2009 International Conference on Computational Science and Engineering

Is the crowd's wisdom biased?

A quantitative analysis of three online communities

Vassilis Kostakos
Department of Mathematics & Engineering
University of Madeira
vk@uma.pt

***Abstract*—We present a study of user voting on three websites: Imdb, Amazon and BookCrossings. Here we report on an expert evaluation of the voting mechanisms of each website and a quantitative data analysis of users' aggregate voting behavior. Our results suggest that the websites with higher barrier to vote introduce a relatively high number of one-off voters, and they appear to attract mostly experts. We also find that one-off voters tend to vote on popular items, while experts mostly vote for obscure, low-rated items. We conclude with design suggestions to address the "wisdom of the crowd" bias.**

Keywords-Voting, rating, quantitative analysis, expert evaluation.

I. INTRODUCTION

As increasing numbers of users connect to the Internet, the potential for exploiting the "wisdom of the crowd" becomes greater. Our long-term research vision is to understand how to organize and utilize online communities to engage in collective problem-solving. A crucial stipulation in this approach remains the effective identification and removal of bias from the crowd's wisdom, and the design of systems that minimizes such bias.

In this paper we take a quantitative approach at identifying bias in three communities where users collectively carry out explicit voting and rating. Here we compare votes and reviews from Imdb and Amazon (on movies) and BookCrossings (on books). A quantitative approach can highlight important aspects of aggregate behavior, and more crucially such aspects cannot be identified by a qualitative approach operating on small samples. We show that despite the large community size of each website, there exist significant biases in users' voting and rating behavior. Here we frame these biases in terms of the design of the voting mechanisms at each website. We conclude this paper by providing design suggestions to help identify and eradicate bias in large online communities.

II. RELATED WORK

The analysis of aggregate voting & rating behavior has been extensively researched in the context of recommender systems [e.g. 7]. Further work has considered modeling the dynamics of collective rating behavior [5], and developing algorithms for detecting malicious users or groups of users that try to manipulate rating systems [8]. Such work, however, typically considers the "back end" system and is rather independent of the user interface design of the voting mechanisms.

More recently, researchers have begun to explore users' perceptions of recommendation engines [6], and the effect of recommendations on users' shopping behavior [4]. It is important to note that a number of factors, such as coordination [3] and level of difficulty of task [2] have been shown to improve the quality of aggregate user performance.

While previous work has considered the interface and visual design of online communities in the context of voting and rating, typically it relies on rather small population samples engaged either via interviews, lab studies, or observation. In this study we take an holistic approach and examine the behavior of all users of each website, and analyze their voting behavior in terms of the voting mechanisms. We perceive our work as orthogonal to traditional empirical evaluation focusing on individual users and qualitative feedback. Here we consider the big picture.

III. DESCRIPTION

For the present study we collected two sets of data. The authors carried out an expert evaluation of each website, focusing on the the voting mechanisms of each. In addition, we collected data on all votes cast on each website. For Imdb and BookCrossings we used previously published data, while for Amazon we collected movie ratings directly via public APIs.

A. IMDB

Imdb is an online movie database. Each of its items has a short listing which is shown in search results and does not include any rating information. Each item also has its own page where the full listing appears. This shows total number of ratings the movie has received and their average (out of 10), as well as a small number of user reviews. In addition, every item has a more detailed ratings page that displays a histogram of votes, and a detailed comments page where all user comments are accessible.

Users need to register in order to vote on Imdb. Once registered, users can cast votes without leaving the movie's full listing page, simply by clicking on the graphic that represents the movie's current rating in terms of stars. Depending on which start is clicked an appropriate rating is cast on behalf of the user. Ratings do not require an associated text review or justification, but users can choose to cast a vote with an associated review on a separate page. Ratings with associated reviews can themselves be characterized as Useful / Not Useful by other users, hence pushing them up on the list of reviews. Another feature of Imdb is that each user has a profile page where all their reviews are shown. Users themselves are not directly rated in any way.

978-0-7695-3823-5/09 $26.00 © 2009 IEEE
DOI 10.1109/CSE.2009.491

B. Amazon

Amazon is an online retailer of various goods including movies, books, and electronics. This website generates a short listing for each item appearing in search results, which includes the average rating of the item (1-5 stars with half stars in between) and the number of ratings. The full listing for each item adds a small v-shaped graphic next to the star rating, which unveils a histogram of votes when the mouse cursor hovers above it. This page also includes some user reviews of this item. In addition, each item has an associated review page that provides access to all reviews.

Users need to register in order to vote on Amazon. To cast a vote, users need to find the item they wish to vote for and click on the button "Create your own review". This takes them to a separate page where they enter a rating (1-5 stars), a title for their review, and a written or video review. All reviews can be rated as Helpful / Not helpful by other users, and can also be commented on. Users have a profile page showing their reviews, friends, people they find interesting, tags they use, and products they have tagged. In addition, users can have badges, e.g. "Top 500 reviewer" (by number of reviews), "Real Name" (their profile name is their real name), "Vine Voice" (gets early releases to comment on them). These badges are shown wherever their name appears on the website.

C. BookCrossings

BookCrossings (BC) is an online book community where users share their reviews about books. An interesting distinction that BookCrossings makes is that each individual copy of a title is treated uniquely. Individual copies are differentiated by custom printed labels that users stick on books. These labels have a BCID (BookCrossings ID) which is unique to each copy. The website operates on the principle that users can find such books in public spaces like a cafe, read them, and then return them to some other public space. Each title has a star rating (1-10 full stars) that only appears in the full listing page. In addition, the listing page provides text reviews of the book.

BookCrossings requires users to register in order to vote. Furthermore, users must have a valid BCID in order to vote, and these do not appear on the site, but only on the labels that are stuck physically on books. If a user has purchased a new copy, then they must go through the process of registering the book (by entering its ISBN and author/title information), thus generating a new BCID label which they must print and stick on the book. Then they can use this new code to write their review for the book. Users also have a public profile, with basic demographics and the number of books they have reviewed, with further links to the books themselves. Finally, BookCrossings has a "high score" page showing the users with the all-time highest number of registered books, as well as the users with most registered books in the previous week.

IV. RESULTS

A. Expert analysis

We completed an expert analysis of the three websites, which highlighted important differences in the voting mechanisms across the three websites. In particular, we found differences in the barrier to casting a vote, quality control, and the motivation mechanisms.

Given that none of the websites allow for anonymous voting, Imdb offers the least barrier to casting votes, as this can happen directly on the items' description page and with no need for a textual justification. Amazon comes next, as it requires users to enter a text or video review with each rating. BookCrossings has by far the highest barrier to vote, as in addition to a written justification it requires users to generate unique serial numbers for items they wish to rate, or requires users to have physical access to the items at the time of voting.

The quality control mechanisms also vary across the three websites. Amazon has the strongest quality mechanism as it requires users to write a review in order to justify their rating. Also, it enables users to rate other users' reviews as helpful or not, and additionally write a meta-review. Imdb gives users the option to submit a textual justification for their vote, and it allows users to rate such reviews as useful or not. On the other hand, BookCrossings offers no mechanism for peer review.

In terms of motivation, Amazon comes first with the series of "badges" that users can earn based on their performance, and additionally allows users to specify friends within the community. BookCrossings has explicit "high score" lists of users based on their performance, while Imdb allows users to specify their friends within the community.

B. Quantitative analysis

In addition to expert analysis, we analyzed vote records from each website. Table 1 shows the number of records that we analyzed per website. The Imdb dataset was obtained from the Imdb website and is updated regularly. We do not have data grouped by user as this is not made available. The Amazon data was collected by the authors during March 2008 from the American version of the website by using public APIs. The BookCrossings data is available from http://www.informatik.uni-freiburg.de/~cziegler/BX and was collected during August-September 2004.

Table 2 shows the differences between novice and expert behavior across the three websites. Specifically, it shows the percentage of users with only one vote (i.e. novices), and the least number of votes for a user (or movie) in the top 5% of the

	Imdb	Amazon	BC
Items	233,106	21,880	271,379
Users	-	134,272	278,858
Votes	101,281,733	379,651	1,149,780

Table 1. The number of items, users and votes that we analyzed. Note that BookCrossings items refers to unique ISBNs.

	Novice vs. Expert behavior					
	Imdb		*Amazon*		*BC*	
	Users	Items	Users	Items	Users	Items
1 vote	-	7%	58%	23%	56%	58%
top 5% of population	-	> 662 votes	> 7 votes	> 10 votes	> 30 votes	> 10 votes

Table 2. The percentage of users and items that had only 1 associated vote, and the maximum number of votes for the bottom 95% of the population of users or items.

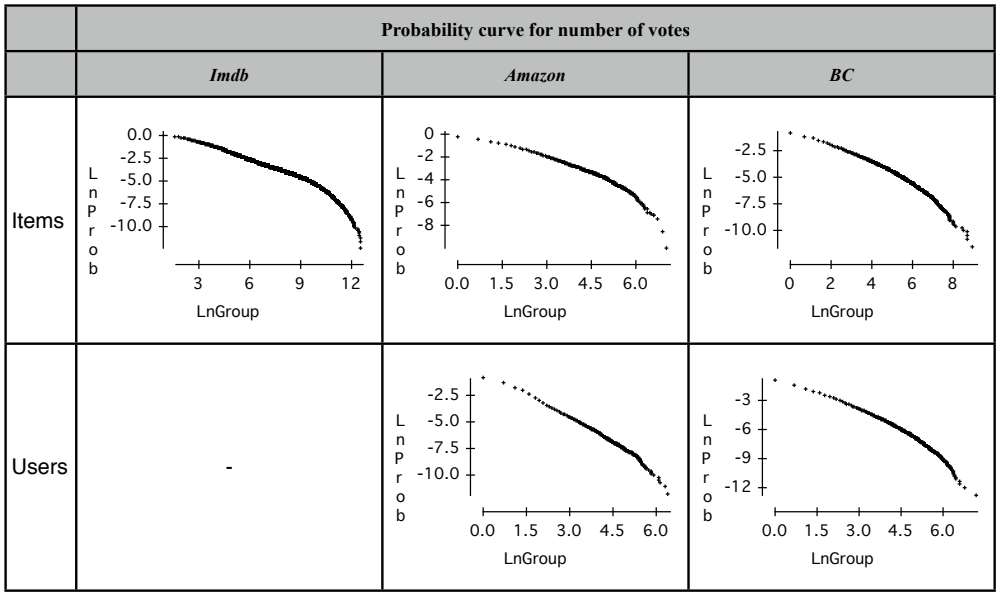

Table 3. The probability distribution of number of votes per item and per user, shown on ln-ln plots. X-axis is ln(number of votes), Y-axis is ln(probability) of an item having less than x-axis number of votes.

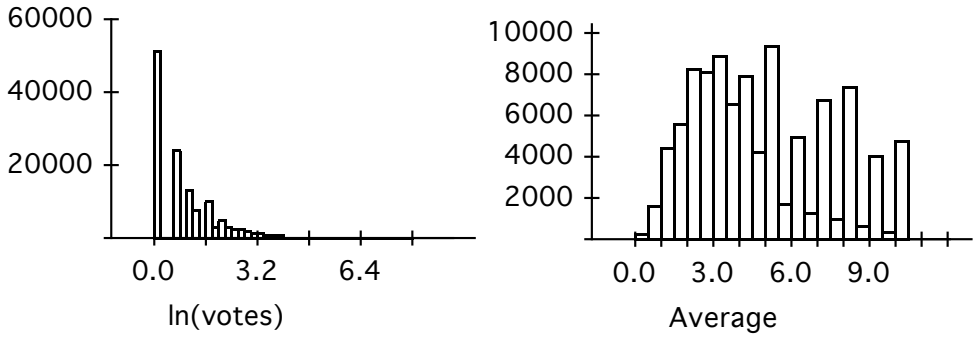

Figure 1. Focusing on the items that the top 5% of users vote for on BookCrossings (users with >30 votes). On the left a histogram of total votes per book (including votes of non-expert users), and on the right a histogram of average rating (including votes of non-expert users).

population (i.e. experts). A Mann–Whitney U test showed that the distribution of number of votes differed significantly between the users on Amazon and BC (U=6612209442, p<0.0001). Similarly, the distribution of number of votes per item is significantly different between Amazon and BC (U=6312, p<0.0001), Amazon and Imdb (U=3837597876, p<0.0001), and BC and Imdb (U=1708, p<0.0001).

Table 3 shows the ranked probability plot for the number of votes per user and per movie. This shows the probability that a user, or item, has a certain number of votes associated with them. The probabilities follow a power law with an exponential cutoff. Specifically, for Imdb Items the distribution has exponent alpha = -0.64, R^2=0.998, p<0.0001; for Amazon items alpha = -0.94, R^2=0.999, p<0.0001; for Amazon users alpha = -1.56, R^2=0.998, p<0.0001; for BC items alpha = -1.13, R^2=0.992, p<0.0001; for BC users alpha = -1.66, R^2=0.99, p<0.0001. In addition, an ANOVA showed a significant effect of website on the number of votes cast by users (F(1,239553)=1080, p<0.0001), as well as on the number of votes received by each item (F(2,707593)=2449.4, p<0.0001)

Table 4 presents histograms of the average rating of items, as well as the average rating of each users. "User rating" is based on the votes cast by users about items, not about other users. A Mann–Whitney U test showed that the normalized rating distribution differed significantly between the users on Amazon and BC (U=241644704, p<0.0001). Similarly, the distribution of normalized item ratings differed significantly

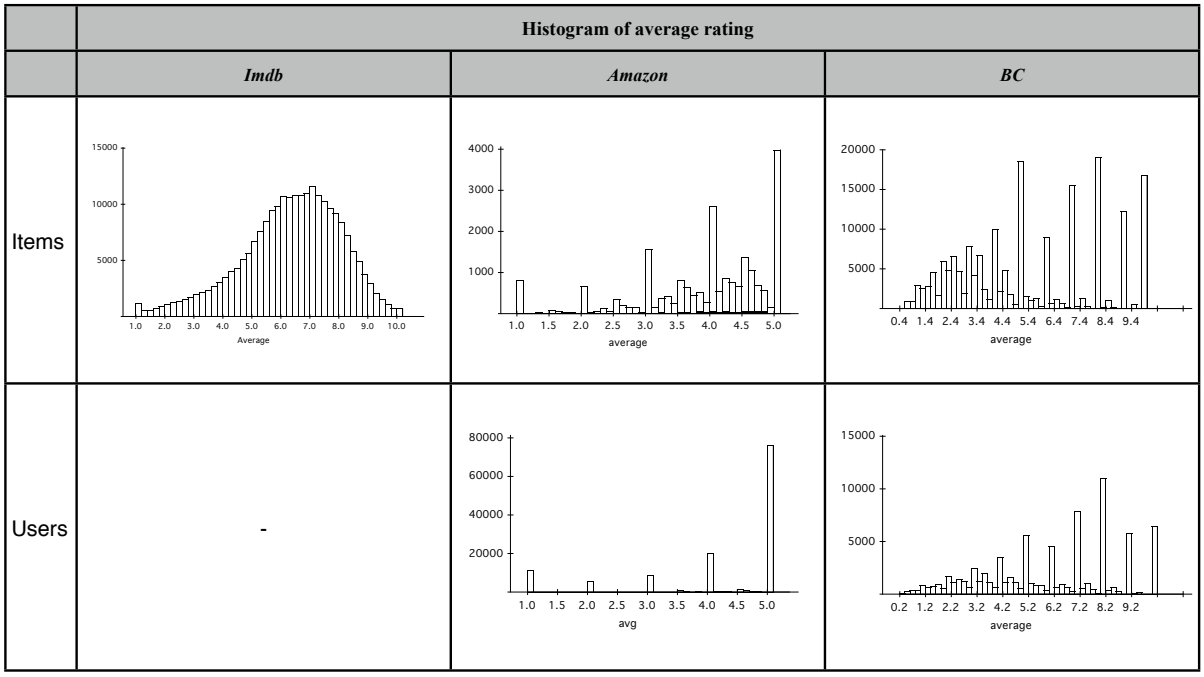

Table 4. Histograms of average rating per item and per user. X-axis is rating, Y-axis is frequency.

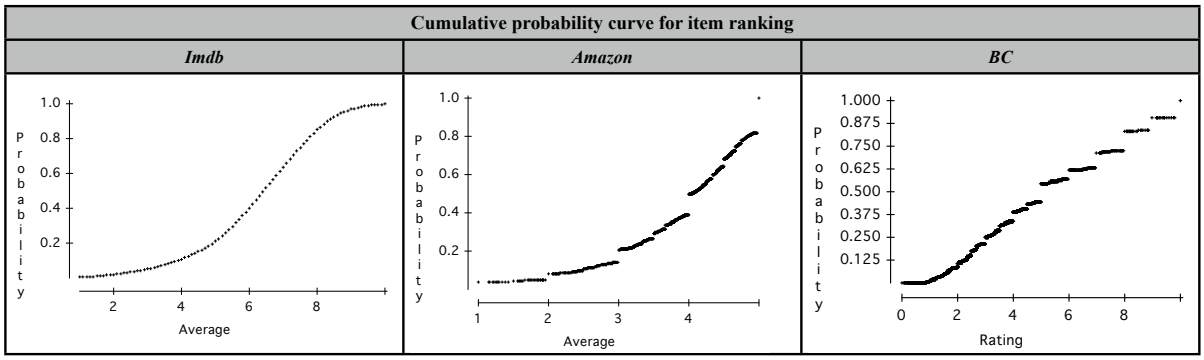

Table 5. Cumulative probability curve describing the ranking of items. X-axis is rating, Y-axis is percentage of items with rating less than the x-axis value.

between Amazon and BC (U=1411255125, p<0.0001), Amazon and Imdb (U=247742947, p<0.0001), and BC and Imdb (U=1137135, p<0.0001). The skewness and kurtosis for each distribution are: Imdb (items: -0.57, 0.16), BC (items: 0.74, -0.84; users: 0.02, -1.3), Amazon (items: -1.1, 0.77; users: -1.43, 0.82). In addition, an ANOVA showed a significant effect of website on the average rating per user (F(1,239553)=103123, p<0.0001) as well as on average rating per item (F(2,595201)=117972, p<0.0001).

Figure 1 focuses on the books that the top 5% (or expert) users reviewed on BookCrossings. On the left is a histogram of the number of votes (show in ln scale) that these books received (including votes of non-expert users), and on the right is a histogram of their average rating (including votes of non-expert users). The rating distribution of expert users on BC was significantly different from the overall distribution of user rating on BC shown in Table 4 (Mann-Whitney U=1791, p<0.0001).

Table 5 presents a cumulative probability plot of item rankings. Each point on the curves shows the percentage of items (y-axis) that have a smaller rating than the corresponding x-axis value. A Mann-Whitney U test showed that the distributions are significantly different (Amazon-BC U=421721, Amazon-Imdb U=72185, Imdb-BC U=22284, all at p<0.0001). In addition, an ANOVA showed a significant effect of website to the difference in the means of each distribution (F(2,5375)=256.04, p<0.0001), while a

V. DISCUSSION

A significant result highlighting the advantage of a quantitative approach is the power law distribution (with an exponential cut-off) of the number of votes cast by users and received by items (Figure 2). This suggests that while most items and users typically have few votes associated with them, we should not be surprised to find individuals or items with orders of magnitude more votes. Given a small sample, such extreme data could possibly be discarded as outliers, but our analysis suggests that in fact they are not. Hence, in analyzing online communities we must keep in mind that inevitably some users will cast huge number of votes, and some items will receive tremendous number of votes, while most will have very few.

In particular, from Table 2 we find that more than 50% of users on Amazon and BC cast only one vote. Similarly large portions of items receive a single vote, with the exception of Imdb where only 7% of items are rated only once. Furthermore, we see that the experts (top 5% of voters) cast as few as 7 votes on Amazon, while on BC they cast 30, suggesting that BC consists of proportionately more expert users within the community. Even so, the top 5% of popular items receive as little as 10 votes on both sites. Conversely, Imdb's popular items receive at least 662 votes. This distribution of votes can be framed as a consequence of the barrier to vote, with Imdb's lower barrier resulting in lots of votes, while BC's elaborate mechanism results in a much steeper distribution with most votes given by a very small portion of users.

Such "heavy tailed" distributions have certainly been identified in the past, and in fact have been proposed as a business opportunity for selling "less of more" [1]. Our results give us insight into who constitutes the heavy tail. We considered the top 5% of users on BC, and analyzed the items that they vote for (Figure 1). We found that experts vote for mostly obscure titles with few votes (Figure 1 left), but of both high and low rating (Figure 1 right). This can be seen as a reinterpretation of the long tail hypothesis [1], in that many obscure items of small popularity are likely to be of interest to a few select experts, while the crowds -- made up of people who buy only few times -- are likely to be interested in "popular" items.

Our analysis also highlighted significant skewing on both Amazon and BC (Table 4). As opposed to Imdb, where ratings follow a normal distribution, Amazon and BC ratings are skewed to the right, indicating overly positive voting on behalf of users. Regardless of the elaborate quality control mechanism of Amazon, with both meta-voting and meta-reviewing, we find that in fact most items receive 5 starts, and most users vote 5 stars. On the other hand, BC's distribution is less skewed, despite the complete lack of peer reviewing. These findings suggest that the quality mechanisms on Amazon, meant to eradicate unfair voting, is compelling users to be too positive when voting.

The bias of these results is also visible in Table 6, where Amazon and BC diverge considerably from the standard S-shaped curve. This table also suggests that a more appropriate way to present ratings is in terms of the percent of the population that an item is better than. Such a measure would transfer well across websites, since ratings on an absolute scale differ quite considerably.

In terms of motivational mechanisms, we find that Amazon's approach has resulted in increased votes, but only for a small portion of expert users. From Table 2 we find that even though Amazon's top 5% of products have at least 10 votes -- hence identical to BC -- the number of items with only one vote is 23%, i.e. less than half of that of BC. This suggests that many items have received an extra 2 or 3 votes, mostly from experts, as also indicated by the shallowness of the probability curve for Amazon items in Table 3.

There are a number of design recommendations we can make based on our analyses. First, it should be expected that in an online community some users will vote an order of magnitude more times than the bulk of the population. Our analyses suggests that these users can be responsive to motivational mechanisms, and can help raise the standard of the website by reducing the number of items with a single vote. An even better approach to increase the amount of voting is reducing the barrier to vote. We also find the quality control mechanisms contribute to users being too positive in their voting, and hence these mechanisms should be redesigned to help condemn too positive as well as too negative reviews. We also find that in a community of mostly experts, such as BC, voting is less biased despite the absence of quality mechanisms. In terms of representing ratings, an approach that transfers well is the use of relative scales instead of absolute, as we have shown absolute scales to be skewed.

VI. CONCLUSION

We have carried out an expert evaluation of three websites, and quantitatively analyzed users' voting behavior. We have shown considerable bias in this behavior, and in addition have framed this bias in terms of the voting mechanisms on each website. Our analysis suggests that when harnessing the "crowd's wisdom", the design features of the system should be carefully considered for their effect on aggregate behavior. Finally, our quantitative analysis indicates that such aggregate behavior cannot be captured adequately using qualitative analyses, hence these two approaches should be employed in parallel.

VII. REFERENCES

[1] Anderson, C. (2006). The Long Tail: Why the Future of Business is Selling Less of More, Hyperion.

[2] Kittur, A., Chi, E.C. and Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. CHI 2008, pp. 453-456.

[3] Kittur, A. and Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: Quality through coordination. CSCW'08, (to be published).

[4] Leino, J. and Raiha, K.J. (2007). Case amazon: ratings and reviews as part of recommendations. RecSys '07: Proceedings of the 2007 ACM conference on Recommender systems, pp. 137-140.

[5] Lerman, K. (2007). Dynamics of collaborative document rating systems. WebKDD/SNA-KDD '07:, Workshop on Web mining and social network analysis.

[6] Ozakca, M. and Lim, Y.K. (2006). A study of reviews and ratings on the internet. CHI '06, pp 1181-1186.

[7] Resnick P. and Varian H.R. (1997). Recommender systems. Communications of the ACM, vol. 40 (3): 56-58.

[8] Yang, Y. Sun, Y. Ren, J. and Yang, Q. (2007). Building Trust in Online Rating Systems Through Signal Modeling. CDCSW'07, pp. 23.