Goodreads vs Amazon: The Effect Of Decoupling Book Reviewing And Book Selling

Stefan Dimitrov[1], Faiyaz Zamal[1], Andrew Piper[2], Derek Ruths[1]
stefan.dimitrov@mail.mcgill.ca, faiyaz.zamal@mail.mcgill.ca,
andrew.piper@mcgill.ca, derek.ruths@mcgill.ca
[1]School of Computer Science, [2]Department of Languages, Literatures, and Cultures
McGill University

Abstract

Book reviewing is a commonplace activity on many e-commerce sites. However, it is nested within the broader context of book buying and selling. Goodreads, an on-line platform for social curation of book collections, provides an opportunity to observe on-line book reviewing in an environment that is not (at least overtly) focused on commercialization. In this study, we perform a careful comparative study of reviewer behavior and engagement in Goodreads and Amazon.com, constrained to a single genre (biography), including 21,394 books and 2.5 million reviews. We discover marked differences between the platforms that suggest disparate population composition and objectives of review-writing across the two platforms. Our findings suggest an important and generalizable principle: that two platforms engaging users on the same task (e.g., book review writing) may elicit quite different behavior depending on the implicit or explicit context and motivation present.

User-contributed product and service reviews have become exceedingly common on a wide range of on-line platforms. This is due, in large part, to the significant value that such reviews carry for various stakeholders. For visitors to a site, reviews provide valuable information (typically in the form of recommendations); for platform owners, reviews drive traffic to the site; for review authors, writing reviews can feel altruistic and can also confer status within the on-line platform; and for producers of the products and services being reviewed, such reviews can influence sales — both on the platform and more broadly.

Because on-line reviewing has become widespread and carries both economic and social value, significant research has been done to characterize and understand reviewing behavior, with a great deal of work focusing on topics such as review authenticity (Ott, Cardie, and Hancock 2012; et al. 2013), review relevance estimation (Martin and Pu 2014), and rating prediction (McAuley and Leskovec 2013).

Of course, insofar as reviews are intended to represent the candid views of users (who have experienced the product/service in some way), a central question in this domain is the extent to which reviews - of the same products and product class - are strongly conditioned by the context in which they are written. Given two sites eliciting reviews on the same products - how similar or different might we expect their respective reviews to be? To our knowledge, very few studies have approached this question - the most relevant work of which we are aware considered the effect of exogenous factors (weather and demographics) on the content and style of recommendations (Bakhshi, Kanuparty, and Gilbert 2014). While relevant, prior work such as this has not addressed the question of how platform context shapes review writers and reviews themselves.

In this study, we investigate this question within the context of book reviews. Books continue to represent one of the largest sectors of the culture industry, with most recent reported annual U.S. sales of just over 27 billion USD (AAP 2013); U.S. music sales by comparison were just under 4.5 billion USD (IFPI 2014). Amazon and Goodreads are the two largest English-language platforms that allow users to write and publicly share reviews of books. The stated purpose of these sites, however, are quite different. Amazon is an e-commerce site, such that book reviews presumably inform the book purchases that web visitors make. Goodreads, by contrast, is a book "social cataloging" site in which users can keep track of books they have read/are reading/would like to read, organize books into collections, and add comments to books (which function as reviews insofar as comments are public to the entire Goodreads community, include star ratings, and are aggregated at the book-level).

Thus, a key difference between Goodreads' and Amazon's e-commerce site is that Goodreads does not link the review-writing process to purchasing decisions in any overt way (the interested Goodreads user can find a link to booksellers, but this information is very discretely positioned on the book's webpage). It is worth noting that while Goodreads is owned by Amazon, there appears to be little or no integration between the two services (other than a more prominent link to Amazon.com than to other booksellers). By studying Goodreads and Amazon reviewing behavior at large-scale, we have the opportunity to observe how overt review-commercialization impacts review writing and review construction. Notably, while work has been done on Amazon reviews, this is the first published study to perform large-scale analysis of any kind on Goodreads data (e.g., (Chevalier and Mayzlin 2006; Hu, Jok, and Reddy 2014)).

In this first exploratory study, we limit our analysis to a

specific genre: biography. Genre labeling represents a complex problem in its own right with regard to the different review platforms, motivating us to focus our attention on a single, well-defined genre. Goodreads, for example, has over 790 genre labels (ranging from "fiction" to "fantasy" to "fruits-and-veggies"), while Amazon uses a hierarchical classification scheme that is not immediately aligned with genre ("books → business & investing → high-tech" is one example). We choose to focus on biography because it represents both a high degree of generic coherence (biography, unlike categories like art books or even history, has a tightly constrained focus on a single individual's life) and it enjoys a high degree of popular interest (one need only think of the cultural significance of Steve Jobs' biography).

Within this genre, our analysis considers a range of statistics including review length, sentiment, and vocabulary distributions. These measurements support a number of broad conclusions. Foremost, our work suggests that Amazon reviewers do, indeed, tailor their reviews for purchasing. This contrasts with Goodreads, which maintains a more engaged reviewing population - with users contributing more reviews on average than Amazon reviewers - whose reviews are more oriented towards discussion of book characteristics. We also find that, paradoxically, rating distributions are quite different, but that the average sentiment of reviews are nearly identical between Amazon and Goodreads. We take this to be an indication that readers on both sites agree on the nature of reviews, but systematically make different choices about rating values.

Data

We crawled the Goodreads and Amazon websites for books and reviews.

Goodreads. First, we crawled every book in the 'biography' genre from Goodreads. It is worth noting that genres in the Goodreads community are user defined. Thus, we were able to identify 10,820 books labelled 'biography' by at least one Goodreads user and with at least one review. Crawling the reviews for these books was very challenging, primarily due to the request authenticity verification system employed by Goodreads. Still, we were able to craft AJAX requests for every review page, giving us a total of 1,600,471 reviews for biography books. We decided to work with the full set of reviews as opposed to a sample, because the Goodreads website uses a proprietary (unknown) algorithm for review sorting. Thus, any review sampling would not have been random and could not be fairly compared with a random sample of reviews from Amazon.

Amazon. We downloaded all book pages and reviews from Amazon using the ISBN numbers from our Goodreads dataset: a total of 10,574 book pages, or 246 books fewer than our Goodreads dataset. The reason for this difference is either that these books had no ISBN number on Goodreads or they are not listed on Amazon. Our Amazon dataset contains 945,548 reviews. Some of these reviews relate to verified purchases on Amazon, however we did not consider the difference in review types for the current study.

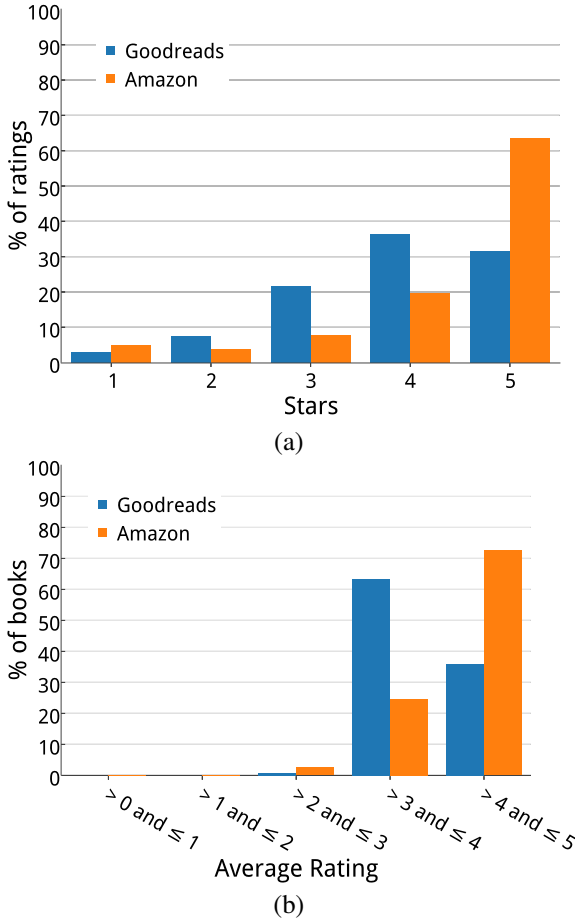

Figure 1: The distribution of stars (a) over all ratings and (b) averaged by book.

Results

Reviewer abundance. We were able to extract the unique user identifier for every review in both of our datasets (581,409 users on Goodreads and 631,922 users on Amazon). However, we found that 44,023 of our Amazon reviews were written by anonymous users; we excluded these reviews in our distribution calculations. We discovered that, on average, Goodreads users write more reviews (2.75) than Amazon users (1.51). There are few very engaged users who write disproportionally high number of reviews on both websites (371 reviews for one user on Goodreads and 699 on Amazon).

Review abundance. For every book in our dataset, we recorded the total number of reviews. Amazon displays reviews from all editions on the page of every edition, resulting in duplicate reviews (46,454 of the Amazon reviews were excluded from the experiments as duplicates). We found that on average books have more reviews on

Statistic	Goodreads	Amazon	Difference
# of Reviews*	1,600,471	945,548	654,923
# of Users*[1]	581,409	631,922	50,513
Avg # Reviews/User*	2.75 ± 5.35	1.51 ± 2.90	1.24
Avg # Reviews/Book*	147.92 ± 357.15	95.65 ± 355.50	52.27
Avg Review Length* (words)	87.09 ± 131.10	117.84 ± 164.36	30.75
Avg Review Length* (sent.)	5.0 ± 5.78	5.95 ± 6.79	0.95
Avg Sentence Length*	99.25 ± 124.44	110.44 ± 96.61	11.19
Avg Stars*	3.86 ± 1.04	4.33 ± 1.09	0.47
Avg Stars/Book*	3.88 ± 0.31	4.27 ± 0.51	0.20
Avg Sentiment/Review	0.04 ± 0.08	0.04 ± 0.07	0
Avg Sentiment/Book	0.04 ± 0.02	0.04 ± 0.02	0

Table 1: The distribution of statistics computed. All starred (*) statistics have statistically significant differences ($p < 0.00001$).

Goodreads (147.92) than on Amazon (95.65). As expected, very popular books have a large number of reviews (3,000 for our top Goodreads book and 15,990 for Amazon).

Review length. We compared the length of reviews on Goodreads and Amazon, by looking at word count. We found that Amazon reviews are longer (including 1st quartile, mean, median, 3rd quartile), which was contrary to what we expected. Thus, we considered other measures of review length: number of sentences in a review and words in a sentence. These experiments confirmed our finding: reviews for biography books on Amazon are significantly longer than on Goodreads. Of note, Amazon reviews have previously been observed to be longer than other book reviewing sites (Chevalier and Mayzlin 2006).

Review star ratings. Both Goodreads and Amazon allow users to rate books on a scale from 1 to 5 stars. We considered the average rating per book and found that on Amazon books have higher average rating (4.27) compared to Goodreads (3.88). The medians of these distributions are 4.4 and 3.9 respectively. We also looked at the overall rating distributions (see Figure 1) and found that Amazon users are much more likely to give 5 stars (63.60% of all ratings) than Goodreads users (31.55% of ratings), who give 4 stars more often (36.26%) and much more often than Amazon users (19.73%). Amazon users, on the other hand, are more likely to express dissatisfaction (4.82% of ratings are 1 star) compared to Goodreads users (only 2.84% ratings 1 star). We split the average ratings per book into 5 categories (0-1, 1-2, 2-3, 3-4, and 4-5 stars) and found that the highest proportion of Goodreads books falls in the range 3 to 4 stars (63.26%) while on Amazon the highest proportion of average ratings are between 4 and 5 stars (72.56%). This finding reaffirms our observation that readers are more likely to give 5 stars to biography books on Amazon than they are on Goodreads.

Review sentiment. After measuring review lengths, we performed sentiment analysis to identify whether the greater brevity of Goodreads reviews corresponded with the expression of stronger feelings. We split every review into words and performed a look-up in the SentiWordNet sentiment dictionary (Baccianella, Esuli, and Sebastiani 2010). We took the posterior polarity of the most popular meaning (Guerini, Gatti, and Turchi 2013) for words with multiple definitions. We looked at four metrics: the avg. positive review sentiment (normalized sum of all positive terms in a review), the avg. negative review sentiment, the absolute sentiment in a review, and the difference between positive and negative scores in a review. Comparing the overall sentiment rate (all reviews) on Goodreads and Amazon does not show a significant difference for any of the four metrics - only two of these statistics are shown in Table 1. We also looked at a subset of shorter (10 to 20 words) reviews, but the results were identical between the two sites. As a final permutation, we considered the impact of appending the title to the text of shorter reviews on Amazon, still we did not identify a significant difference in the expression of sentiment on Goodreads compared to Amazon.

Platform-specific review language. In order to get a general sense for the kinds of review content that typifies each platform, we performed a Wilcoxon rank-sum test on all review vocabulary. The top indicative words for each platform are revealing.

- *Amazon*: buy, bought, will, purchased, reader, gift, purchase, ordered, highly, reviewers, price
- *Goodreads*: goodreads, shit, interesting, pretty, memoir, bit, listened, funny, definitvely, didn't, parts

The word sets suggest that Amazon reviews have a uniquely strong affinity for discussing book buying.

Key Findings

Goodreads has a more engaged reviewer base. Reviewers write more reviews and books have more reviews. That said, Goodreads has a smaller reviewer base (this is definitely the case since there were 44,023 anonymous Amazon reviews that could not be included in the count).

Amazon users write more purchase-oriented reviews. We expected Goodreads to have longer reviews, with the idea that the Goodreads population would be more interested

in discussing books, rather than writing arguments for or against buying a book. Based on the literature in the field, we expected the participatory platform to exhibit more expressive breadth than the commercial platform (Jenkins 2006). In fact, however, reviews are significantly longer on Amazon and contain, on average, longer sentences per review. These details suggest a greater degree of expressive complexity which could be consistent with nuanced arguments for/against buying decisions. This possibility is further supported by the platform-indicative review words, for which Amazon showed an enrichment for buying-related vocabulary. In contrast, we suspect that Goodreads reviews are less invested in convincing readers to take a particular action (buy/not-buy) and more reflective of journaling practices or community conversations that tend to be more impressionistic in nature. This would also account for the greater number of reviews per reviewer and for the enrichment in words relating to book quality.

Amazon generates more extreme valued reviews. We found that Amazon ratings are more extreme, suggesting an intent to sway a decision-making process rather than convey nuanced feelings about a reading experience. While the percentage of 2-4 star reviews on Goodreads greatly outnumber those on Amazon (with the middle ground of 3 stars exhibiting the greatest difference), 1 and 5 star reviews occur almost as twice as often in the case of 1 stars on Amazon and more than twice as often in the case of five stars. In particular, this agrees with prior work showing that extreme-valued ratings are most persuasive where buying decisions are concerned (Chevalier and Mayzlin 2006).

Sentiment is stable across platforms. Curiously, even though Amazon star-based reviews are more extreme, the strength of the sentiment (whether positive, negative, or both) is almost identical across the platforms. This suggests either a limitation in the dictionary used for this particular domain or that ratings are not strongly connected with a sentimental vocabulary. Further exploration is needed to understand how it is that people are saying things with the same sentimental valence, but giving different numeric ratings.

Discussion

The overall image that emerges from our analysis is that Amazon and Goodreads reviews are fundamentally different in ways that reflect the different orientations of the platforms on which they are written.

Amazon reviews have characteristics indicating that review writers are trying to "sell" the book. The length of reviews, the tendency to choose extreme rating values, and a propensity to use terms that concern purchasing behavior all support this observation.

In contrast, attributes of Goodreads reviews reflect the content-orientation of the platform. The vocabulary of reviews favors words that highlight attributes of books or of the experience of reading, reviews tend to be shorter and more journalistic. And ratings for books tend to be more moderate, reflecting both a more nuanced approach to rating and, possibly, a lesser sense of needing to use ratings to "convince" other readers of the reviewer's position (Chevalier and Mayzlin 2006).

Future Work

While the statistics collectively present a coherent picture of the two platforms and their reviews, a number of open questions remain.

Foremost, Amazon's tendency towards longer reviews is curious and deserves further investigation. In addition, a more thorough investigation of affective expression might identify nuanced ways in which sentiment and emotion are employed differently on the different platforms. More broadly, this study needs to be expanded to include other genre; this would have the effect of establishing whether the trends reported here generalize to the platform as a whole. A broader study of the two platforms would also permit cross-genre analysis with the idea that different populations of readers and reviewers engage different genre and possibly view or treat reviewing differently.

References

AAP. 2013. *Bookstats*, volume 3. Association of American Publishers.

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, 2200–2204.

Bakhshi, S.; Kanuparty, P.; and Gilbert, E. 2014. Demographics, weather and online reviews: a study of restaurant recommendations. In *Proceedings of WWW*. Exogenous factors like weather and demographics can affect reviews.

Chevalier, J. A., and Mayzlin, D. 2006. The effect of word of mouth on sales: online book reviews. They observed that Amazon reviews are longer than Barnes and Nobles, reviewers seem to depend on content, not summaries; lower ratings carry more weight than higher ratings.

et al., A. M. 2013. Spotting opinion spammers using behavioral footprints. In *Proceedings of KDD*.

Guerini, M.; Gatti, L.; and Turchi, M. 2013. Sentiment analysis: How to derive prior polarities from sentiwordnet. *arXiv preprint arXiv:1309.5843*.

Hu, N.; Jok, N. S.; and Reddy, S. K. 2014. Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales.

IFPI. 2014. *Recording Industry in Numbers*. IFPI.

Jenkins, H. 2006. *Fans, Bloggers, Gamers: Exploring Participatory Culture*. NYU Press.

Martin, L., and Pu, P. 2014. Prediction of helpful reviews using emotions extraction. In *Proceedings of AAAI*. They find that influential users tend to use more affective words.

McAuley, J., and Leskovec, J. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of RecSys*.

Ott, M.; Cardie, C.; and Hancock, J. 2012. Estimating the prevalance of deception in online review communities. In *Proceedings of WWW*.