9^{th} International Conference on Computational Social Science IC^2S^2
July 17-20, 2023, Copenhagen, Denmark

The Good, The Bad and The Picky: Reference Dependence and the Reversal of Product Ratings

Keywords: Online Learning, Review Systems, Behavioral Biases.

Extended Abstract

Online consumer ratings have become a ubiquitous driver of choice. But, to what extent can we trust their informational content? Because consumer ratings largely measure subjective satisfaction, they can reflect characteristics of their writers just as much as of what is being reviewed. If individuals' characteristics correlate with their choices, a self-selection bias in ratings will arise.

The mechanism we propose is simple: consumers differ in their expertise, which has two interrelated effects. On one hand, more experienced consumers identify and buy higher quality products. On the other hand, the level of satisfaction consumers get from the products they purchase (and, thus, their ratings) depend negatively on their standards, or the level of quality they are used to. Therefore, the ratings of higher quality products reflect the higher standards of their buyers, and are thus downward biased relative to the ratings of their inferior alternatives.

We first propose a simple theoretical model building on the idea that a combination of heterogenous consumer expertise and expectation-based reference-dependent individual utility gives rise to a compression of ratings, effectively penalizing high quality products compared to their lower quality alternatives. The model shows that ratings need not only compress quality differences, they can actually reverse quality rankings. We show that whether this occurs depends on the combination of three of the model's primitives: the share of Expert consumers, the importance of standards in shaping utility, and the size of the utility gap due to expertise.

The reason for this is that, because the share of elite Experts is – by definition – small, increasing their weight further exacerbates the adverse selection problem faced by high quality products, worsening their relative ratings. If only 10% of consumers were Experts, for instance, then double-counting their opinions would bring their effective weight to over 18%, increasing buyers' heterogeneity. Conversely, when only a few inexperienced consumers are on the market, then overweighting the already prevalent Experts' opinions would help decreasing buyers' heterogeneity, and thus ratings' biases.

One obvious criticism is that if these biases in ratings are predictable, shouldn't consumers correct for them and draw proper inference about qualities? Indeed, in many real life scenarios we are aware of our recommenders' characteristics and can thus make inference about the informational content of their advice: if they developed a reputation for being easily satisfied, we will discount part of their enthusiasm. While this may be true for very educated consumers in physical markets, online learning makes this joint inference much more complicated. In most cases, consumers ignore the identities of the reviewers, and hence have limited capacity to extract the relevant information from their posted opinions. Moreover, these selection effects depend on quality itself, which is unknown in the first place. Therefore, while each rating is the combination of both the product's and the reviewer's characteristics, the latter are often ignored, producing an incorrect estimate of the former.

We describe and model these mechanisms in a vertical market. In other words, all consumers would individually rank all products in the same way. Nevertheless, most markets in

which reviews are employed – and particularly the movie market, which we focus on empirically – have a substantial degree of product differentiation. We have two key motivations for our choice, one conceptual, the other empirical.

First, the low correlation between consumers and critics' opinions have often been ascribed to systematic differences in taste. By obtaining analytical results in a model without taste differences, we show that low correlation can be achieved in a much broader context, when consumers' opinions are incorrectly aggregated.

Second, our data is extremely stark in this respect, as shown in Figure XX: it is not that more and less experienced consumers rate different movies differently. The correlation between the two categories is high. We believe that the real story Figure XX tells is that Experts rate virtually *every* movie (for instance, over 99% of our sample) lower than Non-Experts, irrespective of genre, year, or popularity. If Experts simply had a different taste from Non-Experts, this fraction would be much lower, and we would identify several genres, or at least movies, preferred by Experts.

We empirically substantiate the severity of this bias – as well as its drivers and consequences – by studying consumer movie ratings. In particular, we scraped detailed aggregate data for over 9000 movies from IMDb, the most popular movie rating platform in the world, and complemented it with a massive (24 million reviews for 15,700 movies left by 162,000 individual users) individual-level rating dataset from MovieLens, a platform for movie discovery and personalized recommendations.

For each movie, on top of the overall score, an advanced search reveals more detailed averages concerning only specific subgroups of consumers. Of special interest for us is the Top1000 Users category, which groups together the 1000 users who have posted the most ratings on the platform. We will henceforth refer to them as Experts and contrast their behavior with that of all other users (Non-Experts). Experts watch, on average, higher-quality movies. To proxy movie quality, we employ external sources such as the nominations and awards for the most relevant festival and industry awards around the world, and critics' reviews aggregated by Metacritic. Moreover, Experts are harsher than Non-Experts in their ratings. This is not only true on average, but the result holds for almost 98% of movies in our sample. This rating gap is statistically and economically significant: for a given movie on IMDb, Experts award, on average, over half-star less than Non-Experts. The combination of Experts' quality-based self-selection and their stringent rating behavior implies that aggregate ratings penalize high quality movies compared to their inferior alternatives, as predicted by our theoretical model.

Last, we propose a fixed-point recursive algorithm to debias the ratings. We exploit the full history of individual ratings and compute a stringency score for each user. Then, we subtract user-specific stringency from each of the user's ratings. We then use these corrected individual ratings to re-compute all movie ratings. We iterate this process until it converges, that is, until individual stringencies and movie ratings are self-confirming. Upon completing this process, we find that, as predicted by our theory, our debiased aggregate ratings better correlate with external measures of quality, such as nominations and awards, and reviews by critics. This correction is appealing in that it does not require us to take a stance on which ratings (or which users) are more or less accurate, nor on which movies are more or less high quality.