What If We Ask A Different Question? Social Inferences Create Product Ratings Faster

Eric Gilbert
School of Interactive Computing & GVU Center
Georgia Institute of Technology
gilbert@cc.gatech.edu

ABSTRACT

Consumer product reviews are the backbone of commerce online. Most commonly, sites ask users for their personal opinions on a product or service. I conjecture, however, that this traditional method of eliciting reviews often invites idiosyncratic viewpoints. In this paper, I present a statistical study examining the differences between traditionally elicited product ratings (i.e., "How do you rate this product?") and *social inference ratings* (i.e., "How do you think other people will rate this product?"). In 5 of 6 trials, I find that social inference ratings produce the same aggregate product rating as the one produced via traditionally elicited ratings. In all cases, however, social inferences yield less *variance*. This is significant because using social inference ratings 1) therefore converges on the true aggregate product rating faster, and 2) is a cheap design intervention on the part of existing sites.

Author Keywords

product reviews; ratings; ecommerce; social psychology

ACM Classification Keywords

H.5.3 [Group and Organization Interfaces]: Asynchronous interaction - Web-based interaction;

INTRODUCTION

Consumer product reviews fundamentally shape commerce online. From broad retail sites like Amazon to niche marketplaces like Etsy to the underground Silk Road, product ratings are the building blocks of search, recommendations and purchase decisions [1, 2, 3, 4, 5, 7, 9, 16]. The importance of product ratings is evident in the lengths to which sellers sometimes go to manipulate them, leading to an entire thread of research on detecting spurious reviews (e.g., [11]).

Most commonly, sites ask people for their personal opinions on a product or service. On Amazon, for example, the site asks its users, "How do you rate this item?" Among other problems—such as generating bimodal ratings distributions [15]—this practice can lead to particularly idiosyncratic reviews. The blog Least Helpful humorously catalogs a number of them. In one case, a reviewer gives the classic novel *Animal Farm* only one star because the reviewer "never enjoyed books or movies where animals talk"[1]. In another case, a user poorly rates a set of measuring spoons because his ferret ate them and ended up needing a trip to the hospital[2].

These are extreme examples, of course. At the same time, I conjecture that the traditional method of eliciting reviews often invites idiosyncratic viewpoints. This may be fine when the objective is to personalize a user's own experience with a service (i.e., Netflix's personalization engine). However, in cases where we rely on the collective wisdom of other people to inform our future decisions—such as Amazon star ratings, Reddit upvotes and Etsy seller ratings, etc.—designers truly want to know *how much other people will like it.* In a recent, foundational piece of work, Shaw et al. [14] demonstrate that among a number of treatment conditions, the Bayesian Truth Serum approach [12] of asking people to prospectively consider peer ratings significantly improved the accuracy of inexpert ratings. This is non-obvious; in many contexts, reviewers may misjudge established norms, a concept known in the social sciences as pluralistic ignorance [10, 13].

This paper builds on the results of Shaw et al. In the context of product reviews, I consider not only accuracy, but also *variance*. I present a statistical study built on Mechanical Turk respondents examining the differences between traditionally elicited product ratings (i.e., "How do you rate this item?") and *social inference ratings* (i.e., "How do you think other people will rate this item?"). I find, like Shaw et al., that social inferences usually produce an accurate answer; in 5 of the 6 items I compared, social inference ratings produce statistically indistinguishable mean product ratings. The novel contribution of this work is that I show that social inferences substantially reduce variance.

This is noteworthy for two reasons. First, because social inference ratings originate from a distribution with considerably less variance (i.e., it has fewer outliers), you need less of them to converge on an aggregate product rating. The majority of products receive relatively few ratings [8], so efficiently using the ones you already have is a primary concern. Second, since designers can easily control the way they ask for ratings—a

CHI 2014, April 26–May 1, 2014, Toronto, Ontario, Canada.

ACM 978-1-4503-2473-1/14/04..$15.00.
http://dx.doi.org/10.1145/2556288.2557081

[1] http://goo.gl/yqwjFb
[2] http://goo.gl/eCpYmm

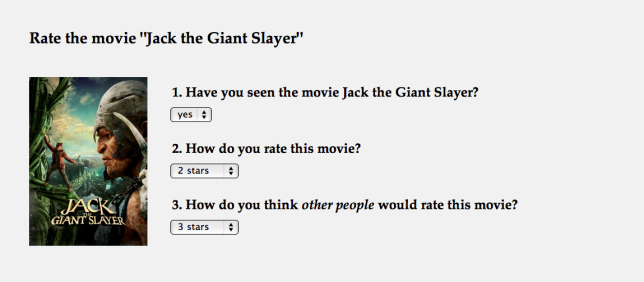

Figure 1. Screenshot of experimental task, as Mechanical Turk workers experienced it.

change only requires a quick edit of an HTML page—social inferences are a cheap design intervention.

Next, I outline my experiment, the data it generated, and the statistical bootstrapping technique I used to test for a reduction in variance across the products reviewed in this study. I conclude the paper by reflecting on the theoretical and design implications raised by this work.

EXPERIMENT

I designed an online experiment to examine variance in the context of product reviews. The experimental task asked participants to give movies their individual ratings, and to also prospectively infer the likely ratings of others. Movies were chosen as the product category for this experiment to make recruiting participants more tractable. (The intuition was that more participants would be able to rate popular movies than, for instance, an obscure kitchen utensil.)

I recruited participants from Amazon's Mechanical Turk, paying them each $0.25 to complete my task. For each of six movies, I asked 100 Turkers to provide ratings of it. I restricted the participant pool to include only those workers with a "HIT Approval Rate" above 90% to filter for reliable, well-regarded Turkers. (The HIT Approval Rate is the percentage of times a requester has marked a Turker's work as acceptable.)

I selected six recent, popular movies from Amazon's list of bestselling movies, a list Amazon updates hourly. The first six movies from this list with more than 100 customer reviews were chosen for the experimental task: *Identity Thief*, *Jack the Giant Slayer*, *Silver Linings Playbook*, *The Hunger Games*, *What to Expect When You're Expecting* and *This Is 40*. While not explicitly targeted (i.e., the strategy here is random sampling of an ever-changing list), these movies exhibit a span of average Amazon ratings (3.0 – 4.2), as well as a range of quality (Rotten Tomatoes scores: 20% – 92%; IMDB ratings: 5.5 – 7.9). I discuss further the implications and limitations of this particular sample of movies in the *Discussion*, as this set does differ from the broader universe of products.

For each movie, I asked participants three questions:

1. Have you seen the movie [movie title]?
2. How do you rate this movie?
3. How do you think other people would rate this movie?

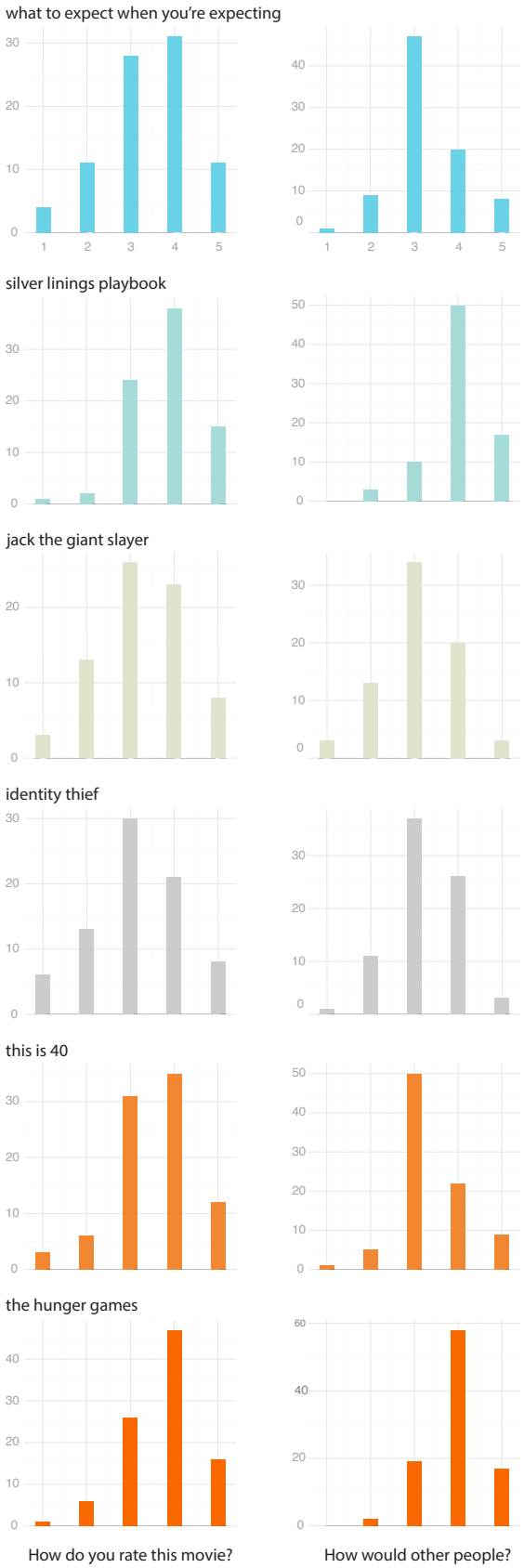

Figure 2. Ratings distributions across all six movies, by question.

To ease recall, the page Turkers interacted with also included the same image used by Amazon to promote the movie. The first question allows us to filter out anyone who reports they did not see the movie. (Some Turkers did respond "no" to this question.) As discussed earlier, the second question, "How do you rate this movie?", mirrors the language Amazon uses to elicit ratings from its customers. For questions two and three, Turkers could pick a rating from the range "1 star" to "5 stars"—again mimicking how Amazon (and many other sites) quantify consumer sentiment.

I left question three, "How do you think other people would rate this movie?", intentionally vague to permit participants to imagine anyone relevant. However, one might consider asking about "people like you," "other people interested in this movie" or "other knowledgable people," instead of simply "other people." I had no theoretical material pointing toward any of these directions, and therefore asked the higher-level, but also vaguer "other people" variant.

Difficulty of eliciting social inferences

In a parallel experiment, I asked Turkers how difficult they found providing social inference ratings, as compared to the standard individual ratings. A separate experiment was performed to avoid contaminating the main objective of this work—experimenting with variance-reduction approaches—by framing the tasks in participants' minds as difficult.

To measure how difficult of a time participants had providing social inference ratings, I adapted the well-known NASA-TLX workload inventory [6]. In this phase, in addition to the three questions above, I asked Turkers two other questions:

4. How mentally demanding was providing your own individual rating?
5. How mentally demanding was providing the rating you think other people would give?

If any benefit accompanies eliciting social inferences from product reviewers, it would be offset by any cost associated with doing so. These two questions aim to capture that difference in reviewer workload.

Participants

600 participants were recruited from Amazon's Mechanical Turk. After cleaning the response data for people who responded in too little time, and after removing those Turkers who provided ratings but also claimed not to have seen the movie, 93 participants were removed. Thus, the dataset contains the individual ratings and social inference ratings of 507 people. I did not ask participants for demographic information, as I didn't see it as particularly relevant for this experimental task.

RESULTS

The main statistical question in this work is whether social inference ratings differ in variance from traditionally elicited ratings, alongside secondary questions of accuracy and difficulty. Since these data do not conform to normal distributions, I turn to nonparametric statistical methods to compare the two questions: the Wilcoxon test for central tendencies, Levene's test to assess variance homogeneity, and a bootstrapping technique I discuss shortly.

Movie	$\mu_t - \mu_s$	W	p	$\sigma_t - \sigma_s$	F	p
Identity	-0.09	521	0.34	0.28	6	0.02
Jack	0.17	435	0.18	0.14	3.7	0.06
Silver	-0.21	366	0.04	0.11	5.6	0.02
Hunger	-0.2	408	0.01	0.15	7.9	0.01
Expect	0.11	391	0.16	0.19	7.8	0.006
40	0.16	491	0.05	0.14	4.2	0.04

Table 1. Pairwise comparisons between movies' traditional ratings and their social inference ratings. μ_t and σ_t refer to traditionally elicited ratings; μ_s and σ_s refer to social inferences.

Table 1 summarizes my pairwise findings. Where the means differ (i.e., *Silver Linings Playbook*, *The Hunger Games*), social inference ratings produce a higher mean (3.77 vs. 3.98, and 3.71 vs. 3.91, respectively). I include this table to suggest the pairwise difference at the movie level. However, in this context, the movies themselves simply represent a grouping of a pseudorandom sample—not an independent variable in their own right.[3] Rather, we have a single hypothesis (e.g., whether variance differs across questions) to test in the context of independent experiments (i.e., the six movies, which I will assume not to have statistical dependencies). Hence, family-wise error rate controls such as Bonferroni corrections are inappropriate in this specific context.

Without a ready-made, omnibus statistical technique on which to rely, I turn to bootstrapping to estimate whether a reduction in variance exists across all the movies. Let σ_{jk} be the standard deviation for the j-th movie and the k-th question, with $j = 1, \ldots, 6$ and $k = 1, 2$[4]. We can reformulate this paper's main hypothesis as testing the null hypothesis $H_0 : \exists j : \sigma_{j1} = \sigma_{j2}$. To test H_0, we bootstrap the statistic j, the number j such that $\hat{\sigma}_{j1} > \hat{\sigma}_{j2}$; that is, the j movies whose traditional rating variance is greater than its social inference variance. Via a Monte Carlo simulation comprising 1M resamplings, I find with 95% confidence that $j \in (5.214, 6]$. As j must be an integer, therefore $j = 6$. In summary, across all movies in this experiment, we see a significant reduction in variance by asking for social inference ratings. Finally, using a similar method, I find no difference in difficulty across the two question types, with $j = 0$.

DISCUSSION

Eliciting product ratings as social inferences produces (arguably) the same product ratings, but with considerably less variance, and without increasing a reviewer's mental workload. Speaking in visual terms, this means that the ratings distributions in the left column of Figure 2 are wider, while the ones in the right column are thinner. Where the means do differ (i.e., *Silver Linings Playbook*, *The Hunger Games* in Table 1), social inference ratings produce a mean rating closer to the actual Amazon product rating than the traditional "How do you rate this movie?" In the case of *The Hunger Games*,

[3] One could imagine in the future constructing a 2x2 design, where product type (i.e., comedy, drama, action) is its own object of study.
[4] In other words, σ_{41} would represent the standard deviation of *Identity Thief's* traditionally elicited ratings, while σ_{42} would represent the standard deviation of its social inferences.

for example, 6,245 Amazon reviewers have given the movie an aggregate rating of 4.1 stars. In this sample, Turkers gave the movie a social inference rating of 3.91, but only a mean rating of 3.71 via the customary variant. Based on these data, I conclude that social inference ratings can produce the same product ratings as traditionally elicited ones, only with less noise.

I believe this is a finding with widespread applicability. Sites all over the internet ask users to rate all manner of things: products, experiences, posts, comments, news stories, etc. Next, I review some of the theoretical social computing questions raised by this work. I also offer designers guidance on what these results could mean for the way sites elicit feedback from users about products.

Theoretical Implications

I believe this work raises a number of theoretical points and questions for social computing. First and foremost, as ratings pervade the internet, we need to better understand how social inference ratings work in different contexts. Do we observe the same effects for different objects? Perhaps it is impossible to replicate these findings within the context of user-generated comments on a site like Reddit, for example. As the conclusions presented here rely on only a small segment of the ratings space, we need further exploration to generalize to the broader space of products.

Moving to the operational: Do we need to ask both questions? Does asking one influence the other? Should the question ask about vague "other people" or some more-targeted variant, such as "people you know" or "your friends?" We know from various anchoring and ordering effects studies that subtle framing differences like these can alter results substantially.

Design Implications

Designers have full control over the questions they use to elicit ratings. One reason I find these results so compelling is that an intervention based on them requires such a simple change on the part of existing sites. That is, simply add a label and a drop down list to the page that asks users for the product ratings.

While some "superstar" products attract many thousands of reviews, because of the nature of how social data is often distributed [8], the majority of products receive relatively few ratings. Using them most efficiently is therefore a primary concern. By reducing variance (i.e., subjecting means to fewer outliers) in the ratings distribution, designers can converge on the true, eventual mean product ratings faster than with the typical question. While the present research only offers preliminary evidence, the impact to today's social sites could be widespread and significant.

Limitations and Future Work

I base these findings on a sample drawn from Mechanical Turk. While convenient, it remains unclear how different this sample is from the broader population of product reviewers; an intervention study on Amazon, for instance, would help make this clear. Moreover, for tractability, I chose movies as my object of study. It also remains unclear how products used differently (e.g., ones used over and over) might affect eliciting social inference ratings.

REFERENCES

1. N. Archak, A. Ghose, and P. G. Ipeirotis. Show me the money!: deriving the pricing power of product features by mining consumer reviews. In *Proc. KDD*, pages 56–65, 2007.
2. N. Christin. Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In *Proc. WWW*, pages 213–224, 2013.
3. C. Dellarocas, X. M. Zhang, and N. F. Awad. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive marketing*, 21(4):23–45, 2007.
4. A. Ghose and P. Ipeirotis. The economining project at nyu: Studying the economic value of user-generated content on the internet. *Journal of Revenue & Pricing Management*, 8(2):241–246, 2009.
5. E. Gilbert and K. Karahalios. Understanding deja reviewers. In *Proc. CSCW*, pages 225–228, 2010.
6. S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human mental workload*, 1(3):139–183, 1988.
7. J. A. Konstan and J. Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, 2012.
8. J. Laherrere and D. Sornette. Stretched exponential distributions in nature and economy:"fat tails" with characteristic scales. *The European Physical Journal B-Condensed Matter and Complex Systems*, 2(4):525–539, 1998.
9. Y. Liu, X. Huang, A. An, and X. Yu. Modeling and predicting the helpfulness of online reviews. In *Proc. ICDM*, pages 443–452, 2008.
10. D. T. Miller and C. McFarland. Pluralistic ignorance: When similarity is interpreted as dissimilarity. *Journal of Personality and social Psychology*, 53(2):298, 1987.
11. M. Ott, C. Cardie, and J. Hancock. Estimating the prevalence of deception in online review communities. In *Proc. WWW*, pages 201–210, 2012.
12. D. Prelec. A bayesian truth serum for subjective data. *Science*, 306(5695):462–466, 2004.
13. D. A. Prentice and D. T. Miller. Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *Journal of personality and social psychology*, 64(2):243, 1993.
14. A. D. Shaw, J. J. Horton, and D. L. Chen. Designing incentives for inexpert human raters. In *Proc. CSCW*, pages 275–284, 2011.
15. A. Talwar, R. Jurca, and B. Faltings. Understanding user behavior in online feedback reporting. In *Proc. EC*, pages 134–142, 2007.
16. Q. Ye, R. Law, and B. Gu. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180–182, 2009.