WINTER 2014 VOL.55 NO.2

MITSloan
Management Review

The Problem With Online Ratings

By Sinan Aral

REPRINT NUMBER 55224

Online, social influence can bias ratings.

The Problem With Online Ratings

Studies show that online ratings are one of the most trusted sources of consumer confidence in e-commerce decisions. But recent research suggests that they are systematically biased and easily manipulated.

BY SINAN ARAL

THE LEADING QUESTION
How reliable are consumer-generated online ratings?

FINDINGS

- Social influence can lead to disproportionately positive online ratings.
- Subsequent raters are more likely to be influenced by previous positive ratings than by negative ones.
- Take positive online ratings with a grain of salt.

A FEW MONTHS AGO, I stopped in for a quick bite to eat at Dojo, a restaurant in New York City's Greenwich Village. I had an idea of what I thought of the place. Of course I did — I ate there and experienced it for myself. The food was okay. The service was okay. On average, it was average.

So I went to rate the restaurant on Yelp with a strong idea of the star rating I would give it. I logged in, navigated to the page and clicked the button to write the review. I saw that, immediately to the right of where I would "click to rate," a Yelp user named Shar H. was waxing poetic about Dojo's "fresh and amazing, sweet and tart ginger dressing" — right under her bright red five-star rating.

I couldn't help but be moved. I had thought the place deserved a three, but Shar had a point: As she put it, "the prices here are amazing!"

Her review moved me. And I gave the place a four.

As it turns out, my behavior is not uncommon. In fact, this type of social influence is dramatically biasing online ratings — one of the most trusted sources of consumer confidence in e-commerce decisions.

The Problem: Our Herd Instincts

In the digital age, we are inundated by other people's opinions. We browse books on Amazon with awareness of how other customers liked (or disliked) a particular tome. On Expedia, we compare hotels based on user ratings. On YouTube, we can check out a video's thumbs-up/thumbs-down score to help determine if it's worth our time. We may even make serious decisions about medical professionals based in part on the feedback of prior patients.

For the most part, we have faith in these ratings and view them as trustworthy. A 2012 Nielsen report surveying more than 28,000 Internet users in 56 countries found that online consumer reviews are the second most-trusted source of brand information (after recommendations from friends and family).[1] According to the survey, more than two-thirds of global customers say they trust messages on these platforms — a 15% increase in four years.

But this trust may be misplaced. The heart of the problem lies with our herd instincts — natural human impulses characterized by a lack of individual decision making — that cause us to think and act in the same way as other people around us.[2] On two different days in April 2013, for instance, the price of gold fell more than it had in three decades. At the time, market watchers offered all sorts of justifications as to why the metal's price plunged so precipitously, but none of them was particularly compelling. "It is hard to escape the conclusion that gold investors sold off because other investors were selling — in other words, herd instinct kicked in," wrote Sarah Gordon of the Financial Times.[3]

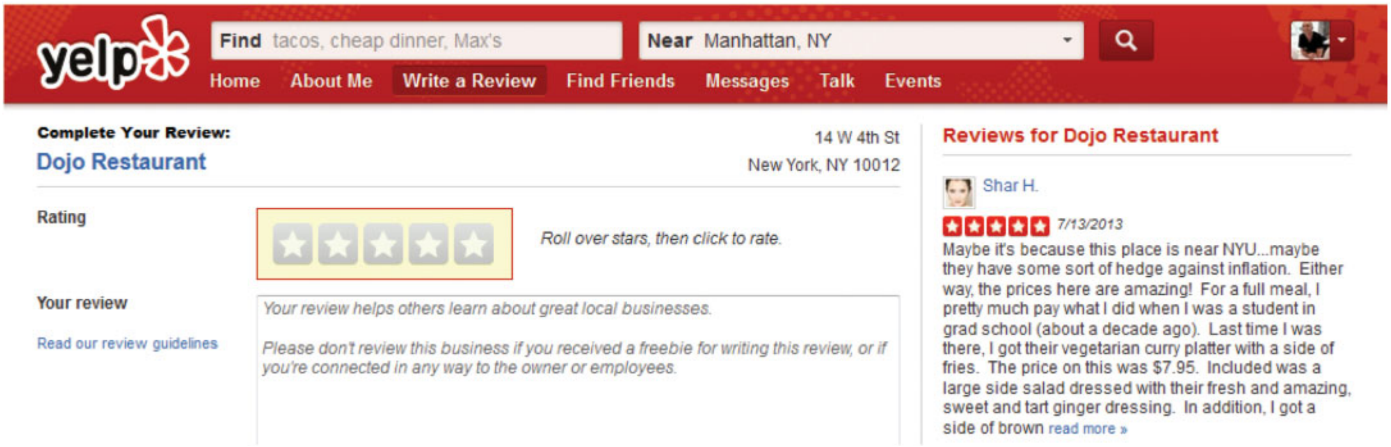
Social Influence Bias

When it comes to online ratings, our herd instincts combine with our susceptibility to positive "social influence."[4] When we see that other people have appreciated a certain book, enjoyed a hotel or restaurant or liked a particular doctor — and rewarded them with a high online rating — this can cause us to feel the same positive feelings about the book, hotel, restaurant or doctor and to likewise provide a similarly high online rating.

Recently, my colleagues Lev Muchnik, a senior lecturer at the Hebrew University of Jerusalem's School of Business Administration, and Sean J. Taylor, a doctoral student at New York University's Stern School of Business, and I designed a simple randomized experiment on a social news-aggregation website.[5] (For more details about our experiment, please see "About the Research," p. 50.) On the site, users rate news articles and comments by voting them up or down based on how much

AN EXAMPLE OF SOCIAL INFLUENCE

I had planned to give Dojo Restaurant a three-star review until I saw other raters' glowing reviews, such as this one from Shar H. I ended up giving the restaurant four stars instead of three in my review.



Complete Your Review:
Dojo Restaurant
14 W 4th St
New York, NY 10012

Roll over stars, then click to rate.

Your review helps others learn about great local businesses.

Please don't review this business if you received a freebie for writing this review, or if you're connected in any way to the owner or employees.

Reviews for Dojo Restaurant

Shar H.
★★★★★ 7/13/2013
Maybe it's because this place is near NYU...maybe they have some sort of hedge against inflation. Either way, the prices here are amazing! For a full meal, I pretty much pay what I did when I was a student in grad school (about a decade ago). Last time I was there, I got their vegetarian curry platter with a side of fries. The price on this was $7.95. Included was a large side salad dressed with their fresh and amazing, sweet and tart ginger dressing. In addition, I got a side of brown read more »

they enjoyed them. We randomly manipulated the scores of comments with a single up or down vote. Up indicated the "user" enjoyed the comment; down indicated the "user" didn't. We then measured the impact of these small manipulations on subsequent scores.

The results were alarming. The positive manipulations created a positive social influence bias that persisted over five months and that ultimately increased the comments' final ratings by 25%. Negatively manipulated scores, meanwhile, were offset by a correction effect that neutralized the manipulation: Although viewers of negatively manipulated comments were more likely to vote negative (evidence of negative herding), they were even more likely to positively "correct" what they saw as an undeserved negative score.

This social influence bias snowballs into disproportionately high scores, creating a tendency toward positive ratings bubbles. Positively manipulated scores were 30% more likely than control comments (the comments that we did not manipulate) to reach or exceed a score of 10. And reaching a score of 10 was no small feat; the mean rating on the site is 1.9. A positive vote didn't just affect the mean of the ratings distribution; it pushed the upper tail of the distribution out as well, meaning a single positive vote at the beginning could propel comments to ratings stardom.

Ratings Bubbles

These findings could help explain the online ratings bubbles recently observed by several different research teams, which some scientists have described as the "J-shaped distribution" of online ratings. It turns out that online ratings tend to be disproportionately positive. The distributions of product ratings on Amazon.com include far more extreme positive (five-star) than negative (one-star or two-star) or generally positive (three-star or four-star) reviews. Trends toward positivity have also been observed in restaurant ratings and movie and book reviews on a variety of different websites.[6]

The social influence bias that we observed in our experiment helps to explain these bubbles. If social influence creates positive herding but not negative herding, ratings bubbles could be caused by an asymmetry in our cognitive biases toward the prior positive opinions of others: We tend to herd on positive opinions and remain skeptical of the negative ones.

In one study that examined the skewed distribution of online ratings, researchers Nan Hu, Paul Pavlou and Jennifer Zhang also conducted a small side experiment.[7] They invited students to a lab to rate a single music CD selected at random from Amazon and compared the resulting ratings to the ratings of the same CD on the Amazon site. They did this to see if the distribution of (an albeit small) random sample of actual opinions about this item (66 students in a university lab) matched the distribution of ratings given on Amazon. What they found was puzzling: The ratings from their experiment were approximately normally distributed, like a standard bell curve, cresting in the middle (reflecting the higher frequency of two-star, three-star and four-star reviews) and sinking at the extremes (reflecting the comparable paucity of one-star and five-star reviews). Meanwhile, the distribution of ratings on Amazon for the same item followed the J-shape (with the frequency of five-star reviews more than doubling that of one-star, two-star, three-star and four-star reviews).

The authors interpreted these findings as evidence that Amazon's buyers are more likely to be positively predisposed to a product because they had voluntarily purchased it, creating a selection bias toward more positive ratings. Selection bias is a potentially good explanation for the J shape (if reviews come from purchasers and if purchasers are, indeed, positively predisposed). But here's the catch: Amazon does not require users to buy items before rating them. So I wondered: Were prior ratings in this experiment shown to raters before they rated? The paper makes no mention of this aspect of the experimental setup. I wrote to the authors and asked them whether prior ratings were visible to users during the rating process. They replied that they were not. Examining social influence bias was not part of their study. In other words, the simulated environment they had created mimicked Amazon's interface — with one crucial difference: The raters did not see the distribution of prior ratings, or any information on prior ratings for that matter, before they rated any item.

In this context, I reconsidered the four-star rating

ABOUT THE RESEARCH

In decision making, our society is increasingly relying on the digitized, aggregated opinions of others. Online ratings provide one important example. Our understanding of the impact of this type of social influence on collective judgment is limited, however, because distinguishing influence from uninfluenced agreement on true quality is nearly impossible without randomized experimentation. For example, popular products may be popular because of the irrational effect of past positive ratings, or alternatively the best products may become popular because they are of the highest quality. Unfortunately, very few large-scale randomized experiments examine herding effects in society.[i]

We therefore conducted such an experiment to quantify the effects of social influence on users' ratings and discourse on a social news aggregation website similar to reddit.com. (Detailed results of our study were reported in an article in Science.[ii]) Users of the site write comments in response to posted articles, and other users can then "up-vote" or "down-vote" these comments, yielding an aggregate current rating for each posted comment equal to the number of up-votes minus the number of down-votes.

Over five months, 101,281 comments submitted on the site were randomly assigned to one of three treatment groups: up-voted, down-voted or control. Up-voted comments were artificially given an up vote (a +1 rating) upon the comment's creation, while down-voted comments were given a down vote (a –1 rating) upon the comment's creation. As a result of the randomization, comments in the control and treatment groups were identical in expectation along all dimensions that could affect users' rating behavior except for the current rating. This manipulation created a small random signal of positive or negative judgment by prior raters for randomly selected comments.

We observed the trajectory of the control and treatment group comments' scores over a period of five months to estimate how prior ratings were biasing future ratings. The 101,281 experimental comments were viewed more than 10 million times and rated 308,515 times by subsequent users. The experiment uncovered five key results:

1. The positive manipulation increased the likelihood of positive ratings by 32% and created accumulating positive herding that increased final ratings by 25% on average.

2. Positively treated comments were also significantly more likely than those in the control group to accumulate exceptionally high scores. Up-treated comments were 30% more likely to reach or exceed a score of 10. (The mean rating on the site is 1.9.)

3. Positive social influence and negative social influence created asymmetric herding effects. Negatively treated comments received down votes with a significantly higher probability than control comments. But this effect was offset by a larger correction effect, whereby these comments were up-voted with a significantly higher probability than were the control comments. This correction neutralized social influence in the ratings of negatively manipulated comments.

4. Herding effects and ratings bubbles varied by topic, implying that some product or service categories are more susceptible to social influence bias than others. While comments on business, culture and society, and politics were highly susceptible to popularity bubbles, those in general news, IT, economics and fun were not.

I had given Dojo. Would I have given four stars had I not seen the glowing reviews of other raters? I was preparing to give Dojo three stars before I saw those other reviews. I wondered: What if everyone's reviews — not just mine — were being swayed in a positive direction as a result of social influence bias?

Many factors could be creating the J-shaped distribution of online ratings: selection bias, fraud or social influence bias. These processes may also be working in tandem to create ratings bubbles. But the key concern raised by social influence bias, in particular, is that it creates a runaway bandwagon effect: The impact of one fake positive review is not compartmentalized. Instead, it dramatically affects future ratings. That fact, on its own, is quite striking. Think of it this way: Even if websites police fake reviews and remove them, their "legacy" lives on in future real reviews whose ratings they have biased. This type of damage is very hard to undo.[8]

Prospective Solutions

How to circumvent these human tendencies in ways that decrease the possibility for fraud and bias is a natural area for further research. At a time when online ratings systems are having a systemic and profound effect on consumer decision making, it is incumbent on scientists to learn how these social processes work and on designers to create systems that curb bias and manipulation.

The incentive to fake ratings often combines with website design (or website rules) to affect the likelihood of fraud and positive-ratings bubbles. Consider a comparison between TripAdvisor and Expedia. While anyone can post a review on TripAdvisor, a consumer can only post a review of a hotel on Expedia if he or she actually booked at least one night at the hotel through the website. An excellent study by Dina Mayzlin, Yaniv Dover and Judith Chevalier shows that hotels with a high incentive to submit fraudulent reviews (independent brands owned by single-unit owners) have a greater share of five-star reviews on TripAdvisor relative to Expedia than do hotels with a lower incentive to fake (franchise brands or chains, which benefit less from reviews and have a greater reputational risk from committing fraud).[9]

Site design and policy can influence ratings bubbles. In 2013, Reddit changed the design of its website to create an option for moderators "to

5. Friendship moderated the impact of social influence on rating behavior. Friends tended to herd on current positive ratings and to correct comments that had negatively manipulated ratings, while enemies' ratings were unaffected by our treatment. (However, this could be because of the small sample of potential first ratings by enemies. Though there are a substantial number of enemies in the community, they were less active, yielding a smaller sample of enemies' ratings.)

We also conducted several tests that ruled out other seemingly commonsensical explanations of our results. For example, one might think users are simply predisposed to rate others' comments positively. The public nature of online ratings — posting a comment for the world to see — might make it more likely that people would rate positively than negatively. And such tendencies certainly exist in our data. Up votes were 4.6 times more common than down votes on this site, with 5.13% of all comments receiving an up vote by the first viewer of the comment and only 0.82% of comments receiving a down vote by the first viewer.

But such tendencies cannot explain the results of our experiment. Though positive votes were more prevalent, our results record the greater-than-expected positivity inspired by a prior positive vote and the correction effect created by a prior negative vote. As comments were manipulated at random, the tendency toward positivity and the relative positivity or negativity of the comments themselves was held constant in our estimates.

One might also suspect that users would fear recriminations for their negative votes if their votes were potentially traceable back to them, making them unlikely to vote negatively and more likely to vote positively in response to treatment. But we designed the experiment to avoid this confounding factor. Users were unaware of the manipulation and unable to trace votes to any particular user, mitigating the potentially chilling effects of fear of recrimination.

Another potentially complicating factor is that users may be more likely to notice, read or vote on comments that have already been voted on than those that have not, creating a selection effect toward comments with more votes or with more positive or negative votes (such as the votes in our manipulations). On some websites, for example, articles with more votes are algorithmically pushed to the top of the list. But in our experiment, users did not observe the comment scores before clicking through to comments — each impression of a comment was always accompanied by that comment's current score, tying the comment to the score during users' evaluations — and comments were not ordered by their popularity or vote score, mitigating any selection bias on high- (or low-) rated comments.

We differentiated between opinion change and selective turnout as drivers of social influence bias. We analyzed changes in turnout (the likelihood of rating) and changes in positivity (the proportion of positive ratings) to identify variance explained by selection effects and opinion change respectively. This analysis ruled out selection effects (differential turnout by different types of raters, such as positive or negative raters) and suggested that a combination of opinion change and greater turnout combined with a natural tendency to up-vote on the site together created the herding effects we observed.

obscure the vote counts on comments for a predetermined amount of time after their submission." The stated goal of this feature was to "curtail and minimize the effects of bandwagon voting, both positive and negative."[10] Stock markets have implemented similar policies. For example, the New York Stock Exchange imposes "circuit-breaker" rules that halt trading at certain thresholds for single-day declines in the market, all in an effort to stave off negative herding.

Understanding the Implications

Executives — both in their lives as business leaders and as consumers — should consider taking positive online ratings with a grain of salt. While a healthy skepticism of positive ratings might cognitively correct for social influence bias, such a correction may not be necessary for negative ratings. In addition, managers should encourage and facilitate as many truthful positive reviews as possible in the early stages of the ratings process. Systematic policies to encourage satisfied consumers to rate early on could change the minds of future consumers to feel more positively toward the products or services they are rating.

Beyond that, it behooves executives to consider a few policy implications of all of this research regarding herding and ratings bubbles. Herding is a system dynamic that when seen broadly can help leaders in a variety of settings beyond the (admittedly important) scope of online ratings. For example, if equity prices work the way positive ratings do, executives should be aware of how such herding dynamics could affect a company's stock price.

Digging deeper into the behavioral mechanisms explaining our results, we found that friends were quicker to herd on positive ratings and to come to their friends' rescue when those friends' ideas were poorly rated. This implies that the structure of social networks helps guide the structure of ratings bubbles. As websites like Facebook and Google encourage more social ratings — "likes" or "+1" indications by friends and more shared endorsements or "friendorsements" in advertising — the likelihood that social influence bias will propel ratings bubbles is increased.

Our experimental manipulations increased total turnout, which, when combined with the general preference for positivity on the site, pushed scores even higher. But we also found evidence that our manipulations actually changed people's opinions, rather than simply inspiring more positively predisposed raters to rate. How do we know? We analyzed the changes in turnout (the likelihood of rating) and in positivity (the proportion of positive ratings). We wanted to know whether the ratings changes that the experiment produced could simply be explained by more positively predisposed people providing ratings, or alternatively whether more negatively predisposed people were actually changing their ratings to positive ones. What we found was that the latter was taking place. Negative raters were becoming positive raters. Opinions were changing.

Now consider this positive opinion change along with the herding effects already discussed. It is clear why our results concerned us when we thought about them in the context of large-scale, opinion-aggregation tasks in society. We happened to conduct the analysis during the 2012 U.S. presidential election. As we heard electoral poll results of likely voters on the radio, we couldn't help but wonder: Do these types of polls predict or rather drive election results?

Many of our recent economic crises have herding and bubbles at their core — from housing to mortgage-backed securities. Understanding how herding and bubbles work is the first step toward averting their effects in a multitude of settings. More theoretical and experimental research on social influence bias and the ratings bubbles that result from it could therefore contribute not only to the management and marketing of online ratings and the user design of ratings sites but also to policies that keep human social systems from running off the rails. Policy makers and website designers should focus on understanding herding scientifically and creating policies and website designs that short-circuit herding behaviors and help prevent ratings bubbles. It is possible that such policies could reduce herding biases in everything from elections to equity markets and beyond.

***Sinan Aral** is the David Austin Professor of Management and an associate professor of information technology and marketing at the MIT Sloan School of Management. Comment on this article at http://sloanreview.mit.edu/x/55224, or contact the author at smrfeedback@mit.edu.*

REFERENCES

1. "Nielsen: Global Consumers' Trust in 'Earned' Advertising Grows in Importance," April 10, 2012, www.nielsen.com.

2. S. Bikhchandani, I. Welch and D.A. Hirshleifer, "A Theory of Fads, Fashion, Custom and Cultural Change as Informational Cascades," Journal of Political Economy 100, no. 5 (October 1992): 992-1026 ; and M.J. Salganik, P.S. Dodds and D.J. Watts, "Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market," Science 311, no. 5762 (February 10, 2006): 854-856.

3. S. Gordon, "Call in the Nerds — Finance Is No Place for Extroverts," Financial Times, April 24, 2013.

4. S. Aral and D. Walker, "Identifying Influential and Susceptible Members of Social Networks." Science 337, no. 6092 (July 20, 2012): 337-341; and S. Aral and D. Walker, "Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks," Management Science 57, no. 9 (September 2011): 1623-1639.

5. L. Muchnik, S. Aral and S.J. Taylor, "Social Influence Bias: A Randomized Experiment," Science 341, no. 6146 (August 9, 2013): 647-651.

6. See M. Luca and G. Zervas, "Fake It Till You Make It: Reputation, Competition and Yelp Review Fraud," Harvard Business School NOM Unit working paper no. 14-006, Boston, Massachusetts, November 8, 2013; Y. Liu, "Word-of-Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," Journal of Marketing 70, no. 3 (2006): 74-89; and J.A. Chevalier and D. Mayzlin, "The Effect of Word of Mouth on Sales: Online Book Reviews," Journal of Marketing Research 43, no. 3 (August 2006): 345-354.

7. N. Hu, J. Zhang and P.A. Pavlou, "Overcoming the J-Shaped Distribution of Product Reviews," Communications of the ACM 52, no. 10 (October 2009): 144-147.

8. Special thanks to Georgios Zervas of Boston University for thoughtful discussions about this particular insight.

9. D. Mayzlin, Y. Dover and J. Chevalier, "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," American Economic Review, in press.

10. Splattypus, "Why Are Comment Scores Hidden?," June 2013, www.reddit.com; see also Deimorz, "Moderators: New Subreddit Feature — Comment Scores May Be Hidden for a Defined Time Period After Posting," May 2013, www.reddit.com.

i. One important exception is the seminal work of Salganik, Dodds and Watts, who conducted a large-scale lab experiment in an "artificial cultural market." See Salganik et al., "Experimental Study of Inequality and Unpredictability." Our work takes this research one step further by examining herding effects on a live website "in the wild" and by examining both negative and positive herding.

ii. Muchnik et al., "Social Influence Bias."

Reprint 55224.

Copyright © *Massachusetts Institute of Technology, 2014. All rights reserved.*

MITSloan
Management Review

PDFs ■ Reprints ■ Permission to Copy ■ Back Issues

Articles published in MIT Sloan Management Review are copyrighted by the Massachusetts Institute of Technology unless otherwise specified at the end of an article.

MIT Sloan Management Review articles, permissions, and back issues can be purchased on our Web site: *sloanreview.mit.edu* or you may order through our Business Service Center (9 a.m.-5 p.m. ET) at the phone numbers listed below. Paper reprints are available in quantities of 250 or more.

To reproduce or transmit one or more MIT Sloan Management Review articles by electronic or mechanical means (including photocopying or archiving in any information storage or retrieval system) **requires written permission.**
To request permission, use our Web site:
sloanreview.mit.edu),
or
E-mail: smr-help@mit.edu
Call (US and International):617-253-7170
Fax: 617-258-9739

Posting of full-text SMR articles on publicly accessible Internet sites is prohibited. To obtain permission to post articles on secure and/or password-protected intranet sites, e-mail your request to smr-help@mit.edu.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.