Proceedings, The Fourth AAAI Conference on
Human Computation and Crowdsourcing

# Leveraging the Contributions of the Casual Majority to Identify Appealing Web Content

**Tad Hogg**[1] and **Kristina Lerman**[2]
1. Institute for Molecular Manufacturing, Palo Alto, CA 94302
2. USC Information Sciences Institute, Marina del Rey, CA 90292

## Abstract

Users of peer production web sites differ greatly in their activity levels. A small minority are engaged contributors, while the vast majority are only casual surfers. The casual users devote little effort to evaluating the site's content and many of them visit the site only once. This churn poses a challenge for sites attempting to gauge user interest in their content. The challenge is especially severe for sites focusing on content with subjective quality, including movies, music, restaurants and items in other cultural markets. A key question is whether content evaluation should use opinions of all users or only the minority who devote significant effort to reviewing content? Using Amazon Mechanical Turk, we experimentally address this question by comparing outcomes for these two approaches. We find that the larger numbers of less informed users more than offset their noisy signals on content quality to provide rapid evaluation. However, such users are systematically biased, and the speed of their assessments comes at the expense of limited collective accuracy.

The users of peer-production and other social media web sites create far more content than any single person can absorb. To address this issue, these sites often endeavor to help their users find interesting items among the torrent of information. This is especially difficult when content quality is subjective, as is the case for music, movies, and other cultural artifacts (Salganik and Watts 2009). In these cases there are no reliable algorithmic methods for identifying higher quality items directly from the content itself. Instead, web sites crowdsource content evaluation by using peer recommendation, that is, by aggregating opinions of their users.

A major challenge for estimating content quality from user opinions is that users vary widely in how they evaluate content. For instance, users can have different *preferences* for content types, such as travel vs. sports stories, or national news vs. updates from friends. Sites can adjust for heterogeneous preferences through personalization, that is by identifying groups of users with similar preferences and recommending content to a user based on reactions by similar users.

Users can also differ in their *expertise* with respect to the web site's content. This variation is particularly important for question answering sites such as stackexchange.com or quora.com. For people repeatedly contributing to such sites, reputation systems can identify users likely to give helpful answers or, more generally, have the expertise to evaluate content. However, when content evaluation is mainly subjective, differences in expertise are not particularly relevant. Moreover, identifying content quality is important even in situations where users visit a site only occasionally or anonymously, preventing the site from establishing a user's reputation from his or her history on the site.

Even among users with the same preferences and relevant expertise, people differ greatly in their *motivation* to participate on the site and evaluate its content. Generally, only a small proportion of users devotes significant effort to the web site. For example, a small fraction of users is responsible for the majority of the news stories submitted to the social news aggregator Digg, and a small fraction of editors is responsible for the vast majority of edits of Wikipedia articles (Wilkinson 2008). Similarly, in a peer recommendation task, users devote vastly different levels of effort to evaluating the content on the sites.

A web site could attempt to identify motivated users by their behaviors on the site. These behaviors can include the time the user spends on the site, requests to view details of particular content, frequency and number of repeated visits to the site, substantial contribution of content (for peer production web sites), and ratings by other users (for sites with reputation systems). Nevertheless, contributors to peer production sites have a variety of motivations (Benkler and others 2015), which can be difficult to infer from readily observable behavior on the site.

In this paper, we focus on one aspect of diversity of motivation that is relatively easy to determine, namely the extent to which users access additional information on content they choose to rate. In particular, we focus on the common situation where the web site provides a short summary of content to users, along with links to additional information. Users viewing the content can choose to rate it based on the summary alone, or spend more time learning about the content by following the provided links. We call the latter users *informed* with respect to that content.

An important question for aggregating user opinions is how to account for these user differences. In particular, to what extent can web sites incorporate opinions of all users to estimate subjective content quality, for faster evaluation,

without sacrificing the accuracy that comes from considering only the opinions of the minority of users who make the effort at detailed evaluation? Furthermore, how does this tradeoff depend on the desired information about quality, e.g., precise values or relative rankings, either for all for just some content? For instance, the web site may only require identifying the highest quality content, as opposed to accurately evaluating all content, since that is what most users will find interesting.

Addressing this question is difficult due to the many confounding factors that affect whether and how users react to content. A major factor is how content is presented to users. Due to the cognitive heuristic known as position bias (Payne 1951), people pay more attention to items at the top of a web page or a list of items than those below them (Lerman and Hogg 2014). Other factors affecting outcomes of the recommendation task include heterogeneity of content quality, its changing relevance (novelty), commonality of user preferences (homophily), and social influence (when showing users a summary of prior users' reaction to the content).

To address some of these confounding factors, we studied this tradeoff for quality estimation with randomized experiments on Amazon Mechanical Turk (Mturk), which is a popular platform for experimental behavioral research (Bohannon 2011; Horton, Rand, and Zeckhauser 2011; Mason and Suri 2012; Crump, McDonnell, and Gureckis 2013; Kittur et al. 2013). These experiments allow us to compare different approaches to aggregating users' reactions to the same content while controlling content ordering. The particularly challenging case, which is the focus of this paper, is when the site attracts many new users, users rarely return, or are anonymous. Responding to heterogeneity of user motivation in this case requires the web site classify the likely accuracy of users' ratings from their behavior during a single session on the site.

In the remainder of this paper, we first discuss related studies of user motivation, then describe our experiments and quality estimation procedures. We then show how these procedures perform for our experiments and discuss consequences for web sites. Identifying varying levels of motivation and accounting for them algorithmically could improve the performance of using crowdsourcing to rate content with subjective quality.

## Related Work

Previous studies attempted to estimate content quality by modeling available data of user reactions to content. Specifically, some studies used observational data from well-established social media sites to estimate parameters of stochastic models that include quality. For instance, studies of news aggregators (Hogg and Lerman 2012; Stoddard 2015) used such models to estimate quality and use it to predict users' future response to content. These studies did not differentiate between users based on their motivation nor account for the varying levels of effort users devote to the evaluation task.

Experiments using MTurk allow evaluating some consequences of manipulating user motivation in crowdsourcing tasks. One such study varied wage rate and the required task effort to crowdsource evaluation of search result relevance (Kazai 2011). This allowed evaluating the effect of MTurk task design on the quality of results, defined by agreement with expert judges. This contrasts with our study of situations common in peer production web sites where quality is defined by the user community itself. Another study varied the framing of an MTurk image classification task (Chandler and Kapelner 2013). This showed how varying the apparent meaningfulness of the task affected the quality of the crowdsourced outcome, again in a setting with independent expert assessment of content quality. This setting, in which the task designer creates the rationale for the task in users' minds, contrasts with user activity on peer production web sites based on their own interest in the site's content.

Other studies explicitly accounted for how a group's diversity affects its performance. Page (Page 2010) identified three types of diversity: variation of attributes, variation in the amount of the attribute individual has, and variation in the composition of the group. While studies have attempted to quantitatively characterize individual's expertise, in order to automatically identify varying levels of expertise in a group, fewer studies looked directly of motivation (Ginsca and Popescu 2013; Pal, Chang, and Konstan 2012; Gkotsis et al. 2014; Yang et al. 2014).

A series of studies examined how to aggregate decisions of many people to produce outcomes that are far more complete and accurate that what any individual knows (Hemmer, Steyvers, and Miller 2010; Miller et al. 2009; Yi, Steyvers, and Lee 2012). These studies assumed that what each individual within a crowd knows is a noisy version of the ground truth, and used Bayesian methods to aggregate many noisy judgements. We use a similar approach to aggregate the judgements of uninformed users about the quality of items, viewing these judgements as a noisy version of an informed users's judgements. Unlike previous works, we also explore the accuracy vs. speed tradeoffs in the estimation of quality using judgements of uninformed users. These tradeoffs are of practical interest to web site providers.

## Experiments

We created a simplified version of peer production web sites and focused on how users *rate* content rather than produce it. Thus we used a fixed set of content, thereby removing variability due to the site's content changing between successive user visits. Specifically, we presented people with a list of science articles and asked them to recommend those that they thought report important scientific problems. This task serves as a proxy for users of peer production sites choosing content they view as important to like or forward to other users. The user recruitment and vetting procedures are described in (Lerman and Hogg 2014).

The experiment's user interface, shown in Figure 1, displayed the title and summary of each story. The title was linked to the full story via its url. Thus people could read the full story by clicking on the title, but were not required to do so. We refer to such actions as "url clicks".

The optional nature of url clicks is key to our study of how users' information affects their behavior. Users decide

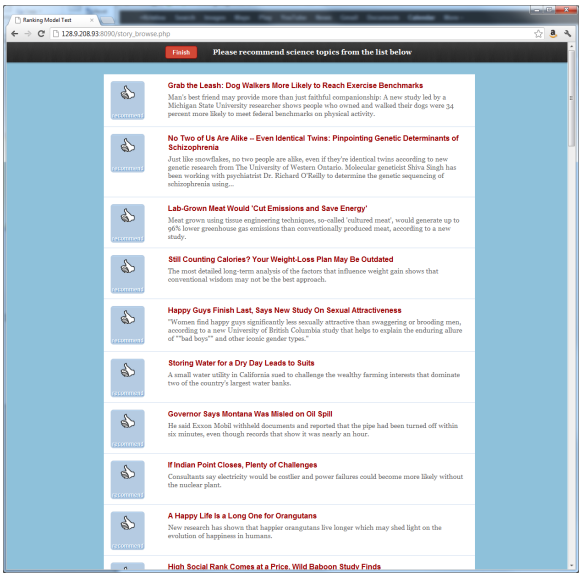


Figure 1: Screenshot of a web page shown to users.

for themselves whether they wish to gain more information about a story to evaluate it. Thus users self-select into groups with more or less information about a story. This corresponds to available options for users on web sites where they can rate content based on a short summary or examine the content in more detail. In addition, url clicks provide a measurable experimental signal about which users viewed the full story. This contrasts with the lack of experimental measurement of which story summaries a user views if they do not click on the url.

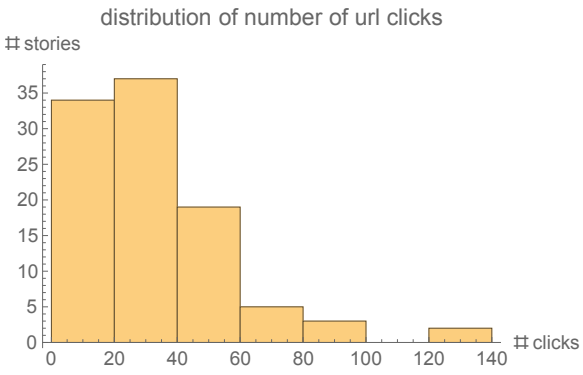


Figure 2: Distribution of number of url clicks on the 100 stories used in the experiments.

Figure 2 shows the range in number of url clicks the stories received. The number of clicks ranged from 4 to 134, with a median of 25 url clicks per story. Of the 3498 total users who participated in the experiments, 816 clicked on at least one url.

Prior experiments (Lerman and Hogg 2014) measured how story position within the list of stories affects how many votes it receives and compared several policies for ordering

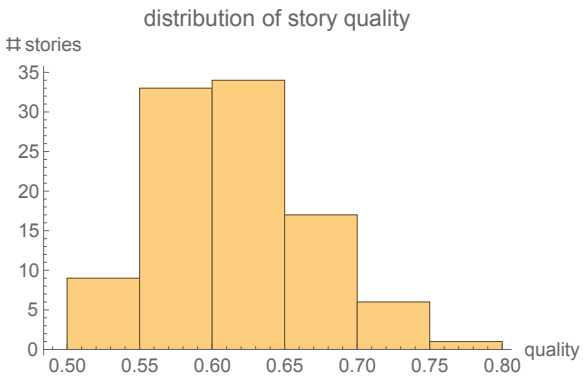


Figure 3: Distribution of quality values for the 100 stories.

the stories. In this paper, we focus on content evaluations made with two of the orderings, which involve a total of 485 users (see Table 1). First, the *random* ordering presented the stories in a new random order for each user. This is a useful control condition in averaging over the effect of position bias. Second, the *recency-of-activity* ordering presented stories in chronological order of the latest vote they received, with the story with the most recent vote at the top of the list. This is similar to Twitter's policy of displaying the most recently tweeted or retweeted post at the top of the followers' streams.

## Evaluating Story Quality

Evaluating the tradeoff between speed and accuracy of estimating story quality with different groups of users has two requirements. First, determining the actual quality of each story to compare with estimates as users arrive at the site. To do so, we use the full history of user reactions to the stories. Therefore, this is a retrospective measure, which is only available long after a story was first presented on the web site. The second requirement is a method to estimate quality as a function of time as users arrive on the site after a story is first presented. The remainder of this section describes these methods.

### Measuring Actual Quality

We operationally define the *quality* of a story as the probability a user votes for it after viewing the full content by clicking on that story's url. This measure reflects the subjective view of the user community in our experiments, and is not necessarily related to writing style or the broader importance of the story's topic. This operational measure is suitable for the situation discussed in this paper, where quality is a subjective evaluation by the user community (Stoddard 2015).

Specifically, to evaluate the quality of a story $s$, we identify users in our experiments who clicked on that story's url. We call such users the *informed users* for that story. The distribution of the number of such users among the stories is same as the distribution of url clicks shown in Figure 2.

Suppose a given story has $M$ informed users, who give the story $V$ votes. The simplest way to use the response by

informed users to estimate story quality is $q = V/M$, i.e., the fraction of such users who voted for the story. However, this ratio of votes to clicks is a noisy estimate, particularly for stories that receive few clicks. To reduce this noise and improve prediction of quality for new stories, we combine the observed behavior of informed users with a prior distribution. Since we define quality as a probability, a convenient prior is the beta distribution $B_{a,b}(q)$ for $q \in (0,1)$, whose probability density is proportional to $q^{a-1}(1-q)^{b-1}$. A uniform prior corresponds to $a = b = 1$.

We estimate story qualities in two steps: determining the prior distribution and then using the prior to determine the quality of each story. For a web site, the prior distribution could come from retrospective evaluation of many items previously presented on the site. Assuming user community preferences and the nature of the content both change relatively slowly, this historical approach provides a helpful prior distribution for evaluating new content. In our experimental setting, we use a similar procedure based on cross validation. For a given set of stories as a training set, we could determine the parameters $a, b$ of the prior distribution as those maximizing the log-likelihood $L$ of the observed number of votes by the informed users on each story in the training set. However, this tends to overfit to the training set. To reduce this overfitting, we include a regularizer favoring a uniform prior, specifically finding parameters $a, b$ that maximize $L_{\text{reg}} = L - \lambda((a-1)^2 + (b-1)^2)$ for a given choice of weight $\lambda$. We determine $\lambda$ by 10-fold cross validation (Abu-Mostafa, Magdon-Ismail, and Lin 2012). Specifically, we split the 100 stories into 10 groups of 10 stories each. Taking each of the groups in turn as the test set, we use the other nine groups as the corresponding training set. We pick the weight $\lambda$ that gives the maximum accuracy in predicting the number of votes received by stories in the test set when averaged over the ten choices for the test set. We then select the prior distribution parameters $a, b$ that maximize $L_{\text{reg}}$ with this value of $\lambda$.

Using this prior distribution, we determine the quality of each story as its maximum likelihood value. Specifically, a story with quality $q$ and $M$ informed users receives $V$ votes according to the binomial distribution $\text{Bi}(M, q; V)$. Combining with the prior for $q$, we estimate the quality of this story as the value of $q$ that maximizes $\text{Bi}(M, q; V)B_{a,b}(q)$, which is

$$\frac{a + V - 1}{a + b + M - 2} \tag{1}$$

This reduces to $V/M$ for a uniform prior. Figure 3 shows the distribution of quality values resulting from this procedure.

## On-line Estimation of Quality

The above estimation of story quality is an off-line procedure, determined after presenting the stories to all users in our experiments. In practice, web sites need to rapidly identify the quality of new content. This requires estimates based on only the reactions of the small set of users who visit the site soon after the story is posted.

One option for such on-line estimation is to apply the above quality estimation procedure to the reactions of the informed users as they visit the site. Informed users, by our definition, give the most reliable assessment of story quality. Unfortunately, few users make the effort to examine the full story when deciding whether to vote for it. Hence, applying this procedure to early estimates will only have a few users and hence provide only rough quality estimates. Instead, at these early times, the estimation is dominated by the prior distribution of story quality, giving estimates close to the middle of the quality range. This means stories with particularly high or low qualities will be poorly estimated by this procedure. The lack of informed users thus limits the effectiveness of this approach to early story evaluation.

An alternative estimation method aggregates the reactions of all users, whether or not they click on the story's url. The vast majority of such users are uninformed, i.e., they do not click on the url. Thus the effectiveness of this procedure depends on how well user response to just the story summary reflects its actual quality.

In our experiments, uninformed users are less likely to vote for a story than informed users. This arises from two factors. First, uninformed users may not see the story at all. We refer to this factor as the story's *visibility*, which depends on where the story is presented to the user among the list of all stories. We define $v_p$ as the probability a user views a story presented at position $p$ in the list. Second, even when uninformed users see a story, they do not vote for it as often as informed users do. To quantify this difference, we define the *appeal* $r_s$ of story $s$ as the probability an uninformed user votes for $s$ after viewing it. Thus, the probability an uninformed user votes for story $s$ is $\rho_s = v_{p_s} r_s$ where $p_s$ is the position of the story shown to that user.

Our experiments do not directly measure which stories users view. Instead we estimate visibility $v_p$ and appeal $r_s$ from a model of how users view and vote for stories (Hogg and Lerman 2015). We find that appeal and quality have $42\%$ correlation and are related as

$$r_s = \alpha q_s + \delta_s \tag{2}$$

where $\alpha = 0.423$ and the residual $\delta_s$ is a random variable with zero mean and standard deviation $\sigma = 0.1$. Hence the votes by uninformed users, determined by a story's appeal, give a noisy indication of the story's quality, as determined by informed users.

We use Eq. (2) for estimation by treating votes by uninformed users as a noisy signal of the story quality. For a given story, we start with the prior distribution of story quality. Then with each uninformed user we update the posterior distribution for that story's quality based on whether that user voted for the story and the position of the story in the list shown to that user. This Bayesian procedure involves the probability for a vote $\rho_s = v_p r_s$, described above, and Eq. (2). With this procedure, the actions of users who have visited the site by a given time produce a posterior distribution for the qualities for each story. We take the expected value of those distributions as the quality estimates at that time.

A web site implementing this procedure would have access to the reactions of all users, both informed and uninformed, at a given time. Thus, as a minor improvement, we

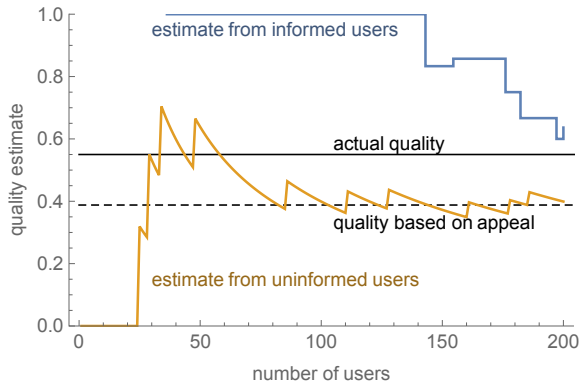


Figure 4: Schematic of quality estimation processes for one story. In this illustration, informed users are about 10% of the user population, which is considerably larger than in our experiments.

extend the procedure for uninformed users to also update the posterior quality distribution with votes from informed users, using the method described above for informed users. In our experiments, informed users are such a small fraction of the user population that these updates have little effect on the result.

### Example of Estimation Methods

The two estimation methods described above – using just informed users and using all users – have differing tradeoffs between speed and accuracy. Before comparing these procedures for our experiments, we illustrate the tradeoff with simulated data using simplified versions of the procedures.

As an illustration of this behavior, consider a story for which informed users vote for with probability $q$. A simple estimate of quality is the fraction of such users who vote for the story. If $M$ informed users give $V$ votes, the estimate $\hat{q} = V/M$ has expected value $q$ and variance $q(1-q)/M$. Thus the error in this estimation decreases to zero as $1/\sqrt{M}$. However, informed users are only a small fraction of visitors to the web site, so achieving an accurate estimate requires a long time.

Uninformed users vote for the story with probability $\rho$, which for this illustration we take to be the same for all uninformed users. This amounts to averaging over the visibility of a story in all possible positions, i.e., taking $\rho = \langle v_p \rangle r_s$ for story $s$.

In this case, suppose that $N$ uninformed users give $U$ votes. Then $\hat{\rho} = U/N$ has expected value $\rho$ and variance $\rho(1-\rho)/N$. This allows estimating the story's appeal and hence its quality from Eq. (2) as $\hat{q} = \hat{\rho}/(\alpha \langle v_p \rangle)$. The expected value of this estimate is the story's quality. There are two contributions to the variance of this estimate. First, the variance associated with the sampling of users' votes, namely $\rho(1-\rho)/(N\alpha \langle v_p \rangle)$. Second, the variance in the relation between appeal and quality in Eq. (2), i.e., $\sigma^2$. The first of these terms decreases as $1/N$, but the second is independent of the number of users, and reflects the difference between how informed and uninformed users evaluate stories. Thus errors in estimating quality from uninformed users do not decrease toward zero as the number of users increases.

| ordering policy | random | recency of activity |
|---|---|---|
| number of users | 199 | 286 |
| number of votes | 1873 | 2586 |
| number of url clicks | 164 | 246 |

Table 1: Experiments for testing quality evaluation.

As an example of evaluating a single story, Figure 4 illustrates the competition between accuracy of these quality estimates and their timeliness, which depends on the rate users visit a web site. Informed users give unbiased estimates of quality. However, relying only on such users requires a long time for quality estimates to converge, since informed users are only a small fraction of the user population. On the other hand, estimates from uninformed users converge rapidly, but the estimates are systematically biased, limiting the achievable accuracy even after many such users visit the site.

For simplicity, this illustration ignores the prior distribution of quality values and does not account for differing visibility of the story to uninformed users based on the story's position in the list shown to each user. These factors, included with the results described below, do not change the qualitative nature of the tradeoff seen from this discussion.

## Results

To estimate story quality from the reactions of early users exposed to the story, we apply the two estimation procedures described above: 1) estimate quality based only on reactions by informed users, and 2) use reactions from all users.

Table 1 describes the experiments used to evaluate quality estimates as a function of number of users providing evaluations. These include examples of both the random and recency-of-activity story ranking policies. These are just two of the orderings used in our full set of experiments.

### Rapidly Identifying High-Quality Content

As a specific example, we consider the random ordering experiments to illustrate the usefulness of estimating quality with uninformed users. This ordering policy presents the stories in a different random order to each user, thus averaging over positions. Figure 5 compares quality estimates based on all users with those based on informed users only. For each successive user, we estimate quality of every story with the procedures described above and show the average of the error, $|\hat{q_s}(n) - q_s|$, over all the stories. Here $\hat{q_s}(n)$ is the quality estimate for story $s$ after $n$ users have visited the site. Only a small fraction of these users are informed, giving the relatively small change in estimates when considering informed users only.

We find uninformed users are not helpful for estimating quality of most stories, because their quality values are already well-estimated by the prior distribution. However, for

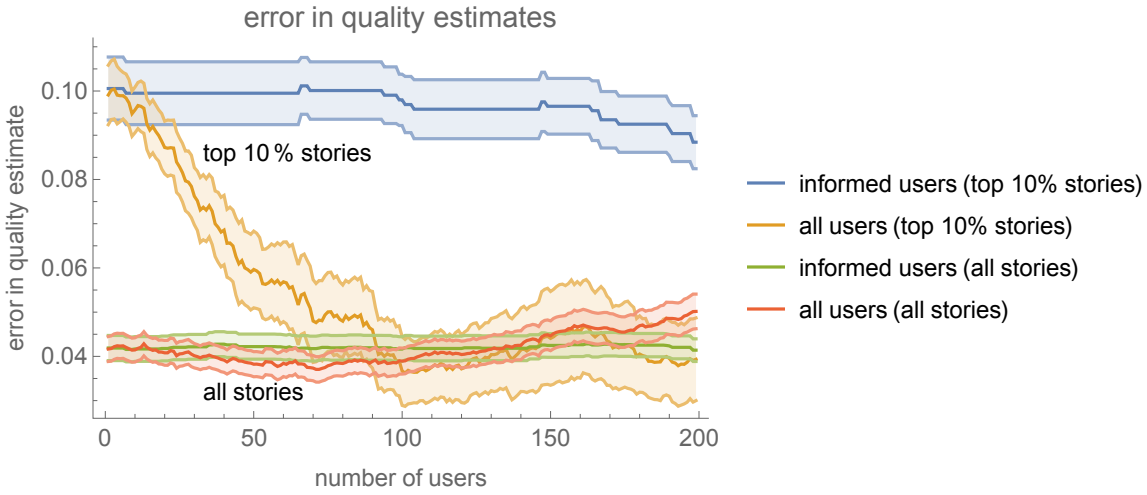


Figure 5: Error of quality estimates vs. number of users, from all users and from only informed users, for all stories and for top stories (highest 10% quality). The thick curves show the average error for the stories in each group, and the shading indicates the standard error of the average, i.e., the standard deviation of the errors divided by the square root of the number of stories (100 and 10 for all and top 10%, respectively).

top quality stories, uninformed users rapidly improve the estimates compared to waiting for a sufficient number of informed users. Moreover, Figure 5 shows there is a limit to the accuracy obtained from estimating quality based on all users, reflecting the lack of perfect correlation between appeal and quality. This behavior is qualitatively the same as schematically illustrated for a single story in Figure 4.

In addition to rapid reduction in error averaged over all stories, estimates based on all users also reduce the standard deviation of the posterior distribution of the quality estimates more rapidly than relying only on informed users, as shown in Figure 6. This results in smaller confidence intervals for the estimates of individual stories. This reduction could be useful for prediction or deciding to highlight content, e.g., where false positives (showing users content that is not actually high-quality) are more harmful than false negatives (missing some high-quality content).

Another perspective on estimation is how well quality estimates allow ordering the stories according to their quality. For helping a web site decide what content to highlight, identifying high-quality content can be more important than knowing the exact quality of the content. In our case, estimating quality after a given number of evaluations gives a ranking to the stories, namely the ordering corresponding to the decreasing order of their estimated quality values.

One measure of ranking performance relevant to highlighting quality content is the precision of the ranking, e.g., what fraction of the first ten positions in the ranked list contain stories with top 10% quality. We find estimates based on all users increase precision more rapidly than waiting for the more accurate estimates of informed users.

Precision evaluates a ranking based on a fixed number of positions, all of which are given the same weighting. However, user attention rapidly decreases with position in the list. Thus rankings that place best (high-quality) stories closer to the top of the list are better, though there is not a sharp threshold as implied by a choice of number of top stories used to define precision. A measure of rankings that captures this variation is the *discounted cumulative gain* One version of this measure is

$$\text{DCG} = \sum_p \frac{2^{q_{s_p}} - 1}{\log_2(p+1)} \tag{3}$$

where the sum is over positions $p$ in the list of stories, in descending order of estimated quality, with $q_{s_p}$ the actual quality of the story $s_p$ in that position. The maximum possible value for DCG is when stories are ordered according to their actual qualities. The *normalized* DCG is the measured value divided by its maximum possible value.

Some stories may have the same estimated qualities, particularly in the early stages when there are only a few users so that most stories have not yet received any votes. All such stories will have the same estimate, namely the expected value of the prior distribution of quality values. For these ties, the stories could equally well be presented in any order. For our evaluation of discounted cumulative gain, we use the average value over all possible orderings of stories with equal estimates.

Figure 7 compares the orderings produced by estimates from all users with those from informed users only. Rankings produced by informed users are little changed during the time shown in the figure, due to the small proportion of informed users. Estimates based on all users, on the other hand, rapidly improve DCG toward its maximum possible value. Thus, the typical estimation error of $0.04$ seen in Figure 5 is small enough compared to the variation in quality values (Figure 3) to give good rankings of the stories. Nevertheless, the relatively narrow range of story qualities means the DCG value is relatively large even before any users have seen the stories. Before any users visit the site, all stories

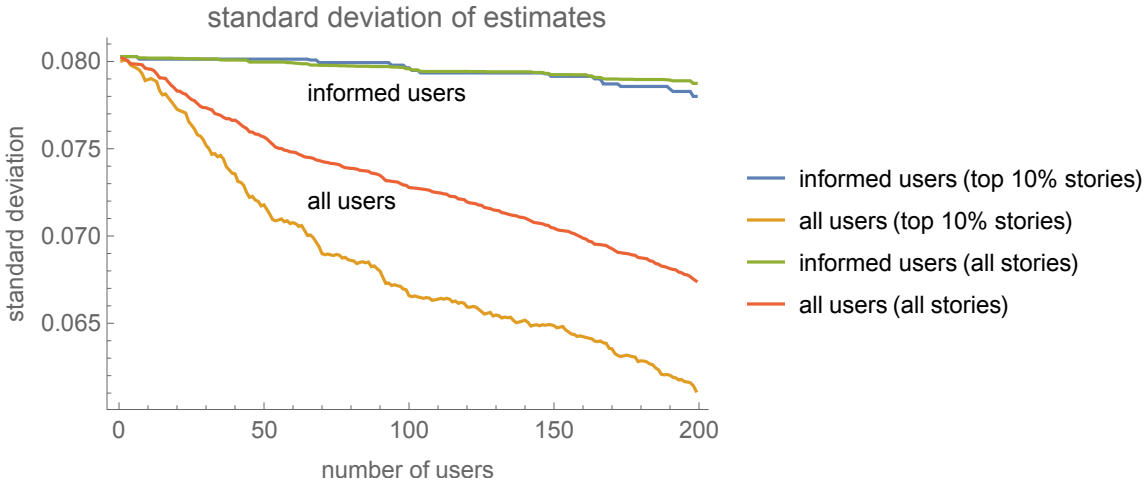


Figure 6: Standard deviation of the posterior distribution of quality estimates, from all users and from only informed users, averaged over all and over just the top 10% of stories.

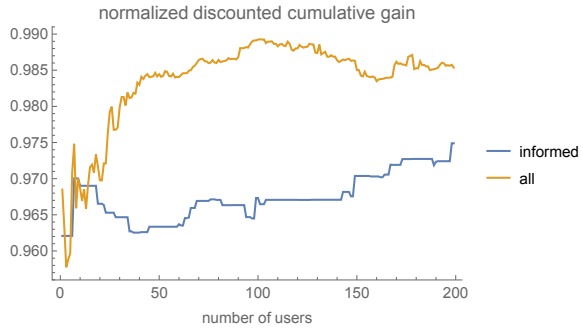


Figure 7: Normalized discounted cumulative gain of stories ordered according to estimated quality vs. number of users, from all users and from only informed users.

have the same estimate, so DCG is the average value over all possible orderings. In our case, this prior value is not much smaller than the value for the optimal ordering.

## Effect of Story Ordering

The above discussion considered showing the stories with random ordering. In practice, web sites present stories in different order to highlight content users are likely to find novel and interesting. For instance, ordering by popularity emphasizes stories highly recommended by previous users.

The random ordering, used above, averages over position bias. For a web site mainly interested in evaluating high-quality content, an ordering policy that tends to place higher quality content in more visible positions at the top of a list will direct user attention to that content, and hence obtain more user reactions to the content. This also has the effect of improving user experience by making it more likely they see high-quality content when they visit the site. One such ordering is the recency-of-activity ordering.

Prior experiments (Lerman and Hogg 2014) show that recency-of-activity ordering gives rapid convergence for estimating a story's appeal. Hence, with the recency-of-activity ordering, uninformed users see higher appeal stories more often than in a random ordering. This suggests recency-of-activity ordering will enhance the performance of uninformed users in the quality estimation because 1) appeal and quality are somewhat correlated, and 2) as we saw with random ordering, uninformed users give a useful signal for quality estimation for high-quality stories.

In contrast to this expectation, the recency-of-activity and random orderings have similar improvement in estimating quality using uninformed users. Specifically, Figure 8 shows the behavior of quality estimation when stories are presented using the recency-of-activity ordering. This is similar to the result for random ordering in Figure 5. Similarly, the two orderings have similar behavior for ranking stories, as shown by comparing Figure 9 with Figure 7. The similar behavior of random and recency-of-activity orderings arises because estimates from uninformed users already reach their best possible values, i.e., accurately estimating appeal, after receiving reactions from about 50 users (see Figure 5). However, the recency-of-activity ordering requires about that many users to collect enough user reactions to stories to significantly affect the ordering (Lerman and Hogg 2014). Thus recency-of-activity ordering does not direct attention to high-quality stories rapidly enough to improve early quality estimation compared to a random ordering.

## Discussion

Our experiments evaluate user behavior repeatedly under different conditions while controlling the content and how it is displayed. Randomly assigning users to these conditions avoids self-selection that may confound observational studies of content evaluation on web sites. As with other experiments on human behavior, a tradeoff for this control is a simplification of actual user experience. Moreover, online experiments have challenges not found in laboratory experi-

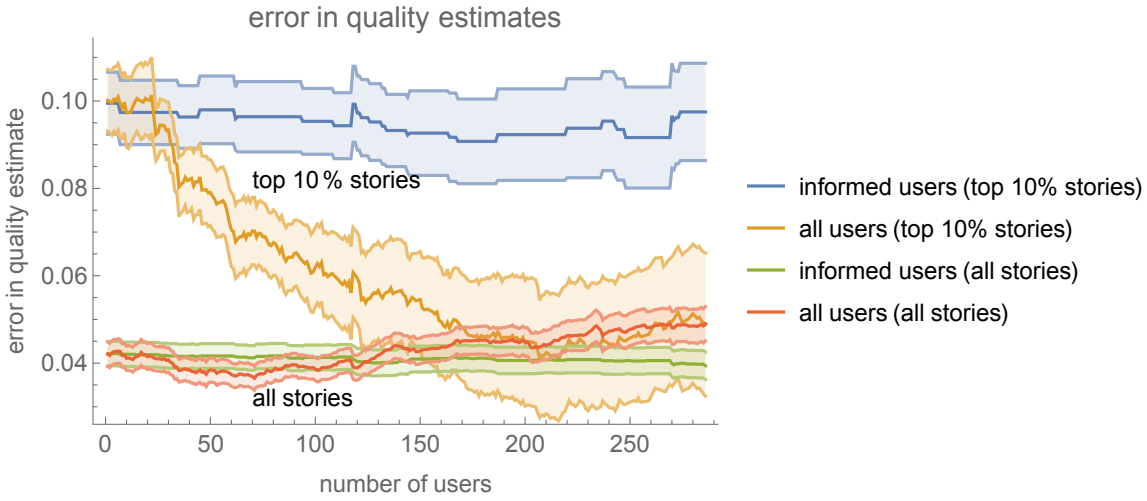


Figure 8: Error of quality estimates vs. number of users for recency-of-activity ordering, from all users and from only informed users, for all stories and for top stories (with top 10% quality). The thick curves show the average error for the stories in each group, and the shading indicates the standard error of the average, i.e., the standard deviation of the errors divided by the square root of the number of stories (100 and 10 for all and top 10%, respectively).

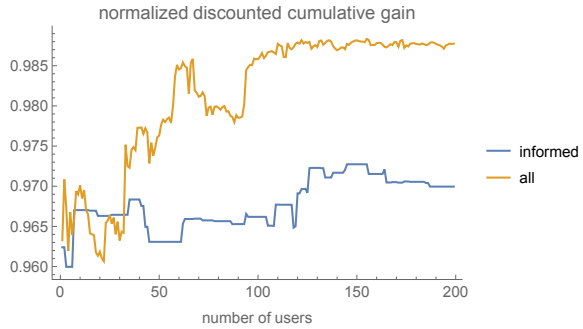


Figure 9: Normalized discounted cumulative gain of stories ordered according to estimated quality vs. number of users for recency-of-activity ordering, from all users and from only informed users.

ments, particularly with respect to controlling the user's environment during the experiment and the possibility of spam users. To help address this problem, MTurk provides a reputation mechanism for its workers, which we included in our user vetting procedures (Lerman and Hogg 2014).

Another difference between MTurk and web sites is that MTurk is a labor market whose participants are paid, whereas peer production users are not. In some cases, publicly visible reputation systems can substitute for financial incentives as users compete for status in the user community (Loch, Huberman, and Stout 2000; Nov, Naaman, and Ye 2009). But in general, payments can alter user motivation (Bowles 2008), e.g., how much people focus on personal gain compared to helping improve the web site for other users by accurately rating content.

These differences are an important caveat to the following discussion of how our results might apply in practice. While online and laboratory experiments produce similar results in at least some cases (Horton, Rand, and Zeckhauser 2011), addressing the significance of the differences for our results on content evaluation will require additional studies.

Most people devote little effort to evaluate quality of content on web sites before deciding whether to endorse or forward it to friends. Nevertheless, we find that the choices of such users can help identify high-quality content more rapidly than waiting for a sufficient number of highly motivated users to visit the site. The extent to which this applies to web sites depends on several key properties of the user population: 1) what proportion of a web site's visitors choose to become informed users, 2) how accurately the site identifies such users from their behavior on the site, and 3) how much better informed users are in determining content quality.

When informed users are fairly common, readily identifiable and much more informative than other users, then the web site can use them exclusively to rate content. In this case, quality estimates will be accurate and rapid enough to allow the site to incorporate quality estimates in delivering content to subsequent visitors. On the other hand, when informed users are relatively rare and other users provide useful, if not entirely accurate, quality estimates, the web site can benefit from incorporating responses from all users. This latter case, which applies to our experiments, is particularly relevant when informed users are rare and there an no reliable alternatives to estimating quality, i.e., when quality is mainly subjective.

Quickly obtaining rough estimates of quality, particularly for the best stories, could allow the web site to highlight the best new content quickly. Moreover, estimated quality values could be shown to users, providing them a better guide to quality than, for example, raw scores (such as number of

votes), which do not distinguish reactions from informed or uninformed users. For instance, for retweeting posts, instead of showing how many people shared an article, show its estimated quality. The estimate, by accounting for differences in votes arising from visibility, should more accurately reflect the quality of the content. Such signals can shift user attention to content with high signal values (Hogg and Lerman 2015), which can increase the number of evaluations for these items.

This observation increases in importance when web sites provide additional information about content that, as a side effect, further reduces the fraction of users who evaluate content in detail. For instance, social influence, from showing users the reactions of prior users, can have this effect. Specifically in our experimental setting, social influence leads to fewer url clicks (Hogg and Lerman 2015). This reduces effort required of users to locate appealing content, thereby benefiting the users. However, social influence also results in fewer informed users. This increases the importance of utilizing the noisy behavior of uninformed users to speed up quality estimation.

We discussed dividing users into two groups: the few users whose behavior indicates more deliberation on their evaluation, and the majority of users who devote less effort. More generally there could be a finer division, or continuous weights, associated with users. Behaviors that could lead to such finer divisions include the number of url clicks or time a user spends on the task. Quality estimation could use these weights to determine how much each user's behavior contributes to the estimates.

A lesson for web site design from these results is to include optional measurable actions users can take that may distinguish knowledgeable and motivated users from others. In our experiments, this optional action was the url click that provided more information on a story. More generally, interfaces could provide users a variety of optional tools to learn more about content and adjust how the site selects content for that user (Bostandjiev, ODonovan, and Hollerer 2012; Forbes, Savage, and Hollerer 2012). Actions a user chooses to take with these options could improve that user's experience on the site. Of more relevance to the discussion in this paper, such actions also indicate the user's level of engagement with the content, so could be used to weight users' reactions when aggregating them to estimate content quality.

Extending beyond our focus on the worst-case scenario of users without extensive history on the site, other types of information include the user's history on the site, e.g., how long since they joined and how active they have been in posting content and responding to others' posts. Comparing their responses to that of the population average could identify users with high motivation, expertise and common preferences to other users. Another approach is for the web site to allow users to rate the helpfulness of other users' activity, e.g., quality of comment or answer to a question, either explicitly (e.g., with a reputation system for users or allowing users to link to others whose ratings they find particularly relevant) or implicitly by giving high ratings (e.g., voting for) content contributed by that user. While these approaches can be useful in exploiting user heterogeneity, they require more action on the part of users than their immediate reactions to new content. Thus it is useful to identify methods of exploiting heterogeneous populations without adding to the complexity of the web site or relying on additional user activities. The scenario considered in this paper is one such approach.

Another lesson for web site design is to attempt to improve the parameters relevant for the tradeoff, namely increasing the number of informed users or the accuracy of reactions from the majority of users. One key aspect of this approach is the amount of effort required by users to provide their opinion on content. Common examples of a range of effort include a simple vote or like, adding a comment, or writing a review. Reducing effort required by users is one way web sites can improve the fraction of users motivated enough to contribute evaluations. Another aspect is the reward a user receives for rating content, particularly if the ratings are helpful to other users.

A future direction for evaluating tradeoffs related to user motivation is extending our experiments to change story ordering based on estimated quality (Van Hentenryck et al. 2016). Such an adaptive visibility ordering could exploit position bias to improve user experience and simultaneously get more user reactions to content that is likely to be of high quality. Specifically, due to position bias, users are more likely to view stories near top of list. Thus, stories shown in those positions get more informative reactions than other stories. This is especially true for interpreting a lack of vote: when the content is in a highly-visible position, a lack of vote is more indicative of lower quality than for stories at less visible positions. Hence a web site could adaptively adjust content ordering to direct user evaluation toward stories requiring additional evaluation. For instance, to speed up identifying top stories, the ordering should emphasize stories whose quality estimate's confidence interval includes top quality values. These would be either new stories with few evaluations so far, or stories that have received enough votes to appear to possibly be top quality, with additional votes needed for confirmation.

The experimental approach described in this paper can help address these extensions in a controlled setting, and thereby compare multiple procedures. Those that appear successful in experimental settings could then be tested on actual web sites.

## Acknowledgments

This work was funded, in part, by the Army Research Office under contract W911NF-15-1-0142 and by the National Science Foundation under grant SMA-1360058.

## References

Abu-Mostafa, Y. S.; Magdon-Ismail, M.; and Lin, H.-T. 2012. *Learning From Data*. AMLBook.

Benkler, Y., et al. 2015. Peer production: A form of collective intelligence. In Malone, T., and Bernstein, M., eds., *Handbook of Collective Intelligence*. MIT Press. 175–204.

Bohannon, J. 2011. Social science for pennies. *Science* 334:307.

Bostandjiev, S.; ODonovan, J.; and Hollerer, T. 2012. Tasteweights: A visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems (RecSys12)*, 35–42. ACM.

Bowles, S. 2008. Policies designed for self-interested citizens may undermine "the moral sentiments": Evidence from economic experiments. *Science* 320:1605–1609.

Chandler, D., and Kapelner, A. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *J. of Economic Behavior and Organization* 90:123–133.

Crump, M. J. C.; McDonnell, J. V.; and Gureckis, T. M. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLos ONE* 8:e57410.

Forbes, A. G.; Savage, S.; and Hollerer, T. 2012. Visualizing and verifying directed social queries. In *Proc. of IEEE Workshop on Interactive Visual Text Analytics*.

Ginsca, A. L., and Popescu, A. 2013. User profiling for answer quality assessment in q&a communities. In *Proceedings of the 2103 workshop on Data-driven user behavioral modelling and mining from social media*, 25–28. ACM.

Gkotsis, G.; Stepanyan, K.; Pedrinaci, C.; Domingue, J.; and Liakata, M. 2014. It's all in the content: state of the art best answer prediction based on discretisation of shallow linguistic features. In *Proceedings of the 2014 ACM conference on Web science*, 202–210. ACM.

Hemmer, P.; Steyvers, M.; and Miller, B. 2010. The wisdom of crowds with informative priors. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*.

Hogg, T., and Lerman, K. 2012. Social dynamics of Digg. *EPJ Data Science* 1(5).

Hogg, T., and Lerman, K. 2015. Disentangling the effects of social signals. *Human Computation Journal* 2(2):189–208.

Horton, J. J.; Rand, D. G.; and Zeckhauser, R. J. 2011. The online laboratory: conducting experiments in a real labor market. *Experimental Economics* 14:399–425.

Kazai, G. 2011. In search of quality in crowdsourcing for search engine evaluation. In *Advances in information retrieval*. Berlin: Springer. 165–176.

Kittur, A.; Nickerson, J. V.; Bernstein, M.; Gerber, E.; Shaw, A.; Zimmerman, J.; Lease, M.; and Horton, J. 2013. The future of crowd work. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, CSCW '13, 1301–1318. New York, NY, USA: ACM.

Lerman, K., and Hogg, T. 2014. Leveraging position bias to improve peer recommendation. *PLoS ONE* 9(6):e98914.

Loch, C. H.; Huberman, B. A.; and Stout, S. 2000. Status competition and performance in work groups. *J. of Economic Behavior and Organization* 43:35–55.

Mason, W., and Suri, S. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods* 44:1–23.

Miller, B.; Hemmer, P.; Steyvers, M.; and Lee, M. D. 2009. The wisdom of crowds in ordering problems. In *Proceedings of the Ninth International Conference on Cognitive Modeling*.

Nov, O.; Naaman, M.; and Ye, C. 2009. Motivational, structural and tenure factors that impact online community photo sharing. In *Proc. of the Third International Conference on Weblogs and Social Media (ICWSM2009)*, 138–145. AAAI.

Page, S. E. 2010. *Diversity and Complexity*. Princeton University Press.

Pal, A.; Chang, S.; and Konstan, J. A. 2012. Evolution of experts in question answering communities. In *ICWSM*.

Payne, S. L. 1951. *The Art of Asking Questions*. Princeton University Press.

Salganik, M. J., and Watts, D. J. 2009. Web-Based experiments for the study of collective social dynamics in cultural markets. *Topics in Cognitive Science* 1(3):439–468.

Stoddard, G. 2015. Popularity dynamics and intrinsic quality in Reddit and Hacker News. In *Ninth Intl. AAAI Conf. of Web and Social Media (ICWSM 2015)*.

Van Hentenryck, P.; Abeliuk, A.; Berbeglia, F.; Maldonado, F.; and Berbeglia, G. 2016. Aligning popularity and quality in online cultural markets. In *Proceedings of the 10th International AAAI Conference on Web and Social Media (ICWSM-16)*.

Wilkinson, D. M. 2008. Strong regularities in online peer production. In *Proc. 9th ACM conference on Electronic commerce*, 302–309.

Yang, J.; Tao, K.; Bozzon, A.; and Houben, G.-J. 2014. Sparrows and owls: Characterisation of expert behaviour in stackoverflow. In *User Modeling, Adaptation, and Personalization*. Springer. 266–277.

Yi, S. K. M.; Steyvers, M.; and Lee, M. D. 2012. The wisdom of crowds in combinatorial problems. *Cognitive Science* 36(3):452–470.