TrueReview: A Platform for Post-Publication Peer Review

WHITE PAPER

Luca de Alfaro
Computer Science department
University of California, Santa Cruz
luca@ucsc.edu

Marco Faella
Electrical Engineering and Information Technologies
University of Naples “Federico II”, Italy
m.faella@unina.it

Technical report UCSC-SOE-16-13
August 28, 2016

Abstract

In post-publication peer review, scientific contributions are first published in open-access forums, such as arXiv or other digital libraries, and are subsequently reviewed and possibly ranked and/or evaluated. Compared to the classical process of scientific publishing, in which review precedes publication, post-publication peer review leads to faster dissemination of ideas, and publicly-available reviews. The chief concern in post-publication reviewing consists in eliciting high-quality, insightful reviews from participants.

We describe the mathematical foundations and structure of TrueReview, an open-source tool we propose to build in support of post-publication review. In TrueReview, the motivation to review is provided via an incentive system that promotes reviews and evaluations that are both truthful (they turn out to be correct in the long run) and informative (they provide significant new information). TrueReview organizes papers in venues, allowing different scientific communities to set their own submission and review policies. These venues can be manually set-up, or they can correspond to categories in well-known repositories such as arXiv. The review incentives can be used to form a reviewer ranking that can be prominently displayed alongside papers in the various disciplines, thus offering a concrete benefit to reviewers. The
paper evaluations, in turn, reward the authors of the most significant papers, both via an explicit paper ranking, and via increased visibility in search.

1 Introduction

Peer review has not always preceded publication. In the times of Galileo, Newton, van Leeuwenhoek, up to Darwin, scientists would share their results via letters or presentations to scientific societies; the results were then discussed among scientists. The current system of pre-publication peer review was widely adopted only relatively recently, starting in the 1940s with the introduction of large-circulation scientific journals [Spi02]. Pre-publication peer review was shaped by the economics of paper journal publishing: as paper journals are slow and expensive to print and ship, peer-review was used to select which articles deserved wide dissemination.

The economics of publishing is very different now. Information nowadays can be disseminated immediately at very low cost, and furthermore, in a manner that makes it open to social interaction: in blogs, wikis, forums, social networks, and other venues, people can both share and comment on information. Yet, for the most part the scientific community still beholds pre-publication peer review as the officially anointed method of disseminating results. Publication in venues such as journals and conferences with a pre-publication peer-review selection process is also the most commonly used measure of scientific productivity, and contributes to shape the near totality of academic and research careers.

Pre-publication peer review has several drawbacks. One of the most salient is the delay imposed on the dissemination of results. In a typical computer science conference, six months may elapse from submission to publication in the proceedings, and this assumes that the conference deadline came just when the paper was ready for submission, and more importantly, that the paper was accepted. To avoid this delay, many authors submit the paper to open repositories such as arXiv\(^1\) at the same time as they submit it to a conference or journal. While this makes it available to other researchers, a submission to ArXiv does not come with the all-important blessing of peer review. As such, papers submitted to arXiv are not generally counted as part of the productivity record of researchers. Submitting to arXiv is not an alternative for submissions to conferences or journals with double-blind review policies, and citing works submitted in arXiv, but not yet peer-reviewed, is not universally perceived as appropriate in science.

A related issue is the one of selection. The current process of scientific reviewing, consequently, aims at deciding which papers to accept for publication,

\(^1\)http://arxiv.org/
and which to reject. Correctness is only one factor in such a decision: commonly, there are many more correct submissions (in the sense of exempt from scientific errors) than can be accepted, and the decision to accept or reject is motivated by judgements on the significance of the submissions. The paper acceptance process is thus of necessity an uncertain process, where a demarcation line needs to be drawn among papers of fairly similar apparent significance. Papers that present correct results, but which do not make the cut, are subjected to a delay as they are re-submitted to different journals of conferences. The process is slow and wasteful of resources.

One last drawback of pre-publication peer review is that papers are not presented to readers in the context of the accumulated knowledge and judgement. While this shields papers from being presented alongside potentially irrelevant reviews, this also means that insightful observations from readers and researchers cannot help understand papers and put them in context.

This white paper presents the design principles and mathematical foundations of TrueReview, an open-source system we propose to build in support of post-publication peer-review. In the next section, we describe the overall motivations and design principles that inspire the development of TrueReview. After a review of related work, we discuss the problem of motivating reviewers, and we describe in Section 4 the reviewer incentive system at the heart of TrueReview. The main challenge in a review system consists in ensuring that all papers receive sufficient and precise evaluations. Our novel incentive scheme promotes reviews that are both truthful and informative, in the sense that they bring novel information into the system rather than merely confirming what is already known. To validate the proposed incentive scheme, we report in Section 5 the result of simulations of the review process with participants having a varying distribution of skills and paper topic expertise. The simulations show that the incentive system is effective in ensuring that all papers receive precise evaluations. We conclude with an overview of the software architecture of TrueReview, and a discussion of some key implementation decisions.

All the code for TrueReview is open source, and it can be found at https://github.com/TrueReview/TrueReview.

2 TrueReview Design Principles

TrueReview will be an open, on-line system, where authors can publish their papers or enter links to their papers, and where reviewers can review and evaluate the papers after they have been published. This would serve the scientific community as a whole, by making the dissemination of results more open, predictable and less
subject to delays, and by helping researchers view papers in light of the accumulated knowledge and wisdom. At the core of TrueReview are the following design principles.

**Driven by scientific communities.** TrueReview does not plan to be a one-size-fits-all solution for post-publication review. Papers published or linked in TrueReview will not be all put into the same “pot” for review and ranking. Each scientific community has norms for the format of published papers, and has well-known researchers that act as standard-bearers for the community: these are the people that today serve on the journal editorial boards and conference program committees. Each community that elects to use TrueReview will decide whether papers are to be submitted to TrueReview directly, or whether TrueReview tracks paper submitted to certain categories in open-access repositories such as Arxiv. The choice of who can review papers in a venue will also be left to each community. In some venues, the senior members might wish to approve who has review privilege, or adopt an invitation system. For other venues, such as venues that correspond to Arxiv categories, it might be sufficient to have published a paper in the same venue to be able to review.

**No delays to publication.** While papers that have just been submitted are unreviewed, this should not prevent their circulation. One natural objection is whether making papers available immediately deprives readers from the quality guarantee conferred by a formal process of paper review. We believe that the benefits of the prompt communication of scientific results far outweigh the drawback of circulating papers in various stages of review. The status of a peer reviewed paper is often assumed by people not familiar with the process to be a seal of approval that guarantees the correctness of the results. In reality, errors in scientific papers are not always discovered by the conference or journal review committees to which the papers are submitted: more often, the errors are discovered by the authors themselves, or by people who try to use or extend the papers results. Only papers that are widely read, and whose results are extensively used, can be trusted to be highly likely to be correct.

**Rank rather than select.** When a paper is submitted to a journal or conference, the question of whether to accept or reject it most often revolves on the relevance of the paper, rather than on its correctness. After the papers that are clearly flawed are eliminated, there are invariably too many papers to fit in the conference or journal format; the committee must then select the papers to accept on the basis of their quality. The committee thus essentially performs a ranking task, applying then
a binary threshold dictated by conference or journal constraints. For the rejected papers that were indeed correct, this process results in a pointless delay to publication; as these are typically resubmitted to other venues, the work that went into ranking them is also wasted.

This summary of the current review process is greatly simplified. In truth, there are many conferences and journals, with different typical quality levels, and authors choose the venue where to submit the paper in order to compromise between the prestige of the venue, and the probability that the paper is accepted. Nevertheless, the process is wasteful of time and work. We believe it would be better to use the reviews and comments for ranking, rather than for selection. There would be no need to artificially set a cut-off line; all papers would be ranked and available on-line as soon as the authors publish them.

Once a ranking of the papers were available (even if approximate), journals and conferences could use the ranking for selection purposes. For example, a conference could gather people interested in a particular field, and allocate paper presentation slots to the 30 highest-ranked papers of the year, and poster presentation space to the next 50 highest-ranked; a journal or book editor could similarly publish (and distribute to libraries in archival form) the best 50 papers of each year. Certainly many users of the system could use the ranking for selection purposes, but the main goal of the system would be to generate a ranking, not a selection.

**Truthful and informative incentives to review.** The main obstacle to post-publication review consists in enlisting expert reviewers, and having them provide accurate ratings and reviews on most papers in a venue. In conference and journals, the enticement to review is provided by the prestige of appearing on the program committee or editorial board of a well known journal or conference. Reviewers need to provide competent reviews, or risk appearing uninformed when their reviews are compared with those of others on the program committee or editorial board. In exchange for being listed as members of the program committee or editorial board, the reviewers also accept to read and review papers that they would not have read out of their interest alone.

We plan to recreate the incentive to review by also giving wide publicity to the most active reviewers, and by attributing merit for the reviews via an incentive system that prizes reviews that both are correct, and provide new information. In each venue, the names of the reviewers that accrued highest review merit will be displayed in the first page, alongside the top-rated papers. Reviewers will be able to link their name to a web page of their choice, such as their academic home page. We hope this visibility and mark of distinction, which mirrors the one currently offered by membership in program committees and editorial boards, will provide
sufficient motivation to actively participate in the system.

The incentive system for reviews will reward reviewers who provide rating that are both truthful and informative. Truthful ratings are those that will be confirmed later on by the consensus opinion on a paper. Informative ratings are those that provide genuinely new information. Examples of informative ratings are the first rating for a previously unreviewed paper, or a rating that differs from the current consensus for a paper, but will be later confirmed to be correct. In contrast, a rating and review that reflects the consensus on a paper that has already been reviewed many times will have low informative value. Considering both truthfulness and informativeness to reward reviewers encourages them to focus where their expertise allows them to give new useful information. This combined incentive should also lead to prompt rating of papers published in venues.

Additionally, for venues with a long life-span (such as venues replicating arXiv categories), users can be encouraged to periodically contribute new reviews by slowly decreasing their accrued score, as long as they do not provide a new review.

3 Related Work

A detailed proposal for a post-publication peer-review model was made by Kriegeskorte [Kri12]. The author advocates signed reviews and multi-dimensional paper evaluations, that can be aggregated by different interested parties in different ways. The dissemination of signed reviews is deemed a sufficient incentive for reviewers to participate in the process. The paper contains also an in-depth analysis of the benefits of post-publication peer review, which are presented in an eloquent way and which are indeed part of the motivation for this study on incentives. The incentive schemes we propose do not require review authors to be publicly visible. This may be beneficial, as there is some evidence that signed reviews may deter prospective reviewers, or dampen the frankness of their opinions [VRGE+99, VRDE10]. The virtues, and drawbacks, of signed reviews have been described in [Gro10, Kha10]; signed reviews can prevent the abuse of review power, but they also can stifle criticism.

The proposal for TrueReview shares its fundamental motivations with [DSdA11], of which it represents an evolution, as well as with [Kri12], while differing in the details of the incentive system. The incentives we propose do not rule out publishing the names of review authors. We propose instead listing, for each publication venue, both the top papers, and the top reviewers as determined by the total of their review bonuses, allowing users of a Web interface to search both papers, and reviewers. Post-publication peer review has also been advocated, on similar grounds as [DSdA11, Kri12], in [Hun12, Her12, dS13]. Even the popu-
lar press has engaged in the discussion [Mar14], with the CEO of Academia.edu mentioning the possibility of gathering reputation points from reviews.

The idea of evaluating scientific proposals via crowdsourcing reviews and ratings has been proposed as a method for adjudicating telescope time, a central issue in Astronomy [MS09], as well as in the evaluation of some National Science Foundation proposals [NSF13].

ArXiv overlay journals are gaining momentum in several scientific disciplines, including math, physics, and computer science [Gib16]. While their papers are publicly available even before acceptance, their selection process follows the traditional peer review model of printed journals. In the words of Timothy Gowers, Fields medalist and managing editor of the arXiv overlay journal Discrete Analysis, “our journal is very conventional [...] But if the model becomes widespread, then I personally would very much like to see more-radical ideas tried out as well” [Bal15].

Other organizations are indeed pursuing more radical ideas: ScienceOpen publishes articles online under an open-access model and encourages post-publication peer reviews, which include a numerical score. Reviews are publicly attributed to their authors and even assigned a DOI. On the other hand, reviewers do not accrue a numerical reputation for their efforts. Similarly to [Kri12], the incentive for the reviewers consists in having a public collection of their reviews.

O’Peer is a proof-of-concept website where authors-reviewers accrue reputation (called credibility) according to both their publication record and the quality of their reviews.

An incentive system that shares many of the design goals with the one we propose for TrueReview has been proposed by Bhattacharjee and Goel [BG07] in their work on incentives for robust ranking in online search. In the [BG07] proposal, users can place tokens on items in order to place wagers on the quality of the items, much as people can bet on horses at races. If the ratio between the qualities of two items is different from the ratio between the token amounts, an arbitration opportunity arises, and a user can move a token from the over-rated item to the under-rated one and gain reputation (an operation that is roughly equivalent to betting a negative dollar on a horse and a positive dollar on another, if negative bets were allowed). The incentive scheme is truthful, as the incentive is to bring token counts in direct proportionality with qualities, and it also promotes informativeness, as the biggest arbitration opportunities occur for the papers that are most under-valued. We made various attempts at adapting [BG07] for post-publication

---

2http://www.academia.edu
3http://www.scienceopen.com
4http://opeer.org
review, before finally opting for the grade-based scheme we propose for TrueReview. The main problem we encountered is the slow start in properly ranking new papers. When a paper is added, initially it has no tokens. If users can place or move one token at a time, a good paper will require many reviews to receive a proper ranking; if users can move many tokens at once, the vandalism of a single user can cause considerable damage. Another issue was that the truthfulness and innovativeness incentives are tied together by the arbitration opportunity, and their strengths cannot be independently tuned. Ultimately, we felt that the approach proposed in this paper was more flexible and allowed us to better control vandalism. We can independently tune the truthfulness and informativeness incentives, and we can adopt a number of aggregation strategies for reviews.

The idea of validating assertions by considering them wagers on future value, and rewarding thus their accuracy, is the principle at the basis of prediction markets \cite{WZ04,TT12}. The arbitration opportunities in prediction markets are in fact conceptually similar to those in \cite{BG07}, except that by having real money involved, the possibility for vandalism is virtually eliminated. Indeed, the stock market offers a model for crowdsourcing valuations that both is truthful, and that offers a prize for informativeness. However, the full working of the market (including the put and call options that are important in betting on future valuations) are vastly more complex than the simple mechanism we presented in this paper, and arguably over-complicated for the task at hand.

There has been much work on peer evaluation, in classroom settings \cite{Geh00,Geh01,Rob01,STOA13} and in MOOCs \cite{PHC}. In a classroom or MOOC setting, however, the focus is on obtaining precise and fair evaluations, rather than on incentives to select the items (papers, or homework submissions) to review. This because in educational settings, students are usually compelled to perform the peer reviews and evaluations as part of their class work. Furthermore, as homework submissions share all the same topic, the review assignment can be (and usually, is) performed automatically, again obviating the need for an informative incentive system.

4 The TrueReview Incentive System for Reviewers

The crucial challenge for post-publication review consists in ensuring that papers receive adequate reviews and precise evaluations. There are many mechanisms for ensuring that the set of potential reviewers is capable of writing useful reviews: they can be invited to review, or the privilege of reviewing can be granted automatically to people who have successfully published previously in the same venues.

The basic user action in TrueReview consists in a user choosing a paper, and
providing both a written review, and a numerical rating for the paper. The ratings are then aggregated in a single rating for the whole paper. TrueReview rewards the author of a review with a review “bonus”. In each publication venue, reviewers will be listed according to the total of the bonuses they received: we hope this visibility will provide incentive to review.

The incentive scheme used for assigning the review bonuses should be truthful: the strategy for users to maximize their bonuses for each review should be to express their honest opinion about the paper. Furthermore, the incentive scheme should be informative: it should prize new relevant information over repetition of already-known information. For instance, it should value the first review on a paper more than a review confirming the consensus opinion on a paper that has already been reviewed many times. Among papers having the same number of reviews, an informative incentive scheme should value reviews that express opinions different from the consensus, and that will turn out to be correct, more than reviews that are simply confirming the current consensus. Informative incentive schemes lead to a quick convergence to the true valuation for all papers.

The incentive system we propose for TrueReview combines two components: one that depends only on the correctness of the expressed valuation, and one that depends only on the error in the current valuation of a paper before the new review is entered. The first component accounts for truthfulness, the second for informativeness. The relative strength of the two components can be tuned. When the truthfulness component is stronger than the informativeness one, reviewers have more of an incentive to express opinions on papers on which they are experts, rather than on papers with insufficient evaluations; the opposite happens when the informative component is stronger.

We present here the results for average as the rating aggregation mechanism; other mechanisms can be used, as we discuss later, provided that they do not introduce a systematic bias in the results. Let $L(a, b) = (a - b)^2$ be the quadratic loss function. When a user provides a rating $x$ for a paper, we award the user a review bonus equal to:

$$- L(x, q^{\text{cur}}) + L(q^{\text{old}}, q^{\text{cur}}),$$

(1)

where $q^{\text{cur}}$ is the most recent quality estimate for the chosen paper, $q^{\text{old}}$ is the estimate just before the review. Notice that the reputation boost due to having graded a paper is not a static quantity, but changes in time according to the most recent quality estimate $q^{\text{cur}}$ for that paper. Hence, re-evaluating a paper and updating its quality estimate affects the reputation of all the users who have reviewed that paper in the past.
4.1 Inference-truthful strategies

This system of incentives promotes honest reviews and an efficient choice of papers to review. The idea is that the first term in (1), which we call *accuracy*, encourages people to “guess” the future quality estimate of that paper. We can prove that, under mild assumptions, the strategy that maximizes accuracy consists in providing a truthful grade. The second term in (1), called *informativeness*, rewards the users who choose to review the papers that need it the most, because their current quality estimate is farther away from the future consensus. A crucial property of the informativeness term is that its value depends on the chosen paper but not on the grade $x$ issued by the user. This ensures that, once the user has chosen a paper to review, the only remaining incentive is accuracy, which encourages rational users to be truthful.

To formalize the above observations, we idealize TrueReview as a game in which an infinite sequence of users connects to the system. Each user chooses a paper to review, forms a numerical belief $b$ on the quality of that paper, and provides a grade $x$ for that paper, possibly different from $b$. We assume that the belief of each user for a paper is unbiased, that is, has expectation equal to the “true quality” of that paper.

Denoting with $u_1, u_2, u_3, \ldots$ the sequence of users that judges a particular paper, we can model the rating process as a Bayesian game [OR94], in which each user can observe the ratings $x_1, x_2, \ldots, x_{i-1}$ given by previous users, as well as their own belief $b_i$ about the paper quality. On the basis of these observations, user $i$ must in turn provide a rating $x_i$ for the paper. We assume that the goal of the user is to maximize their long-term reputation. Denoting with $z = \lim_{n \to \infty} f(x_1, x_2, \ldots, x_n)$ the limit value of the aggregation of the ratings, the goal of user $i$ is thus to minimize $(x_i - z)^2$.

In this setting, it can be shown that the strategy profile in which each player gives as rating its best estimate of the paper quality constitutes a Nash equilibrium. Precisely, consider the strategy profile in which player $i$ uses the observations $x_1, x_2, \ldots, x_{i-1}, b_i$ as well as the prior distribution on paper quality to reconstruct a best Bayesian estimate $\hat{q}$ of the true quality of the paper, and votes $x_i = \hat{q}$. We call this strategy profile the *truthful inference* profile: user $i$ reports not their raw estimate $b_i$ of the paper quality, but rather, their best Bayesian reconstruction in light of all prior information they have, including the previous ratings received by the paper. It can be shown that the truthful inference profile is a Nash equilibrium of the reviewing game.
4.2 Review aggregation and incentives in practice

Aggregation functions. In the Nash equilibrium arising from truthful inference strategies, once users 1, \ldots, n have rated the paper, user \( u_n \) has summarized in the rating \( x_n \) all previous information available about the paper, in addition to his or her own opinion. Thus, if users could be trusted to play according to this strategy profile, we could simply take as aggregation function \( f(x_1, \ldots, x_n) = x_n \). In practice, putting all faith in the last rating is unsafe: the truthful inference strategy is difficult to play, the last user may have unreliable individual opinion, and individual users cannot be unconditionally trusted to act only with the goal of raising their own review reputation. For this reason, we favor in practice aggregation functions that rely on weighed averages. One approach consists in using the arithmetic average: \( f(x_1, \ldots, x_n) = \text{avg}(x_1, \ldots, x_n) \). Another interesting approach consists in using a geometrically-weighed average, where the most recent rating matters most, and prior ratings are then discounted geometrically. Denoting with \( \langle x_1, \ldots, x_n \mid w_1 \ldots, w_n \rangle \) the weighed average in which rating \( x_i \) is weighed \( w_i \), we let \( f(x_1, \ldots, x_{n-1}, x_n) = \langle x_1, \ldots, x_n \mid a^{n-1}, \ldots, a, 1 \rangle \), for some \( 0 < a < 1 \).

Giving more weight to the latest reviews, as called by second approach above, enables the system to react more quickly when errors in papers are discovered. Consider a paper that receives good ratings \( x_1, \ldots, x_n \), after which reviewer \( n + 1 \) discovers a serious error. If we use simple average as aggregator, we need \( n \) additional reviews simply to bring the paper rating mid-way between its original inflated value, and its revised value after the error has been discovered. If we use the geometrically-weighed scheme, the convergence to the new consensus after the error is accounted for can be faster, since the lastest \( k \) ratings account for a share at least \( 1 - a^k \) of the total weight, independently of the value of \( n \).

Another design decision is the initial value of \( q^{\text{cur}} \), i.e., how to judge a paper that has yet to receive any review. We adopt the conservative approach consisting in assigning the lowest possible grade (zero, in our simulations). In this way, evaluations are skewed towards lower grades and several positive reviews are required to bring a paper to the top of the grade scale. This choice follows the idea that underestimating the paper quality is preferable to over-estimating it. Alternatively, one may set the initial value of each paper to the average value of all the papers that have received at least one review. In other words, each new paper is treated as an average paper. This approach would hasten the convergence to the true value of the papers, but it would slow down the accrual of reviews on new papers, as the informativeness bonus would be lower for them.
Incentives. Practical considerations suggest enriching the incentive structure with two parameters $c > 0$ and $\alpha \in (0, 1)$, obtaining the following:

$$c - (1 - \alpha)L(x, q^{\text{cur}}) + \alpha L(q^{\text{old}}, q^{\text{cur}}).$$

It is easy to see that the truthfulness guarantees discussed in Section 4.1 remain valid if incentives (1) are replaced by (2).

The additive constant $c$ makes sure that even reviews with a low informativeness content still result in a positive reputation boost. Another way to put it is that $c$ compensates users for having examined the paper. Hence, it is intended to motivate reviewers to participate in the scheme. The coefficient $\alpha$ allows to tune the relative importance of accuracy and informativeness.

In the next section, we experimentally evaluate the effect of different values of $\alpha$ on a simulated population of users.

5 Simulations

To validate the proposed incentive scheme, we report the result of simulations of the review process with participants having a varying distribution of skills and paper topic expertise. We also examine a range of user behavior, from fully rational and motivated to increase the total bonuses received, to approximate models that may better account for actual reviewer behavior. We examine the full-rational case in which reviewers take into account both their opinion, and all previous reviews using Bayesian inference, computing then the expected bonus they may expect by reviewing each paper. Moreover, as a fully rational behavior capable of statistical inference cannot be taken as granted, we also simulate a simpler type of reasoning, which is aware of the current evaluation of each paper, but is insensitive to the individual ratings and to the number of ratings collected by a paper.

In both cases, our simulations confirm that an incentive scheme that combines truthfulness and informativeness, once properly tuned, leads to much faster convergence to an accurate evaluation of all papers, than schemes based on truthfulness alone. As an example, for the simpler reasoning model and after an average of 3 reviews per paper, the scheme based on truthfulness alone retains 35% of the initial evaluation error whereas the tuned scheme only retains 7% of the initial error. Our simulations also show how incentives based on truthfulness alone can give rise to a perverse feedback loop in which a few papers end up with a very large number of evaluations, while other remain un-evaluated. This happens because, if accuracy is the only reward criterion, once some papers receive more reviews due to statistical fluctuations, reviewers flock to them, as they can attain precision simply by voting the average grade of all previous reviews.
Moreover, the scheme we suggest exhibits positive correlation between the competence of each reviewer and the reputation she will accrue after a few contributed reviews. As a further positive side-effect, we show that users are motivated to review papers on which they are more competent. Our results also indicate how to tune the relative strength of the truthfulness and novelty incentives to obtain schemes that perform well under a variety of user behaviors.

5.1 Simulation Setup

We simulate the behavior of a population of 1000 users that evaluate a collection of 1000 papers. We assume that each paper has an intrinsic quality $q^{\text{true}}$ which represents our ground truth. At any given time, the system attributes a current rating $q^{\text{cur}}$ to each paper. Such rating starts at zero and is updated as the arithmetic average of the grades provided by the reviewers (including the initial default value of zero).

The reputation resulting from a review is defined by (2). The core component of the simulation is its user model, dictating how simulated users choose a paper to review and a grade for it. In particular, simulated users hold certain beliefs about the papers, which allow them to estimate the expected reputation boost deriving from reviewing a certain paper. Supported by the observations in the previous sections, we assume that users grade papers truthfully, i.e., according to the best reconstruction allowed by the model.

5.2 User Models

We stipulate that each user is interested in a random sample of 100 papers. On each of those papers, the user initially holds the following beliefs: the paper quality $z$ and the expected error $\sigma$ in judging the paper quality. One can think of $\frac{1}{\sigma}$ as the competence of the user on that paper. Moreover, users are aware of the average error $\bar{\sigma}$ among all users and all papers.

Next, we describe how the above parameters are sampled. The true value $q^{\text{true}}$ is sampled for each paper out of a normal distribution. Then, each user is attributed a typical error $\sigma^t$ out of a distribution with mean $\bar{\sigma}$; the typical error indicates the “overall competence” of the user for the papers under consideration. The paper-specific error $\sigma$ is sampled from a distribution with mean $\sigma^t$. Finally, the perceived paper quality $z$ is sampled out of a normal distribution with mean $q^{\text{true}}$ and standard deviation $\sigma$ (denoted by $\mathcal{N}(q^{\text{true}}, \sigma)$). Thus, we model users of varying degrees of average competence, and with each a set of papers that they might consider reviewing.
We present results for two user models. The models coincide on the original beliefs held by the users about the papers, but differ in the way users take into account previous reviews received by a paper. In the first user model, users have an initial belief in the quality of a paper, but then revise the belief to take into account the grades in the previous reviews and their supposed accuracy. In the second user model, users have an initial belief in the quality of a paper, and they follow their belief, without revising it in view of previous reviews.

**First user model.** In our first user model, users look at previous reviews to reconstruct via Bayesian inference the most likely grade for a paper. Based on these beliefs and taking into account previous reviews, users estimate the reputation boost they may receive from evaluating a given paper. Consider a paper with \( n \) previous reviews and current evaluation \( q_{\text{cur}} \). Since a user does not hold a specific belief on the competence of previous reviewers, he assumes they all share the same error \( \bar{\sigma} \).

Then, in the first user model, the best quality estimate \( \hat{q} \) and the corresponding error \( \hat{\sigma} \) are obtained by Bayesian inference with prior \( \mathcal{N}(z, \sigma) \) and observation \( q_{\text{cur}} \) with likelihood \( \mathcal{N}(q_{\text{true}}, \frac{\sigma}{\sqrt{n}}) \). The likelihood follows from assuming previous reviews independent of one another. Accordingly, the accuracy term of the incentive is estimated as \( \hat{\sigma}^2 \) and informativeness as \( (q_{\text{cur}} - \hat{q})^2 \), leading to the reputation boost estimate

\[
c - (1 - \alpha)\hat{\sigma}^2 + \alpha(q_{\text{cur}} - \hat{q})^2.
\]

**Second user model.** In the second user model, reviewers hold the same beliefs as in the first model, but they apply a simplified reasoning when estimating the reputation boosts. In particular, they do not perform Bayesian inference and do not take into account the number of previous grades that a paper has attracted, but only the overall quality estimate \( q_{\text{cur}} \) resulting from those grades. So, accuracy is simply estimated as \( \sigma^2 \) and informativeness as \( (q_{\text{cur}} - z)^2 \), leading to the reputation boost estimate

\[
c - (1 - \alpha)\sigma^2 + \alpha(q_{\text{cur}} - z)^2.
\]

### 5.3 Choice of paper to review

At each round, a user is selected in round-robin fashion and performs a truthful review of a paper. Since we are not simulating the fact that users voluntarily participate in the system, we can ignore the \( c \) parameter and set it to zero. We compare three different scenarios in which users choose which paper to review in the following ways:

- **Random:** uniformly at random among the papers known to the user.
5.4 Performance Measures

Our first performance measure is the global loss of the current quality estimates, computed as the sum over all papers of the squared difference between the paper quality estimate $q_{\text{cur}}$ and the paper intrinsic quality $q_{\text{true}}$.

To illustrate how more expert reviewers receive more reputation (total review bonus points), we also report the Pearson and Spearman correlations between user competence (the user-typical error $\sigma^t$ discussed above) and the reputation at the end of the experiment. Additionally, the fourth column reports the expected error incurred by a user when grading a paper, relative to the typical error of that user. A value close to 1, such as the one obtained by the random choice criterion, implies that users select papers independently of their specific competences. On the contrary, the lower the value the more users are choosing the papers they are more familiar with. Since users in practice are likely to prefer those papers anyway, we see a low value in that column as a desideratum for our incentive scheme. The relative error values are averaged over rounds and data sets, and the last column in our tables displays the standard deviation over data sets.

5.5 Results

We simulated the behavior of 1000 users evaluating a set of 1000 papers. Each user holds beliefs on a random subset of 100 papers that the user is willing to review. We simulate 5000 reviewing rounds. We repeated each simulation run 10 times, in order to measure the standard deviation of the results across the runs.

First user model. Figure [1] shows the value of the global loss, relative to the initial global loss, when paper choice is performed according to the three criteria random, selfish, and optimal. Figure [1] shows that the selfish choice for $\alpha = 0$
performs even worse than the random choice because, when accuracy is the only incentive, users tend to focus on a few paper. This because, the more evaluations a paper has received, the more accurately users can predict its quality simply on the basis of the previous grades received by the paper. This creates a perverse incentive, in which the papers whose quality is best known draw the most evaluations. Indeed, Figure 3 confirms that in that case more than 50 papers receive an exceeding amount of grades, whereas 650 papers are completely neglected. The distribution of the number of grades per paper becomes much more balanced starting from $\alpha = 0.5$, when the informativeness criterion starts to play a role.

The selfish choice performs better the closer $\alpha$ gets to 1, both from the point of view of the global loss (Figure 1) and as far as the correlation is concerned (Figure 5). On the other hand, selecting $\alpha = 1$ would invalidate the truthfulness guarantees. Moreover, any value too close to 1 may encourage people to discount the importance of the accuracy and provide a superficial evaluation of the paper. Hence, in practice we suggest to select a value of $\alpha$ between 0.5 and 0.8. Notice that after 3000 rounds — i.e., 3 reviews per paper on average — the selfish criterion with $\alpha \in \{0.5, 0.8\}$ achieves a remarkable 7% relative global loss, compared to 31% when users choose papers at random and 4% when they choose papers optimally (see the second column in Figure 5).
Figure 2: Relative loss in the second user model.

Figure 3: Distribution of the number of grades per paper in the first user model. The labels 11+ and 21+ stand for the intervals $[11, 20]$, $[21, \infty)$, respectively.
Figure 4: Distribution of the number of grades per paper in the second user model. The labels 11+ and 21+ stand for the intervals $[11, 20]$, $[21, \infty)$, respectively.

<table>
<thead>
<tr>
<th>choice criterion</th>
<th>loss</th>
<th>Pearson</th>
<th>Spearman</th>
<th>rel. error</th>
<th>rel. error std. dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>0.31</td>
<td>-0.061</td>
<td>-0.089</td>
<td>0.999</td>
<td>0.004</td>
</tr>
<tr>
<td>optimal</td>
<td>0.04</td>
<td>-0.026</td>
<td>-0.035</td>
<td>0.806</td>
<td>0.005</td>
</tr>
<tr>
<td>selfish, $\alpha = 0$</td>
<td>0.76</td>
<td>-0.386</td>
<td>-0.369</td>
<td>0.935</td>
<td>0.004</td>
</tr>
<tr>
<td>selfish, $\alpha = 0.2$</td>
<td>0.09</td>
<td>-0.019</td>
<td>-0.042</td>
<td>0.821</td>
<td>0.005</td>
</tr>
<tr>
<td>selfish, $\alpha = 0.5$</td>
<td>0.07</td>
<td>0.110</td>
<td>0.150</td>
<td>0.746</td>
<td>0.003</td>
</tr>
<tr>
<td>selfish, $\alpha = 0.8$</td>
<td>0.07</td>
<td>0.139</td>
<td>0.250</td>
<td>0.767</td>
<td>0.002</td>
</tr>
<tr>
<td>selfish, $\alpha = 1$</td>
<td>0.07</td>
<td>0.146</td>
<td>0.282</td>
<td>0.787</td>
<td>0.002</td>
</tr>
</tbody>
</table>

Figure 5: Summary data for the first user model. The columns contain: the relative global loss after 3000 reviews; the Pearson and Spearman correlations between competence and reputation after 5000 reviews; the average relative error (defined in Section “Performance Measures”) and its standard deviation across 10 data sets.
Figure 6: Summary data for the second user model. The columns have the same interpretation as Figure 5.

We report in Figure 5 the Pearson and Spearman correlations between user competence and user reputation, and the user preference for reviewing papers on which they are most proficient. As we see, there is positive if weak correlation between user competence and final user reputation. The correlation is weak: to accrue reputation, it is often better to be “lucky” and find a paper that has received incorrect evaluations so far, rather than using one’s expertise to provide a precise evaluation of a paper with already many evaluations.

Second user model. Figure 2 shows that the relative global loss in this model is not very different from the one of the previous, more elaborate one. The most significant difference occurs with $\alpha = 0$, where the performance is not as bad as in the first model. Indeed, when users are insensitive to the number of previous grades, they have no reason to amass grades on a few papers. However, they still have a bias towards papers that have received at least one previous review. Figure 4 clarifies the situation, by showing that the distribution of the number of grades after 5000 rounds is more balanced than in the first model, but still features a significant amount of papers with zero reviews.

The value $\alpha = 0.5$ still shows remarkably good behavior in both global loss and number of grades per paper. Correlation between user competence and reputation is similar to the first model. On the other hand, the relative error (penultimate column in Figure 5) in this case is minimized for $\alpha = 0$, but retains a good value for $\alpha = 0.5$ (0.72), slightly better than in the previous model (0.75). Thus, the second user model confirms that 0.5 is a reasonable default choice for $\alpha$. 
References


[NSF13] NSF. Dear colleague letter: Information to Principal Investigators (PIs) planning to submit proposals to the sensors and sensing systems (sss) program October 1, 2013 deadline, 2013.


