analysis, however, raises important questions regarding the ability of existing forums to reflect the real quality of a product. In the absence of clear incentives, users with a moderate outlook will not bother to voice their opinions, which leads to an unrepresentative sample of reviews. For example, Hu et al. [2006] and Admati and Pfleiderer [2000] show that Amazon[1] ratings of books or CDs follow with great probability bimodal, U-shaped distributions where most of the ratings are either very good or very bad. Controlled experiments, on the other hand, reveal opinions on the same items that are normally distributed. Under these circumstances, using the arithmetic mean to predict quality (as most forums actually do) gives the typical user an estimator with high variance that is often false.

Improving the way we aggregate the information available from online reviews requires a deep understanding of the underlying factors that bias the rating behavior of users. Hu et al. [2006] propose the "Brag-and-Moan Model" where users rate only if their utility of the product (drawn from a normal distribution) falls outside a median interval. The authors conclude that the model explains the empirical distribution of reports, and offers insights into smarter ways of estimating the true quality of the product.

In this paper, we extend this line of research, and investigate other factors that contribute to the user's decision of *when* and *what* feedback to submit to an online forum. We consider actual hotel reviews from the TripAdvisor[2] Web site, and use the following sources of information:

—the numerical ratings left by the user for various aspects of hotel quality;

—the textual comment accompanying the review;

—the number of votes left by other users who considered the review helpful;

—the time sequence of reviews;

We first analyze simple linguistic evidence from the textual review that usually accompanies the numerical ratings. We use text-mining techniques similar to those of Ghose et al. [2005] and Cui et al. [2006]; however, we are only interested in identifying *what* aspects of the service the user is discussing, without computing the semantic orientation of the text. We find that users who comment more on the same feature are more likely to agree on a common numerical rating for that particular feature. Intuitively, lengthy comments reveal the importance of the feature to the user. Since people tend to be more knowledgeable

[1] http://www.amazon.com

[2] http://www.tripadvisor.com/

in the aspects they consider important, users who discuss a given feature in more detail might be assumed to have more *authority* in evaluating that feature. This conclusion is supported by the observation that (i) lengthy comments are generally considered more useful by the other users, and (ii) lengthy comments are not associated to outlier reviews and cover the entire spectrum of ratings.

Second, we identify a correlation between the effort spent in writing a review and the risk perceived by the user for the corresponding transaction. The underlying hypothesis is that users feel more compelled to contribute with feedback about transactions that a priori involve higher risk. For example, buying a book online is usually a low-risk transaction—even if the buyer does not receive the book, the price paid is usually low enough to spare the buyer from major losses. The same can be argued about booking a cheap hotel. Even if the hotel is not great, the risk of obtaining a quality level significantly below the expectations is accepted by the traveler. When booking a high-end hotel, on the other hand, the risk of bad service is less acceptable and might even compromise the purpose of the trip. A romantic weekend can turn out a nightmare in a dirty, unfriendly hotel; similarly, a business meeting organized in a hotel without appropriate facilities may waste participants' time. Users seem to recognize that high-end hotels expose the travelers to higher risk (of taking the wrong decision) and are therefore more diligent in reviewing these hotels.

Third, we investigate the relationship between a review and the reviews that preceded it. A perusal of online reviews shows that ratings are often part of discussion threads, where one post is not independent of other posts. One may see, for example, users who make an effort to contradict, or vehemently agree with, the remarks of previous users. By analyzing the time sequence of reports, we conclude that past reviews influence the future reports, as they create some prior expectation regarding the quality of service. The subjective perception of the user is influenced by the gap between the prior expectation and the actual performance of the service [Parasuraman et al. 1985, 1998; Olshavsky and Miller 1972; Teas 1993] which will later be reflected in the user's rating.

The preceding results can be used to improve the way reputation mechanisms aggregate the information from individual reviews. First, understanding the reporting incentives might lead to new user interfaces that (a) make feedback reporting easier and faster, and (b) help the user feel that her contribution helps others take better decisions. Second, the biases present in the submitted reviews can be corrected by customized aggregation algorithms. For example, the mechanism can compute feature-by-feature estimates of quality, where for each feature, it considers only the subset of reviews corresponding to lengthy comments on that feature. As another example, the mechanism could correct the bias introduced by the expectation created by previous reviews when estimating the real quality.

Finally, we observe that rating distributions are skewed, and show how different ways of averaging ratings produce different quality estimates and different rankings of hotels. For example, rating averages computed based on the arithmetic mean often produce rankings that fluctuate over time; one possible explanation is that raters constantly correct the current average with exaggerated ratings. In contrast, the ranking determined by the median rating is much more stable and robust against outliers. Furthermore, we argue that the median is the only aggregator that makes it best for the users to submit their true rating. Provided that users can be made to understand such incentives, this could lead to generally more accurate feedback.

The remainder of this article begins with the description of the dataset we use for our study. Then, Section 3 analyzes textual reviews. We explain the text-mining methods that we use to classify textual reviews and we define this classification. In Section 4, we discuss the correlation between the risk associated to a hotel and the effort spent by reviewers in describing their experience with that hotel. After that, we investigate in Section 5 the influence of past ratings on the current reviewer. Finaly, Section 6 explains why the average rating is not an accurate estimate of quality, and argues for the use of the mode rating instead.

2. THE DATASET

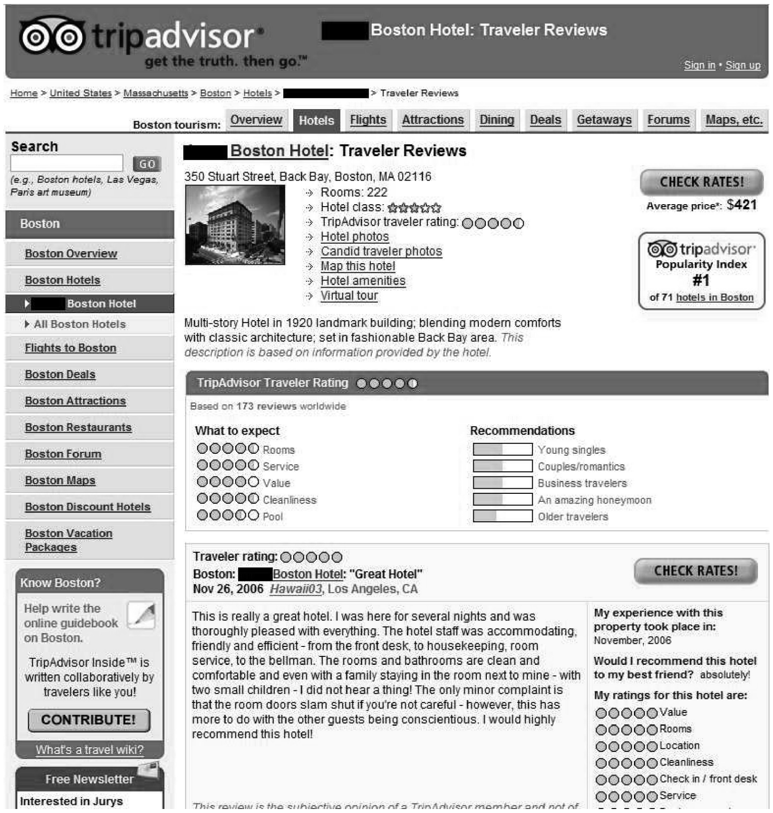
We consider real hotel reviews collected from the popular travel site TripAdvisor. TripAdvisor indexes hotels from cities across the world, along with reviews written by travelers. Users can search the site by giving the hotel's name and location (optional). The reviews for a given hotel are displayed as a list (ordered from the most recent to the oldest), with 5 reviews per page. The reviews contain:

—information about the author of the review (e.g., dates of stay, username of the reviewer, location of the reviewer);
—the overall rating (from 1, lowest, to 5, highest);
—a textual review containing a title for the review, free comments, and the main things the reviewer liked and disliked;
—numerical ratings (from 1, lowest, to 5, highest) for different features (e.g., cleanliness, service, location, etc.);
—the number of votes left by other users for and against the review.

Below the name of the hotel, TripAdvisor displays the address of the hotel, general information (number of rooms, number of stars, short description, etc), the average overall rating, the TripAdvisor ranking, and an average rating for each feature. Figure 1 shows the page for a popular Boston hotel whose name (along with advertisements) was explicitly erased.

We selected four cities for this study: Boston, Las Vegas, New York, and Sydney. For each city we considered all hotels that had at least 10 reviews, and recorded all reviews. Table I presents the number of hotels considered in each city, the total number of reviews recorded for each city, and the distribution of hotels with respect to the star-rating (as available on the TripAdvisor site). Note that not all hotels have a star-rating.

For each review we recorded the overall rating, the textual review (title and body of the review), the number of votes, and the numerical rating on



[redacted] Boston Hotel: Traveler Reviews

[redacted] Boston Hotel: Traveler Reviews

350 Stuart Street, Back Bay, Boston, MA 02116

- Rooms: 222
- Hotel class: ☆☆☆☆☆
- TripAdvisor traveler rating: ●●●●◐
- Hotel photos
- Candid traveler photos
- Map this hotel
- Hotel amenities
- Virtual tour

Average price*: $421

tripadvisor Popularity Index #1 of 71 hotels in Boston

Multi-story Hotel in 1920 landmark building; blending modern comforts with classic architecture; set in fashionable Back Bay area. *This description is based on information provided by the hotel.*

TripAdvisor Traveler Rating ●●●●◐

Based on 173 reviews worldwide

What to expect		Recommendations
●●●●◐	Rooms	Young singles
●●●●◐	Service	Couples/romantics
●●●●○	Value	Business travelers
●●●●◐	Cleanliness	An amazing honeymoon
●●●◐○	Pool	Older travelers

Traveler rating: ●●●●●

Boston: [redacted] Boston Hotel: "Great Hotel"

Nov 26, 2006 *Hawaii03*, Los Angeles, CA

This is really a great hotel. I was here for several nights and was thoroughly pleased with everything. The hotel staff was accommodating, friendly and efficient - from the front desk, to housekeeping, room service, to the bellman. The rooms and bathrooms are clean and comfortable and even with a family staying in the room next to mine - with two small children - I did not hear a thing! The only minor complaint is that the room doors slam shut if you're not careful - however, this has more to do with the other guests being conscientious. I would highly recommend this hotel!

This review is the subjective opinion of a TripAdvisor member and not of

My experience with this property took place in: November, 2006

Would I recommend this hotel to my best friend? absolutely!

My ratings for this hotel are:

- ●●●●● Value
- ●●●●● Rooms
- ●●●●● Location
- ●●●●● Cleanliness
- ●●●●● Check in / front desk
- ●●●●● Service

Help write the online guidebook on Boston.

TripAdvisor Inside™ is written collaboratively by travelers like you!

Interested in Jurys

Fig. 1. The TripAdvisor page displaying reviews for a popular Boston hotel. Name of hotel and advertisements were deliberatively erased.

Table I. A Summary of the Dataset

City	# Reviews	# Hotels	# of Hotels with 1, 2, 3, 4 & 5 stars
Boston	5537	66	2+4+23+15+5
Las Vegas	28553	131	12+31+39+17+7
New York	40676	264	20+20+76+44+19
Sydney	3659	103	0+1+29+19+10

7 features: *Rooms*(R), *Service*(S), *Cleanliness*(C), *Value*(V), *Food*(F), *Location*(L) and *Noise*(N). TripAdvisor does not require users to submit anything other than the overall rating, hence a typical review rates few additional features, regardless of the discussion in the textual comment. Only the features *Rooms*(R), *Service*(S), *Cleanliness*(C) and *Value*(V) are rated by a significant number of users (see Table II). However, we also selected the features *Food*(F), *Location*(L) and *Noise*(N) because they are mentioned by a significant number of textual comments. For each feature, we record the numerical rating given by the user. The typical length of the textual comment amounts to approximately 200 words. All data was collected by crawling the TripAdvisor site in July 2007.

Table II. Number of Reviews Containing Numerical Ratings for Each Feature

City	# Rooms	# Service	# Cleanliness	Value#
Boston	3475	3414	3471	3438
Las Vegas	17376	17098	17387	17271
New York	25535	25050	25506	25270
Sydney	2610	2554	2612	2579

2.1 Formal Notation

We will formally refer to a review by a tuple (r, T) where:

—$r = (r_f)$ is a vector containing the ratings $r_f \in \{0, 1, \ldots 5\}$ for the features $f \in F = \{O, R, S, C, V, F, L, N\}$; note that the overall rating, r_O, is abusively recorded as the rating for the feature *Overall*(O);

—T is the textual comment that accompanies the review.

Reviews are indexed according to the variable i, such that (r^i, T^i) is the i^{th} review in our database. Since we do not record the username of the reviewer, we will also say that the i^{th} review in our dataset was submitted by user i. When we need to consider only the reviews of a given hotel, h, we will use $(r^{i(h)}, T^{i(h)})$ to denote the i^{th} review about the hotel h.

3. EVIDENCE FROM TEXTUAL COMMENTS

The free textual comments associated to online reviews are a valuable source of information for understanding the reasons behind the numerical ratings left by the reviewers. The text may, for example, reveal concrete examples of aspects that the user liked or disliked, thus justifying some of the high, respectively low ratings for certain features. The text may also offer guidelines for understanding the preferences of the reviewer, and the weights of different features when computing an overall rating.

The problem, however, is that free textual comments are difficult to read. Users are required to scroll through many reviews and read mostly repetitive information. Significant improvements would be obtained if the reviews were automatically interpreted and aggregated. Unfortunately, this seems a difficult task for computers since human users often use witty language, abbreviations, cultural-specific phrases, and the figurative style.

Nevertheless, several important results use the textual comments of online reviews in an automated way. Using well established natural language techniques, reviews or parts of reviews can be classified as having a positive or negative *semantic orientation*. Pang et al. [2002] classify movie reviews into positive/negative by training three different classifiers (Naive Bayes, Maximum Entropy and SVM) using classification features based on unigrams, bigrams, or part-of-speech tags.

Dave et al. [2003] analyze reviews from CNet and Amazon, and surprisingly show that classification features based on unigrams or bigrams perform better than higher-order *n-grams*. This result is challenged by Cui et al. [2006] who look at large collections of reviews crawled from the web. They show that the size of the data set is important, and that bigger training sets allow classifiers to successfully use more complex classification features based on *n-grams*.

Hu and Liu [2004] also crawl the web for product reviews and automatically identify product attributes that have been discussed by reviewers. They use Wordnet to compute the semantic orientation of product evaluations and summarize user reviews by extracting positive and negative evaluations of different product features. Popescu and Etzioni [2005] analyze a similar setting, but use search engine hit-counts to identify product attributes; the semantic orientation is assigned through the *relaxation labeling technique*.

Ghose et al. [2005] and Ghose et al. [2006] analyze seller reviews from the Amazon secondary market to identify the different dimensions (e.g., delivery, packaging, customer support, etc.) of reputation. They parse the text, and tag the part-of-speech for each word. Frequent nouns, noun phrases and verbal phrases are identified as dimensions of reputation, while the corresponding *modifiers* (i.e., adjectives and adverbs) are used to derive numerical scores for each dimension. The enhanced reputation measure correlates better with the pricing information observed in the market. Pavlou and Dimoka [2006] analyze eBay reviews and find that textual comments have an important impact on reputation premiums.

Our approach is similar to the previously mentioned works, in the sense that we identify the aspects (i.e., hotel features) discussed by the users in the textual reviews. However, we do not compute the semantic orientation of the text, nor attempt to infer missing ratings.

We define the *weight*, w^i_f, of feature $f \in F \setminus \{O\}$ in the text T^i associated with the review (r^i, T^i), as the fraction of T^i dedicated to discussing aspects (both positive and negative) related to feature f. We propose an elementary method to approximate the values of these weights. For each feature we manually construct the word list L_f containing approximately 50 words that are most commonly associated to the feature f. The initial words were selected from reading some of the reviews, and manually selecting words that refer to our seven features. The list was then manually extended by adding synonyms from an online dictionary[3] and thesaurus.[4] Finally, we brainstormed in our research group for missing words that would normally be associated with each of the features.[5]

We acknowledge the ad-hoc manner of constructing these word lists, and plan to improve this process in our future work. For example, the first improvement would be to automate the process of extending the manually constructed set of seed words. Structured lexicons like WordNet, for example, can allow us to write algorithms that automatically consider the best synonyms and/or antonyms of the seed words. A second direction is to use machine learning and

[3] www.dictionary.com

[4] www.thesaurus.com

[5] For example, the list of words for the feature Rooms was: *room, space, interior, decor, ambiance, atmosphere, comfort, bath, toilet, bed, building, wall, window, private, temperature, sheet, linen, pillow, hot, water, cold, water, shower, lobby, furniture, carpet, air, condition, mattress, layout, design, mirror, ceiling, lighting, lamp, sofa, chair, dresser, wardrobe, closet*. All words serve as prefixes.

derive association between words based on analyzing large corpora of online reviews. This method would not only provide an extensive set of synonyms and/or antonyms, but can also give an indication of the intensity of different words when referring to certain features.

Let $count(l, T^i)$ be the function counting the number of iteration of word l in text T^i. Each term $l \in L_f$ is counted the number of times it appears in T^i, with two exception:

- in cases where the user submits a title to the review, we account for the title text by appending it three times to the review text T^i. The intuitive assumption is that the user's opinion is more strongly reflected in the title, rather than in the body of the review. For example, many reviews are accurately summarized by titles such as *"Excellent service, terrible location"* or *"Bad value for money"*;
- certain words that occur only once in the text are counted multiple times if their relevance to that feature is particularly strong. These were "root" words for each feature (e.g., "staff" is a root word for the feature *Service*), and were weighted either 2 or 3. Each feature was assigned up to 3 such root words, so almost all words are counted only once.

The weight w_f^i is computed as:

$$w_f^i = \frac{\sum_{l \in L_f} count\,(l, T^i)}{\sum_{f \in F \setminus \{O\}} \sum_{l \in L_f} count\,(l, T^i)} \tag{1}$$

To keep a uniform notation, we also define the weight for the feature *Overall(O)* as the normalized length of the entire textual comment associated to a review:

$$w_O^i = \frac{|T^i|}{\max_i |T^i|};$$

where $|T^i|$ is the number of character in the textual comment T^i.

The following is a TripAdvisor review for a Boston hotel (the name of the hotel is omitted):

> "I'll start by saying that I'm more of a Holiday Inn person than a *** type. So I get frustrated when I pay double the room rate and get half the amenities that I'd get at a Hampton Inn or Holiday Inn. The location was definitely the main asset of this place. It was only a few blocks from the Hynes Center subway stop and it was easy to walk to some good restaurants in the Back Bay area. Boylston isn't far off at all. So I had no trouble with foregoing a rental car and taking the subway from the airport to the hotel and using the subway for any other travel. Otherwise, they make you pay for anything and everything. And when you've already dropped $215/night on the room, that gets frustrating. The room itself was decent, about what I would expect. Staff was also average, not bad and not excellent. Again, I think you're paying for location and the ability to walk to a lot of good stuff. But I think next time I'll stay in Brookline, get more amenities, and use the subway a bit more.". The title is: "Good location, but you pay for it."

This numerical ratings associated to this review are $r_O = 3$, $r_R = 3$, $r_S = 3$, $r_C = 4$, $r_V = 2$ for features *Overall*(O), *Rooms*(R), *Service*(S), *Cleanliness*(C) and *Value*(V) respectively. The ratings for the features *Food*(F), *Location*(L) and *Noise*(N) are absent (i.e., $r_F = r_L = r_N = 0$).

The weights w_f are computed from the following lists of common terms:

$count\,(l \in L_R, T)$ ={3 * room}; $w_R = 0.103$
$count\,(l \in L_S, T)$ ={Staff (3x), 2 * amenities}; $w_S = 0.172$
$count\,(l \in L_C, T) = \emptyset$; $w_C = 0$
$count\,(l \in L_V, T)$ ={\$, 2 * rate}; $w_V = 0.103$
$count\,(l \in L_F, T)$ ={restaurant}; $w_F = 0.034$
$count\,(l \in L_L, T)$ ={center (2x), 2 * walk (2x), 5 * location (2x), area};
  $w_L = 0.586$
$count\,(l \in L_N, T) = \emptyset$; $w_N = 0$

The root words ‘staff’ and ‘center’, ‘walk’, ‘location’ were tripled and doubled respectively. The overall weight of the textual review (i.e., its normalized length) is $w_O = 0.197$. These values account reasonably well for the weights of different features in the discussion of the reviewer.

One point to note is that some terms in the lists L_f possess an inherent semantic orientation. For example the word ‘grime’ (belonging to the list L_C) would be used most often to assert the presence, and not the absence of grime. This is unavoidable, but care was taken to ensure words from both sides of the spectrum were used. For this reason, some lists such as L_R contain only nouns of objects that one would typically describe in a room.

The goal of this section is to analyze the influence of the weights w^i_f on the numerical ratings r^i_f. Intuitively, users who spent a lot of their time discussing a feature f (i.e., w^i_f is high) had something to say about their experience with regard to this feature. Obviously, feature f is important for user i. Since people tend to be more knowledgeable in the aspects they consider important, our hypothesis is that the ratings r^i_f (corresponding to high weights w^i_f) constitute a subset of “expert” ratings for feature f.

Figure 2 plots the distribution of the rates $r^{i(h)}_C$ with respect to the weights $w^{i(h)}_C$ for the cleanliness of a Las Vegas hotel, h. Here, the high ratings are restricted to the reviews that discuss little the cleanliness. Whenever cleanliness appears in the discussion, the ratings are low. Many hotels exhibit similar rating patterns for various features. Ratings corresponding to low weights span the whole spectrum from 1 to 5, while the ratings corresponding to high weights are more grouped together.

We therefore make the following hypothesis:

HYPOTHESIS 1. *The ratings r^i_f corresponding to the reviews where w^i_f is high, are more similar to each other than to the overall collection of ratings.*

To test the hypothesis, we take the entire set of reviews, and feature by feature, we compare the standard deviation of the ratings with high weights, against the standard deviation of all ratings. First, we define a high weight as any weight which is in the upper 20% percentile of the entire set of weights for the corresponding feature. Concretely, let W_f be the set of all weights w_f across all hotels and all reviews. We compute the thresholds w^*_f such that 20%

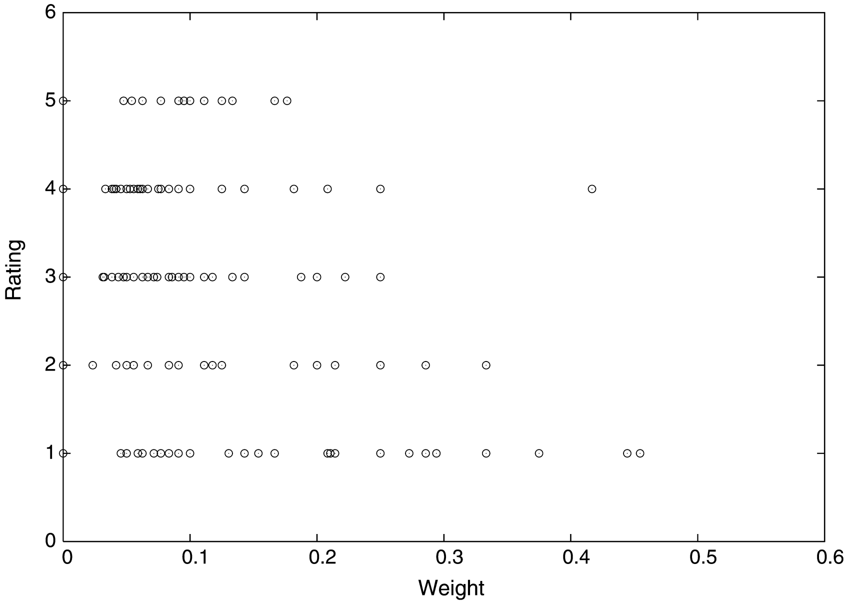


Fig. 2. The distribution of ratings against the weight of the cleanliness feature.

of the values in W_f are above w_f^*. A given weight w_f is said to be a *high weight* if and only if $w_f > w_f^*$. Next, feature by feature, we take every hotel h and identify the reviews where the rating r_f exists and the weight w_f is high. Let $R_f^*(h)$ denote this set of reviews with high weights for feature f and hotel h. Few hotels had fewer than 5 reviews in the set $R_f^*(h)$, and those were ignored in the experiment. Most hotels, however, had many more high weight reviews: the average number of high weight reviews a hotel had for each of the features O, R, S, C and V is 29.82, 21.45, 20.98, 23.35 and 22.81 respectively. The third step of the experiment is to randomly select $N = |R_f^*(h)|$ reviews for the hotel h where the rating r_h for the feature f exists. Let this set be $R_f(h)$. Finally we compute the unbiased standard deviation of the ratings in $R_f^*(h)$ and $R_f(h)$ as:

$$std = \sqrt{\sum_{i=1}^{N} \frac{(x_i - \bar{x})^2}{N-1}}, \tag{2}$$

where $N = |R_f^*(h)|$ is the cardinality of the set $R_f^*(h)$.

City by city and feature by feature, Table III presents the average standard deviation for the two sets. Indeed, the ratings with high weights have lower standard deviation, which confirms our hypothesis. The significance levels were computed with a standard T-test. Also note that only the features O,R,S,C, and V were considered, since for the others (F, L, and N) we did not have enough ratings.

Hypothesis 1 not only provides some basic understanding regarding the rating behavior of online users but also suggests some ways of computing better quality estimates. We can, for example, construct a feature-by-feature quality

Table III.
Feature by feature average standard deviation of *all* ratings, compared to the average standard deviation for ratings with *high* weights. In square brackets, the corresponding p-values for a positive difference between the two

City		O	R	S	C	V
	all	1.215	1.181	1.234	1.174	1.169
Boston	high	0.961	0.942	0.938	0.821	0.963
	p-val	[0.001]	[0.003]	[0.001]	[0.000]	[0.005]
	all	1.192	1.207	1.174	1.159	1.216
Las Vegas	high	0.929	0.976	1.083	1.010	0.955
	p-val	[0.000]	[0.000]	[0.084]	[0.009]	[0.000]
	all	1.160	1.199	1.194	1.120	1.163
New York	high	0.909	0.911	0.893	0.786	0.939
	p-val	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
	all	1.191	1.232	1.114	1.084	1.190
Sydney	high	0.970	0.885	1.238	0.968	1.024
	p-val	[0.000]	[0.000]	[0.943]	[0.070]	[0.021]

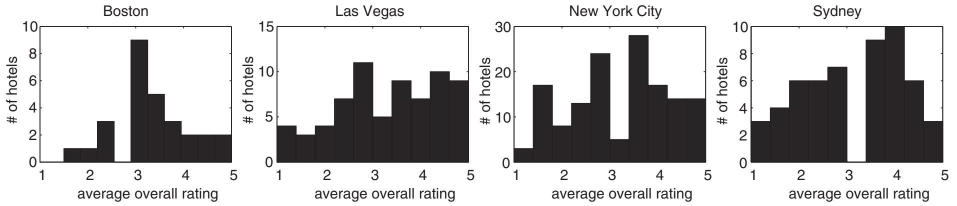


Fig. 3. The distribution of hotels depending on the average overall rating (only reviews corresponding to high weights).

estimate with much lower variance: for each feature we take the subset of reviews that amply discuss that feature, and output as a quality estimate the average rating for this subset. Initial experiments suggest that the average feature-by-feature ratings computed in this way are different from the average ratings computed on the whole data set.

The first objection one might raise against this method is that ratings corresponding to high weights are likely to come from passionate users and are therefore likely to have extreme values. The distribution plotted in Figure 2 supports this claim, as users who write detailed comments about the cleanliness of the hotel are mostly unhappy. Similarly for other hotels, one might expect that users who write a lot about a certain feature agree more, but only on extreme ratings (ratings of 1 or 5).

This, however, does not seem to be the case for the TripAdvisor dataset. As another experiment, we took all hotels from a given city, and for each hotel, h, we computed the average of all ratings $r_f^{i(h)}$ for the feature f, where the corresponding weight, $w_f^{i(h)}$, was high (i.e., belongs to the upper 20% of the weight range for that feature). Figure 3 plots the distribution of hotels for the four cities, depending on the average of the overall ratings corresponding to high weights. For Boston, the average overall rating from long reviews is almost normally distributed around 3.5. New York and Sydney have also the same shape. The former has two peaks around 2.5 and 3.5. In Las Vegas, on the

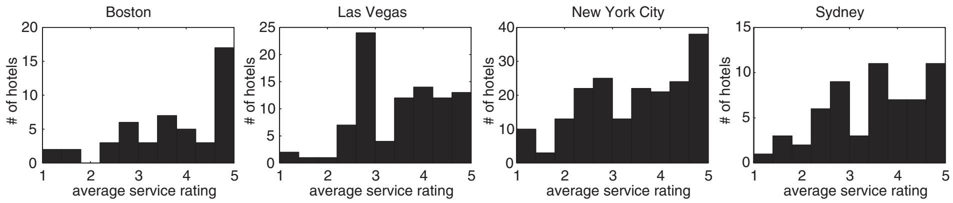


Fig. 4. The distribution of hotels depending on the average service rating (only reviews corresponding to high weights).

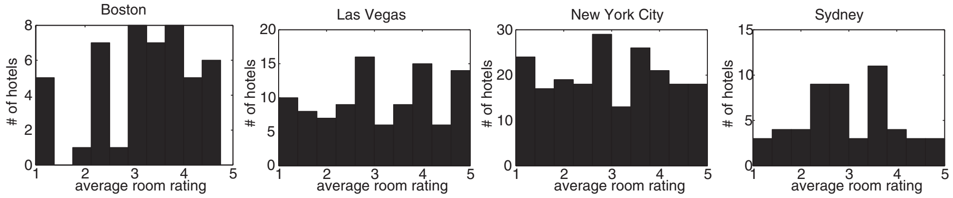


Fig. 5. The distribution of hotels depending on the average room rating (only reviews corresponding to high weights).

other hand, the average overall rating of long reviews seems almost uniformly distributed.

Similar patterns can be seen by analyzing the distribution of hotels depending on the average rating for the *Room* or *Service* (where, again, the average was made only for those reviews discussing a lot the value, respectively the service of the hotel). Figure 4 presents the distribution of hotels as a function of the average rating for *Service*, Figure 5 presents the similar graphs for the ratings on *Room*. In both cases, most of the hotels have an average rating between 2 and 4, which refutes the concern that the users who discuss amply the corresponding features have extreme opinions.

The second objection against building feature by feature estimators who average only the ratings corresponding to high weights, is that users who comment in more detail on a certain feature are not necessary experts of the domain. Consequently, their opinion should not count more than the opinion of other users.

This objection can partly be refuted by the following experiment. Tripadvisor lets users *vote* for the helpfulness of a review. By a simple click, a user can vote whether a particular review was useful or not. The system tracks the total number of votes received by every review, and displays bellow the review a footnote indicating how many of the total votes were positive.

Let the *score* of a review be the fraction of positive votes received by the review. This score can be regarded as an accurate estimator of the review's quality for the following two reasons. First, voting for a review requires almost no effort. The mouse click expressing a vote does not require authentication, and does not perturb the user when browsing for information (i.e., does not trigger a reload of the Web page). As opposed to users who write a full review, the users who vote do not need the internal benefits (e.g., extreme satisfaction or dissatisfaction) to compensate the cost of reporting. Moreover, the same user

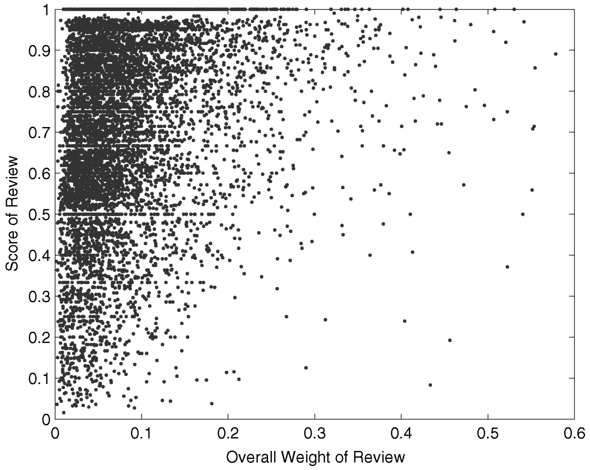


Fig. 6. The score of reviews plotted against their total length.

can easily vote for different reviews, expressing thus a relative order between reviews.

Second, the way the information is displayed (normal font size, below the main text of a review, without any visual pointers) makes it unlikely that the score of the review has an important influence on the final decision of a client. This also means that there probably are few incentives to falsely cast votes, which, together with the first point, leads to the conclusion that the score of a review probably reflects a representative opinion regarding the helpfulness of the review.

Figure 6 plots the score of the reviews in our data set against their weight for the feature *Overall* (which is actually proportional to the total length of the review). Every point in the figure represents a review. One can see that the long reviews tend to have high scores, so they were generally helpful to the other users. However, not all helpful reviews are also longer.

A third objection concerns our choice of defining the weight of a certain feature. The weights computed according to Equation (1) reflect the *relative* importance of certain features in the textual comment accompanying a review. It may well be that two different reviews have the same weight for feature f despite the fact that the comment of the first review mentions three words about f while the second reviewer contains three paragraphs discussing f. At first sight, relative weights are not correlated with the *effort* spent by the reviewer discussing a feature, and therefore cannot be intuitively considered as signals for identifying expert opinions.

We repeated the analysis reported in Table III, but with the weights reflecting the *absolute* length of the text discussing a feature. The results are presented in Table IV. This analysis supports the same conclusion: ratings associated with higher absolute weights have lower variance than the entire set of ratings. We chose relative weights instead of absolute weights because we believe they better reflect the distribution of the user's attention across different aspects.

Table IV.
Average standard deviation for *all* ratings, and average standard deviation for ratings with *high* weights. The weights reflect the *absolute length* of the text discussing a certain feature. The p-values for a positive difference between the two averages are displayed between square brackets. These values are very similar to the results reported in Table III, where the weights reflect the *relative fraction* of the text discussing a certain feature

City		R	S	C	V
Boston	all	1.168	1.196	1.215	1.281
	high	0.892	1.126	0.895	1.064
	p-val	[0.000]	[0.211]	[0.000]	[0.007]
Las Vegas	all	1.173	1.136	1.143	1.133
	high	0.848	1.166	0.980	0.962
	p-val	[0.000]	[0.681]	[0.006]	[0.004]
New York	all	1.176	1.185	1.182	1.165
	high	0.894	0.970	0.819	0.887
	p-val	[0.000]	[0.000]	[0.000]	[0.000]
Sydney	all	1.201	1.209	1.115	1.196
	high	0.880	1.072	0.894	1.038
	p-val	[0.000]	[0.042]	[0.003]	[0.018]

Finally, our analysis is limited to the factors we could observe directly from the TripAdvisor dataset. There can be other factors that are good proxies for *expert* reviews, like reviewer experience in specific categories, or established trust. Our analysis does not account for such parameters, and therefore can contain biased interpretations. A definitive validation of our hypothesis that high weights are good signals for *expert* opinions requires further experiments.

4. CORRELATION BETWEEN REPORTING EFFORT AND TRANSACTIONAL RISK

Another factor behind the feedback bias observed on Amazon is the relatively low value of the items offered in the market. Books and CD's are generally seen as very cheap, so the risk associated with buying a boring book or a bad CD is quite low. Feedback from previous clients decreases the information asymmetry of future buyers, and therefore their risk of choosing the wrong item. However, if the transaction poses little risk in the first place, the contribution of a feedback report to the further decrease of risk is so small, that it does not compensate the effort of reporting.[6]

Feedback reporting is not rational in such environments, since neither the reporter nor the community benefits from the costly action of reporting. Therefore, the feedback that gets submitted is probably a consequence of the strong emotional response of some users who strongly like or disliked the product. This internal emotional motivation can explain the bimodal, u-shaped distribution of ratings observed on Amazon [Hu et al. 2006].

On Amazon, the correlation between reporting incentives and perceived risk is difficult to analyze, since most books and CDs have comparable prices, and

[6] We thank Vincent Schickel-Zuber for this observation.

involve comparable risks. On TripAdvisor, on the other hand, some hotels can be an order of magnitude more expensive than others. Moreover, an inappropriate hotel may ruin a vacation, a business meeting, or a romantic weekend, so the risk of choosing a bad hotel probably exceeds by far the amount paid for the room. The right feedback in this context is very valuable to future users, as it can make the difference between a memorable trip and a dreadful one. This added value to the future travelers should motivate more users to contribute with feedback.

The data collected from TripAdvisor cannot reveal a positive correlation between the risk posed by a hotel and the motivation to submit feedback, since there is no way of estimating the actual percentage of users who left feedback for a particular hotel. However, the TripAdvisor data set can be used to study the correlation between the risk associated to a hotel and the effort spent by previous users in describing their experience with that hotel. Intuitively, we expect that the reviewers of high risk hotels spend more effort in submitting their reviews, as a consequence of feeling stronger motivated to share their experience.

Before presenting the results, let us explain our choices for measuring risk and effort. The effort spent in writing a review is measured by the length of the textual comment accompanying a review, and this choice is simpler to argument. The TripAdvisor feedback submission form is the same for everybody, so the difference between a fast and a careful review is given by the level of detail given in the textual comment. It is true that (i) some users can convey more information in shorter text, and (ii) shorter, concise English is harder to write than a long, sloppy text. Nevertheless, we expect that on the average a longer textual comment contains more information than a short one, and therefore signals more effort.

We will use as a measure for risk the official star-rating of the hotels. The intuition behind this choice is the following. A traveler who books a room in a one- or two-star hotel probably expects as little as a decent bed, a relatively clean room, and a minimum of assistance from the hotel's staff. Given the high competition and the strict hygiene legislation in Boston, Las Vegas, New York, or Sydney (the cities chosen for our study) any hotel is likely to meet these basic requirements, thus keeping the traveler happy. Therefore, the risk taken by the traveler in choosing a low-end hotel is minimum. A four- or five-star hotel, on the other hand, exposes a traveler to a much higher risk. The traveler chooses to pay more because she seeks certain qualities and services that were not offered by the cheaper alternative hotels. A business person, for example, might need reliable communication facilities, a concierge who can recommend good restaurants and arrange local transportation, or a fitness center to relax after a stressful day. These facilities are arguably essential to the success of the trip, and therefore, the business person has a lot to lose if she doesn't obtain them from the chosen hotel.

There are two important reasons for choosing the star-rating as a measure for risk, instead of the more straightforward information on price. First, the price information available on the TripAdvisor is not reliable. TripAdvisor is not a booking agency, and the price they quote is only informative, obtained

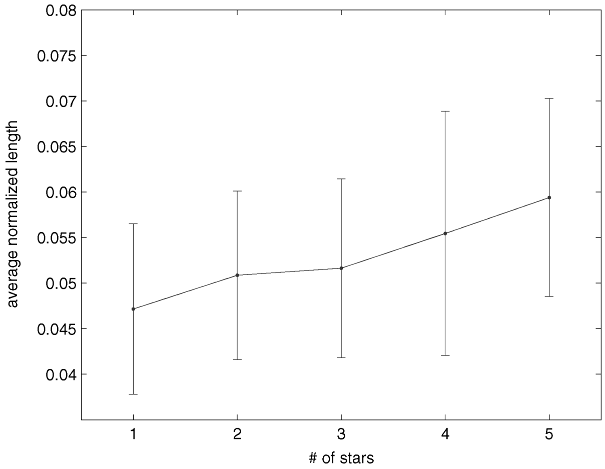


Fig. 7. The average normalized length of all textual comments, as a function of the star-rating.

by averaging prices quoted by their booking partners. The pricing structure of hotel rooms is very complex today (prices depend on the season, on the type of room, but also on the occupancy of the hotel), and is often subject to special offers and discounts. Therefore, the average displayed by TripAdvisor is probably a very poor estimate of what the users really paid for the room. A quick manual scan of some reviews suffices to reveal important differences between the price quoted by the users in their textual comments, and the average price displayed by TripAdvisor.

Star ratings, on the other hand, are fixed, and much easier to verify. Tourist offices or Yellow Pages directories have detailed lists of the hotels available in a certain city, together with their official star-ratings. Eventual errors in the TripAdvisor database are easy to spot by the users, and therefore, will probably be corrected. Again, a manual cross-checking with the information published by booking sites like Expedia or Travelocity will convince you that the star-rating recorded by TripAdvisor is probably correct.

Another problem with assessing hotel expectations based on price is that the price level depends on the location and is typically higher in large cities. Thus, it would require normalization over an unbiased sample of hotels, which is hard to obtain. The start-rating, on the other hand, provides a comparable ranking for all locations, since the distribution of hotels depending on the star rating tends to be the same: the risk of a four-star hotel is the same, whether in Boston, or in Las Vegas, because both cities offer a comparable choice of lower-star hotels.

To study the correlation between the effort spent for writing a review and the risk associated to a hotel, we conducted the following experiment. For each hotel h, we computed the average normalized length of the comments present in the reviews submitted about the hotel h. We then grouped all hotels depending on their star rating, and computed the mean, respectively the standard deviation, of the average review length. Figure 7 plots the two values as a function of the number of stars. Indeed, higher-rated hotels receive, on the average,

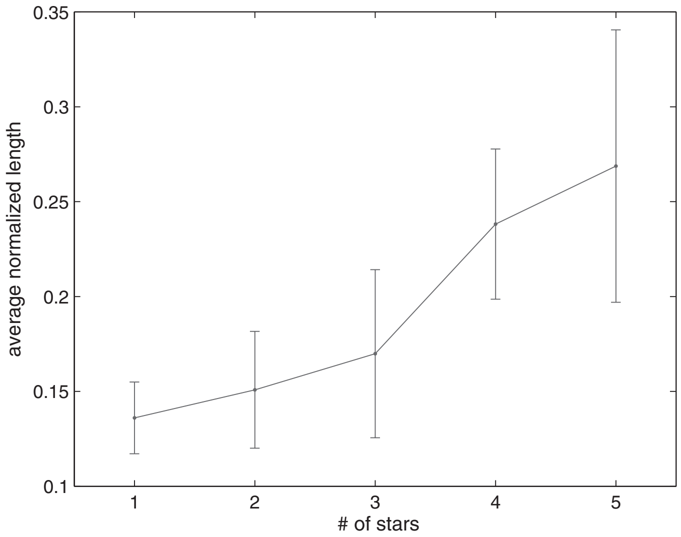


Fig. 8. The average normalized length of the 3 longest textual comments, as a function of the star-rating.

longer reviews, which supports our hypothesis that users spend more effort in reviewing the "riskier" hotels. However, due to the large variance of the review length, we cannot conclude statistically that reviews of x-star hotels are significantly[7] longer than the reviews submitted about $x + 1$-star hotels, where x can be one, two, three, or four. Nevertheless, when hotels are split into two risk groups:

- the *low*-risk hotels have one, two, or three stars;
- the *high*-risk hotels have four or five stars;

there is a significant increase in the length of reviews submitted about the high-risk hotels as compared to the length of the reviews about low-risk hotels. The p-value for the corresponding T-test is $1.8 \cdot 10^{-4}$.

Visually, the difference between the length of reviews received by high-risk and low-risk hotels can be better seen in Figure 8. Here, we only considered the three-longest reviews submitted about every hotel. The plot displays the average and the standard deviation of the 3 longest reviews about all hotels in a certain star category. Clearly, the most detailed reviews of four and five star hotels are significantly longer than the most detailed reviews submitted about the less-than-four-stars hotels.

It is also interesting to consider how the reviews address different aspects of quality depending on the star rating of the hotels. Figure 9 plots the average weight of the features *Cleanliness*, *Rooms*, *Value* and *Service* for hotels with different numbers of stars. The cleanliness, for example, is mostly discussed for the low-end hotels. Four- or five-star hotels are expected to be impeccably

[7] The T-tests were conducted at a 5% significance level.

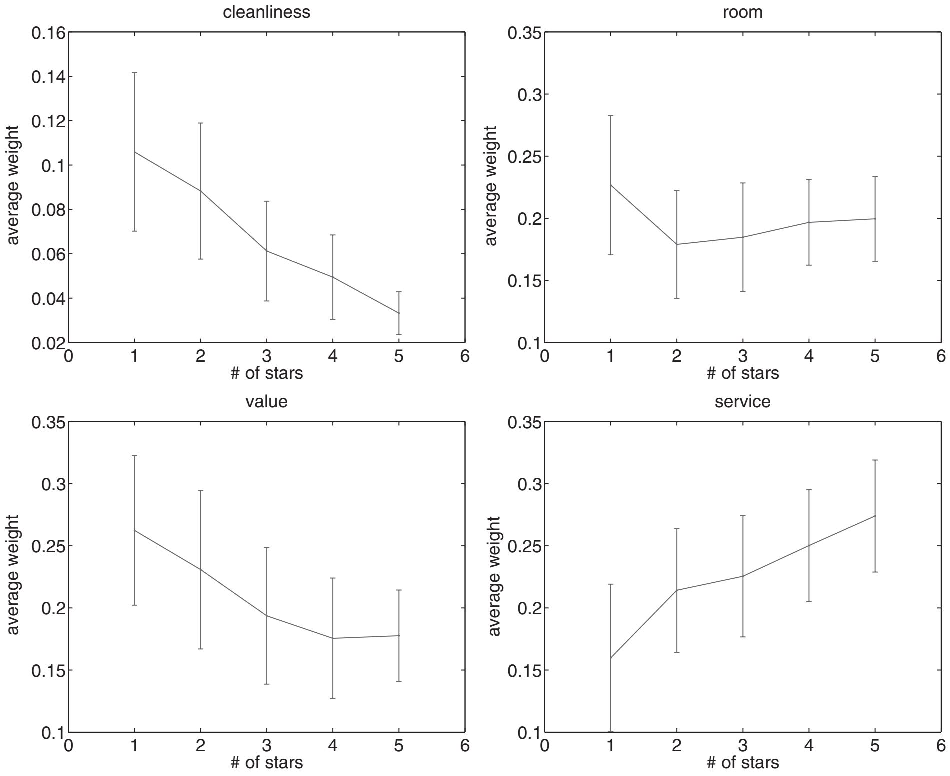


Fig. 9. The fraction of the comment taken by different features, as a function of the star-rating.

clean, and probably are very clean. The cleanliness is definitely not a risk factor for high-end hotel, hence reviewers spend little time discussing it. For low-end hotels, on the other hand, the cleanliness can be a major decision factor, and there is a significantly higher risk of choosing a dirty hotel. Hence the increased fraction of the comments addresses the cleanliness.

The same trend is observable for the fraction of the text discussing the value of the hotel, although overall, the value is discussed much more than the cleanliness. High-end hotels apparently have a very well established tradeoff between price and quality, which makes this feature a low risk factor. For low-end hotels, on the other side, there can be large variations between the value of different hotels, hence the presence of this feature in the reviews. The service, on the other hand, becomes increasingly important for high-end hotels. As argued previously, the travelers who chose fancier hotels do so because they need or like certain services that are not available in the cheaper hotels. The main risk associated to choosing four or five star hotels comes from not getting the services you want, hence naturally, the reviews for these hotels go into more detail regarding the services offered by the hotel.

The quality of the room, on the other hand does not seem to vary a lot depending on the star rating of the hotel. The rooms tend to be discussed more for one-star hotels, however, the difference is not statistically significant.

As a conclusion, one of the important motivation driving users to exert effort when submitting feedback is the desire to reduce the decisional risk of future users. The higher the risk associated to the lack of information on a particular aspect, the more valuable is the contribution of a feedback report, and hence the higher is the motivation of the reviewer to give more details. A natural extension of this observation is to conclude that feedback about risky transaction is not only more detailed, but also more frequent. A proper validation of this hypothesis requires controlled experiments and remains for future work.

Another question we hope to address in our future work is the exact correlation between the reporting effort and the value of the transaction. For all the reasons mentioned in the beginning of this section, we believe that the number of stars of a hotel can reasonably measure risk. Nevertheless, the ultimate measure of risk is the actual price paid by the user. Reliable information about prices would allow a detailed investigation of fine-grained incentives and biases, such as the surprise of a user who booked a high-end room at a bargain price, or the disappointment of a traveler who realizes she has paid for the name, not for the service of a hotel. Unfortunately, hotels and travel agencies are not willing to share actual and historic prices, making this investigation difficult to conduct.

5. THE INFLUENCE OF PAST RATINGS

Two important assumptions are generally made about reviews submitted to online forums. The first is that ratings truthfully reflect the quality observed by the users; the second is that reviews are independent from one another. Anecdotal evidence [Harmon 2004; White 1999] challenges the first assumption,[8] but the problem cannot be further investigated without controlled experiments that provide reliable data about the true experiences. In this section we address the second assumption, and find a dependence between the reviews of different users about the same hotel.

A perusal of online reviews shows that reviews are often part of discussion threads, where users make an effort to contradict, or vehemently agree with the remarks of previous users. Consider, for example, the following review:

> I don't understand the negative reviews... the hotel was a little dark, but that was the style. It was very artsy. Yes it was close to the freeway, but in my opinion the sound of an occasional loud car is better than hearing the 'ding ding' of slot machines all night! The staff on-hand is FABULOUS. The waitresses are great (and *** does not deserve the bad review she got, she was 100% attentive to us!), the bartenders are friendly and professional at the same time...

Here, the user was disturbed by previous negative reports, addressed these concerns, and set about trying to correct them. Not surprisingly, his ratings were considerably higher than the average ratings up to this point.

It seems that TripAdvisor users regularly read the reports submitted by previous users before booking a hotel, or before writing a review. Past reviews create some prior expectation regarding the quality of service, and this expectation

[8] Part of Amazon reviews were recognized as strategic posts by book authors or competitors.

has an influence on the submitted review. We believe this observation holds for most online forums. The subjective perception of quality is directly proportional to how well the actual experience meets the prior expectation, a fact confirmed by an important line of econometric and marketing research [Parasuraman et al. 1985, 1988; Olshavsky and Miller 1972; Teas 1993].

The correlation between the reviews has also been confirmed by recent research on the dynamics of online review forums [Forman et al. 2006].

5.1 Prior Expectations

We define the prior expectation of user i regarding the feature f, as the average of the previously available ratings on the feature f[9]:

$$e_f(i) = \frac{\sum_{j<i, r_f^j \neq 0} r_f^j}{\sum_{j<i, r_f^j \neq 0} 1}.$$

As a first hypothesis, we assert that the rating r_f^i is a function of the prior expectation $e_f(i)$:

HYPOTHESIS 2. *For a given hotel and feature, given the reviews i and j such that* $e_f(i)$ *is high and* $e_f(j)$ *is low, the rating* r_f^j *exceeds the rating* r_f^i.

We define *high* and *low* expectations as those that are above, respectively below a certain cutoff value θ. The set of reviews preceded by high, respectively low expectations are defined as follows:

$$R_f^{high} = \left\{ r_f^i | e_f(i) > \theta \right\}$$
$$R_f^{low} = \left\{ r_f^i | e_f(i) < \theta \right\}.$$

These sets are specific for each (hotel, feature) pair, and in our experiments we took $\theta = 4$. This rather high value is close to the average rating across all features across all hotels, and is justified by the fact that our data set contains mostly high quality hotels.

For each city, we take all hotels and compute the average ratings in the sets R_f^{high} and R_f^{low} (see Table V). The average rating amongst reviews following low prior expectations is significantly higher than the average rating following high expectations.

There are two ways to interpret the function $e_f(i)$:

—The expected value for feature f obtained by user i before his experience with the service, acquired by reading reports submitted by past users. In this case, an overly high value for $e_f(i)$ would drive the user to submit a negative report (or vice versa), stemming from the difference between the actual value of the service, and the inflated expectation of this value acquired before his experience.

—The expected value of feature f for all subsequent visitors of the site, if user i were not to submit a report. In this case, the motivation for a negative

[9] If no previous ratings were assigned for feature f, $e_f(i)$ is assigned a default value of 4.

Table V. Average Ratings for Reviews Preceded by Low and High Expectations. The P-values for a Positive Difference are Given within Square Brackets

City		O	R	S	C	V
Boston	low	3.733	3.889	3.923	4.092	3.763
	high	3.013	3.451	3.304	3.624	3.186
	p-val	[0.001]	[0.015]	[0.000]	[0.006]	[0.001]
Las Vegas	low	3.509	3.600	3.595	3.635	3.686
	high	3.125	3.412	3.248	3.420	3.397
	p-val	[0.001]	[0.086]	[0.005]	[0.086]	[0.011]
New York	low	3.771	3.713	3.807	3.992	3.793
	high	3.432	3.427	3.408	3.670	3.356
	p-val	[0.000]	[0.001]	[0.000]	[0.000]	[0.000]
Sydney	low	3.708	3.734	3.682	3.742	3.695
	high	3.407	3.185	3.117	3.300	3.104
	p-val	[0.018]	[0.001]	[0.000]	[0.008]	[0.000]

report following an overly high value of e_f is different: user i seeks to *correct* the expectation of future visitors to the site. Unlike the interpretation above, this does not require the user to derive an *a priori* expectation for the value of f.

Note that neither interpretation implies that the average up to report i is inversely related to the rating at report i. There might exist a measure of influence exerted by past reports that pushes the user behind report i to submit ratings which to some extent conforms with past reports: a low value for $e_f(i)$ can influence user i to submit a low rating for feature f because, for example, he fears that submitting a high rating will make him out to be a person with low standards.[10] This, at first, appears to contradict Hypothesis 2. However, this conformity rating cannot continue indefinitely: once the set of reports project a sufficiently deflated estimate for v_f, future reviewers with comparatively positive impressions will seek to correct this misconception.

5.2 Impact of Textual Comments on Quality Expectation

Further insight into the rating behavior of TripAdvisor users can be obtained by analyzing the relationship between the weights w_f and the values $e_f(i)$. In particular, we examine the following hypothesis:

HYPOTHESIS 3. *When a large proportion of the text of a review discusses a certain feature, the difference between the rating for that feature and the average rating up to that point tends to be large.*

The intuition behind this claim is that when the user is adamant about voicing his opinion regarding a certain feature, his opinion differs from the collective opinion of previous postings. This relies on the characteristic of reputation systems as feedback forums where a user is interested in projecting his

[10] The idea that negative reports can encourage further negative reporting has been suggested, before [Khopkar et al. 2005].

Table VI. Average of $|r^i_f - e_f(i)|$ when weights are high and low with P-values for the difference in sq. brackets

City		R	S	C	V
Boston	high	0.936	1.166	1.386	1.019
	low	0.679	0.790	0.772	0.758
	p-val	[0.000]	[0.000]	[0.000]	[0.130]
Las Vegas	high	0.892	1.164	1.133	1.083
	low	0.760	0.750	0.782	1.015
	p-val	[0.005]	[0.000]	[0.000]	[0.198]
New York	high	0.984	1.112	1.118	1.289
	low	0.694	0.749	0.778	0.926
	p-val	[0.000]	[0.000]	[0.000]	[0.000]
Sydney	high	1.008	1.151	1.229	0.860
	low	0.664	0.665	0.738	0.709
	p-val	[0.000]	[0.000]	[0.000]	[0.332]

opinion, with particular strength if this opinion differs from what he perceives to be the general opinion.

To test Hypothesis 3, we measure the average absolute difference between the expectation $e_f(i)$ and the rating r^i_f when the weight w^i_f is high, respectively low. Weights are classified high or low by comparing them with certain cutoff values: w^i_f is low if smaller than 0.1, while w^i_f is high if greater than θ_f. Different cutoff values were used for different features: $\theta_R = 0.4$, $\theta_S = 0.4$, $\theta_C = 0.2$, and $\theta_V = 0.7$. *Cleanliness* has a lower cutoff since it is a feature rarely discussed; *Value* has a high cutoff for the opposite reason. Results are presented in Table VI.

This demonstrates that when weights are unusually high, users tend to express an opinion that does not conform to the net average of previous ratings. As we might expect, for a feature that rarely was a high weight in the discussion, (e.g., cleanliness) the difference is particularly large. Even though the difference in the feature *Value* is quite large for Sydney, the P-value is high. This is because only few reviews discussed value heavily. The reason could be cultural or because there was less of a reason to discuss this feature.

6. AGGREGATING RATINGS INTO AVERAGES

One common use of rating information is to rank a list of alternatives (i.e., products or services like hotels or merchants) in the order of decreasing quality. Most systems in use today do this by taking the arithmetic mean of the ratings, thus minimizing the mean square error to the ratings. However, the results of the previous sections provide evidence that online reviews constitute a biased set of opinions, with distributions that do not respect the normality assumptions.

In Garcin et al. [2009] we investigate alternative ways of aggregating ratings using the *mean*, the *median* and the *mode* of the distributions. If ties occur when computing the *mode* or the *median* rating of a hotel, we choose the smaller of the two equivalent alternatives. A second tie-breaking rule is used to determine a unique ranking of the hotels in a given city. When two hotels have the same

Table VII. 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
median - mode	10.182	16.140	16.860	10.340	13.380

Table VIII.
Average number of outliers (with highest ratings 5) required to alter the ranking. In bold, the highest value

	Boston	Las Vegas	New York	Sydney	*average*
Mean	3.328	5.102	8.041	1.948	4.605
Median	**10.297**	**40.602**	**22.639**	3.639	**19.294**
Mode	9.047	23.867	22.309	**3.691**	14.729
p-value	0.000	0.000	0.000	0.000	

average (i.e., mean, median, respectively mode rating), we rank higher the hotel with a larger number of reviews.

A first objective of our study was to compare the final hotel rankings obtained by each aggregator. For each city, we computed three orderings of the hotels in that city according to each of the three aggregators. For each pair of orderings, we computed the average difference in rank of each hotel in the two orderings. These average differences are reported in Table VII, where, for example, the number in the upper left-most cell means that the rank of a hotel in Boston when aggregated by the median differs on average by 7.7 positions from its rank when aggregated by the mean. Likewise, the rank of a hotel in New York changes by an average of 16.9 positions 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. Considering that most feedback websites display only the first 5 or 10 "best" items, the results of Table VII 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.

We suggest that it is important to evaluate different rating aggregators with respect to their *robustness* to outlying reviews. A quick analytical exercise [Garcin et al. 2009] shows that the median and the mode are the most robust against strategic outliers that intentionally try to change the overall rating of a hotel.

Empirically, 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 VIII shows, for example, that the ranking under the mean can be changed with less than 5 reviews (on the average). The median and the mode, on the other hand, are much more robust and require 20, respectively 15 outliers before the ranking changes.

We also looked at how the rank of a hotel (computed according to different aggregators) evolves in time as new reviews arrive. Figure 10 shows this evolution for one hotel in New York, and clearly indicates that the ranking induced by the median or by the mode is much more stable than the ranking induced

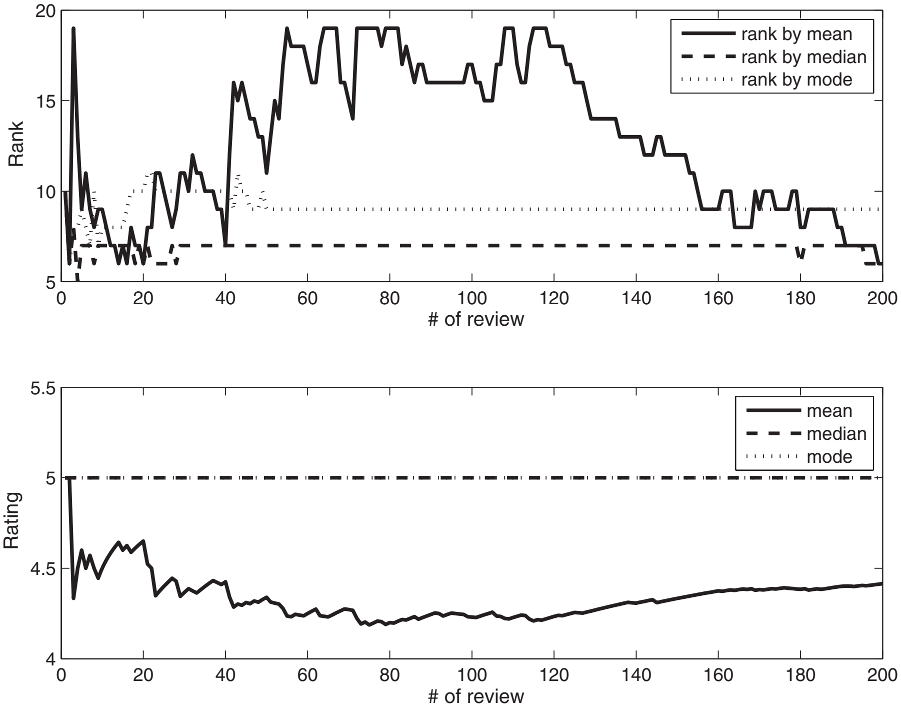


Fig. 10. The evolution of the rank and rating of a hotel in New York.

by the mean. A very similar trend can be observed for most hotels in our data set.

Note that the aggregator also has an influence on the incentives to the raters. A rater would like to bring the expressed average rating of the system as close as possible to a particular value, regardless of whether this value corresponds to an honest opinion. When the aggregate is formed by taking the arithmetic mean, it is usually in the rater's best interest to exaggerate its rating so that the average is moved as much as possible in the direction of the desired value. However, if the average is formed using the median, it is in the best interest of the user to report exactly the desired rating. Let the current average be r and let the rating the user desires be s. Assume without loss of generality that $r > s$. Provided that the current average is based on at least 3 ratings greater than s, any rating of s and below will have the same effect on the median, so there is no interest in reporting anything lower than s. On the other hand, reporting less than s could result in the average drifting below s at some later time, and this would not be in the interest of the rater. In fact, it can be shown [Moulin 1980] that the median is the only way of averaging that incentivizes raters to be truthful.

7. SUMMARY OF RESULTS AND CONCLUSIONS

The main goal of this paper is to advance our understanding of the factors that (i) drive a user to submit feedback, and (ii) bias the rating that a user provides to the reputation mechanism. For this purpose, we used two additional sources of information besides the vector of numerical ratings: the textual comments that accompany each rating, and the reports that have been previously submitted by other users.

Using straightforward natural language processing, we were able to establish a correlation between the weight of a certain feature in the textual comment accompanying the review, and the noise present in the numerical rating. Specifically, it seems that users who discuss a certain feature in detail are likely to agree on a common rating. This observation allows the construction of feature-by-feature estimators of quality that have a lower variance, and are hopefully less noisy. Initial experiments suggest that longer reviews tend to be more helpful to the other users, backing up the argument that reputation estimators should weigh more the corresponding ratings. Nevertheless, further evidence is required to support the intuition that at a feature level, high weight ratings are also more accurate, and therefore deserve higher priority when computing estimates of quality.

Using the same natural language processing of the textual comments associated to reviews, we were able to establish a correlation between the risk associated to a hotel and the effort spent in submitting the review. For the reasons detailed in Section 4 we assume that hotels with higher number of stars present a higher risk for the travelers, in terms of taking a bad decision. The average length of the reviews submitted about high-risk hotels is significantly bigger than the average length of low-end hotel reviews, meaning that users are willing to spend more effort when they perceive a higher risk of taking a bad decision. An immediate extension of this observation is that users will also be more motivated to submit feedback about high-risk transactions, however, we did not have the proper data to validate this assumption.

Second, we considered the dependence of ratings on previous reports. Previous reports create an expectation of quality which affects the subjective perception of the user. We validate two facts about the hotel reviews we collected from TripAdvisor: first, the ratings following low expectations (where the expectation is computed as the average of the previous reports) are likely to be higher than the ratings following high expectations. Intuitively, the perception of quality (and consequently the rating) depends on how well the actual experience of the user meets her expectation. Second, we include evidence from the textual comments, and find that users who devote a large fraction of the text to discussing a certain feature are likely to motivate a divergent rating (i.e., a rating that does not conform to the prior expectation). Intuitively, this supports the hypothesis that review forums act as discussion groups where users are keen on presenting and motivating their own opinion.

Naturally, a question that arises from this study is whether the observed biases can be corrected to obtain a better estimate of quality. One clear result is that a user can be given a more accurate estimate by weighting reviews according to how well the features discussed match those that are important to the user. Furthermore, aggregating ratings using the median rather than the arithmetic mean provides a more stable ranking that better informs users about the quality they can expect from a hotel. Furthermore, it eliminates incentives to manipulate the ranking by providing exaggerated ratings. If users can be made to understand this, they might provide more accurate ratings and thus increase the overall accuracy of review forums.

REFERENCES

Admati, A. and Pfleiderer, P. 2000. Noisytalk.com: Broadcasting opinions in a noisy environment. Working Paper 1670R, Stanford University.

Cui, H., Mittal, V., and Datar, M. 2006. Comparative experiments on sentiment classification for online product reviews. In *Proceedings of the National Conference on Artificial Intelligence*.

Dave, K., Lawrence, S., and Pennock, D. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International Conference on the World Wide Web (WWW03)*.

Dellarocas, C., Awad, N., and Zhang, X. 2006. Exploring the value of online product ratings in revenue forecasting: The case of motion pictures. Working paper.

Forman, C., Ghose, A., and Wiesenfeld, B. 2006. A multi-level examination of the impact of social identities on economic transactions in electronic markets. http://ssrn.com/abstract=918978.

Garcin, F., Faltings, B., and Jurca, R. 2009. Aggregating reputation Feedback. In *Proceedings of the International Conference on Reputation (ICORE)*. 119–128.

Ghose, A., Ipeirotis, P., and Sundararajan, A. 2005. Reputation premiums in electronic peer-to-peer markets: Analyzing textual feedback and network structure. In *Proceedings of the 3rd Workshop on Economics of Peer-to-Peer Systems (P2PECON)*.

Ghose, A., Ipeirotis, P., and Sundararajan, A. 2006. The dimensions of reputation in electronic markets. Working Paper CeDER-06-02, New York University.

Harmon, A. 2004. Amazon glitch unmasks war of reviewers. *New York Times*.

Houser, D. and Wooders, J. 2006. Reputation in auctions: Theory and evidence from eBay. *J. Econ. Manag. Strat. 15*, 353–369.

Hu, M. and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*.

Hu, N., Pavlou, P., and Zhang, J. 2006. Can online reviews reveal a product's true quality? In *Proceedings of the ACM Conference on Electronic Commerce (EC'06)*.

Kalyanam, K. and McIntyre, S. 2001. Return on reputation in online auction market. Working Paper 02/03-10-WP, Leavey School of Business, Santa Clara University.

Khopkar, T., Li, X., and Resnick, P. 2005. Self-selection, slipping, salvaging, slacking, and stoning: The impacts of negative feedback at eBay. In *Proceedings of the ACM Conference on Electronic Commerce (EC'05)*.

Melnik, M. and Alm, J. 2002. Does a seller's reputation matter? evidence from Ebay auctions. *J. Indust. Econ. 50,* 3, 337–350.

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

Olshavsky, R. and Miller, J. 1972. Consumer expectations, product performance and perceived product quality. *J. Market. Resear. 9*, 19–21.

Pang, B., Lee, L., and Vaithyanathan, S. 2002. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-02)*.

Parasuraman, A., Zeithaml, V., and Berry, L. 1985. A conceptual model of service quality and its implications for future research. *J. Market. 49*, 41–50.

Parasuraman, A., Zeithaml, V., and Berry, L. 1988. SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *J. Retail. 64*, 12–40.

Pavlou, P. and Dimoka, A. 2006. The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation. *Inform. Syst. Resear. 17,* 4, 392–414.

Popescu, A. and Etzioni, O. 2005. Extracting product features and opinions from reviews. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

TALWAR, A., JURCA, R., AND FALTINGS, B. 2007. Understanding user behavior in online feedback reporting. In *Proceedings of the ACM Conference on Electronic Commerce (EC'07)*.

TEAS, R. 1993. Expectations, performance evaluation, and consumers' perceptions of quality. *J. Market. 57*, 18–34.

WHITE, E. 1999. Chatting a singer up the pop charts. *Wall Street Journal*.

Received August 2007; revised August 2008, April 2009, December 2009; accepted January 2010