TrueReview:
A Proposal 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
November 4, 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 priviledge, 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. Post-publication peer review has also been advocated, on similar grounds as [DSdA11, Kri12], in [Hun12, Her12, dS13]. Even the popular press has engaged in the discussion [Mar14], with the CEO of Academia.edu[2] mentioning the possibility of gathering reputation points from re-

[2] `http://www.academia.edu`

views. Experiments with post-publication peer review have been conducted in [Bha14], which obtained useful data on the extent on which later reviews are influenced by earlier ones.

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[3] 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[4] 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

[3] `http://www.scienceopen.com`

[4] `http://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* [WZ04, TT12]. The arbitration opportunities in prediction markets are in fact conceptually similar to those in [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 [Geh00, Geh01, Rob01, STOA13] and in MOOCs [PHC$^+$13]. 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.

4.1 Informativeness and Accuracy of a Review

We introduce an incentive system for reviewers that is both truthful and informative. Consider the sequence of ratings $x_0, x_1, x_2, \ldots, x_n$ that have been assigned, in chronological order, to a given paper, where x_0 is the default rating that is assigned by the system to every paper who is added to the system, as a starting point. To define the bonus B_i for the author of rating x_i, for $0 < i < n$, let

$$q_i^{past} = \text{avg}\{x_0, x_1, \ldots, x_{i-1}\}$$
$$q_i^{future} = \text{avg}\{x_{i+1}, x_{i+2}, \ldots, x_n\}$$

be the averages of the ratings preceding and following x_i, respectively. Let L be the quadratic loss function, defined by $L(a, b) = (a - b)^2$. We define the *accuracy loss* and *informativeness* of the rating x_i as follows:

$$\text{Informativeness:} \qquad \delta_i = L(q_i^{past}, q_i^{future}) \tag{1}$$

$$\text{Accuracy loss:} \qquad \theta_i = L(x_i, q_i^{future}) \tag{2}$$

To reward reviewers which are *both* informative and accurate, we let the review bonus b_i be:

$$b_i = \delta_i \cdot f_{\alpha,M}^S(\theta_i) \tag{3}$$

where $f_{\alpha,M}^S$ is a sigmoidal function parameterized by a parameter $\alpha > 0$, and by the maximum rating M that can be given to a paper (see Figure 1). The sigmoidal

function is such that $f^S_{\alpha,M}(0) = 1$, and $f^S_{\alpha,M}(M^2) = 0$, so that perfectly accurate reviewers will get their full bonus, and reviewers with the maximum possible value M^2 of accuracy loss will not get any bonus.

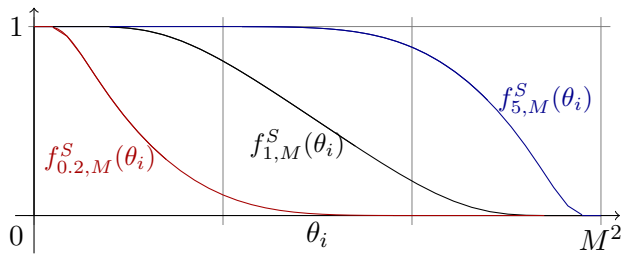

Figure 1: Three instances of $f^S_{\alpha,M}(x)$, for $\alpha = 0.2$, $\alpha = 1$ and $\alpha = 5$.

The informativeness, accuracy loss, and bonus defines by (1)–(3) have a number of important properties.

Informativeness provide an incentive to select papers whose current evaluation is most different from what the future consensus will be. Crucially, the informativeness depends on the *previous* and *future* ratings of the paper, but not on the rating x_i assigned by the reviewer under consideration. Thus, once the reviewer selects a paper to review, informativeness plays no further role, and the bonus depends entirely on the accuracy loss. This decouples choice of paper from accuracy, and will be crucial in proving that the incentive scheme is truthful.

The accuracy loss is computed by comparing the user's rating only with *future* valuations. This eliminates any incentive to provide a valuation that is similar to the known past ones, against the user's true belief about the paper. Consider, for instance, an alternative definition where q^{future} represents the average of all valuations. If the user has some prior knowledge about the typical number of reviews a paper is likely to receive, and most of those have already been written, then the user would be able to achieve high accuracy just by providing a rating that is the average of the past ones.

A consequence of our choice of incentive scheme is that we reward users who discover new information, and present it convincingly in their reviews. For such users, the score they propose is different from past ones, and influences all future ratings. The informativeness will be high, due to the difference between past and future ratings, and the accuracy will not suffer from the difference between the score and previous ratings. Notice how this latter property would not hold if we included all ratings in the computation of q^{future}.

4.2 Truthfulness

Our incentive schemes are not truthful in the strong sense that it is a weakly dominating strategy for players to give ratings that reflect their true opinion of the paper. There are many collusion schemes that form Nash equilibria where deviating reduces the bonus: for instance, the one where all reviewers provide the same, constant, rating. Since there is no ground truth for paper quality independent from reviewer-provided ratings, the inability to ensure that truthful strategies are weakly dominating is unavoidable: reviewers could agree to review a paper as if it were another one; there is nothing intrinsic that ties the reviewer behavior to the paper being reviewed that can be used in the mechanism design. The best we can show is that under some conditions, being truthful is a strict Nash equilibrium, that is, a Nash equilibrium from which deviating is not favorable.

The rating process can be modeled as a Bayesian game [OR94], in which each user i can observe the ratings $x_1, x_2, \ldots, x_{i-1}$ given by previous users to a given paper, as well as their own belief y_i about the paper quality q^{true}. On the basis of these observations, user i must in turn provide a rating x_i for the paper. In formulating our results, we assume that the private estimate y_i of the quality of the paper available to player i is unbiased, and that estimates of different reviewers are uncorrelated. The assumption that the user estimates are not overall biased is unavoidable: there is no way to distinguish between a paper of quality q^{true}, and a paper of quality $q^{true} - \Delta$ for some $\Delta > 0$, which all users over-appreciate on average by an amount Δ. Put another way, there is no way to differentiate good papers from papers that only seem good to all users: the two notions coincide in our system.

In the following theorems, we use the fact that the bonus received by the i-th user is a combination $b_i = F(\theta_i, \delta_i)$, where F is monotonically decreasing in θ_i and monotonically increasing in δ_i. As δ_i is fully determined by the choice of paper, to reason about truthflness, we can reason on the θ_i component only.

Our first result concerns reviewers who rate papers without access to other reviews. In this case, it is easy to show that being truthful is a strict Nash equilibrium.

Theorem 1. *Assume all users form statistically uncorrelated and unbiased estimates of the quality of each paper, and assume that users enter their review without being able to read other reviews first. Then, the strategy profile under which all users rate the paper with their quality estimates is a strict Nash equilibrium.*

Proof. Let $0 < i < n$, and assume $x_j = y_j$ for $0 < j \leq n$, $j \neq i$, so that all users except the i-th rate papers with their individual estimate, and consider the point of view of the i-th user. The user must minimize θ_i. As the private estimates $\{y_j\}_{0<j\leq n}$ are uncorrelated, the expected value of θ_i can be written as the sum of

two variances $v_i + v_f$, where v_i is the variance of x_i with respect to the true value q^{true} of the paper, and v_f is the variance of q_i^{future} with respect to q^{true}. As user i has no influence over v_f, the user must minimize v_i, and this entails voting the best estimate y_i of q^{true} available to the player, so that $x_i = y_i$. □

We can extend this result to the case in which reviewers can read previous reviews, and adjust their submitted ratings according to the previous ratings for the paper. Consider again users $1, 2, 3, \ldots$, with private uncorrelated estimates $y_1, y_2, y_3, \ldots$, whose expected value is the quality q^{true} of the paper. We assume for simplicity that these private estimates all have the same variance v (the general case is similar, and can be obtained by weighing each estimate with the inverse of its variance). In a truthful strategy profile, each user will report the most precise estimates that can be computed from the private information and from the previous ratings. Thus, user 1 will report $x_1 = y_1$, user 2 will report $x_2 = (x_1 + y_2)/2 = (y_1 + y_2)/2$, and in general, user n will report $(n-1)x_{n-1}/n + y_n/n = (y_1 + y_2 + \cdots + y_n)/n$. The next theorem shows that deviating from this truthful strategy yields a lower bonus, so that the truthful strategy profile is a Nash equilibrium.

Theorem 2. *If reviewers have access to previous reviews, and if their private estimates are uncorrelated, being truthful is a strict Nash equilibrium.*

Proof. Consider users $1, 2, \ldots, n, n+1$, with uncorrelated private estimates y_i. We show that it is optimal for user n to be truthful; the general case for $1 \leq i \leq n$ is similar but leads to more complicated notation. If user $n+1$ plays truthfully, while player n deviates from the truthful amount by Δ, we have:

$$x_n = \frac{n-1}{n} x_{n-1} + \frac{y_n}{n} + \Delta \qquad x_{n+1} = \frac{n}{n+1} x_n + \frac{y_{n+1}}{n+1} \, .$$

The expected loss $\mathrm{E}\left[(x_n - x_{n+1})^2\right]$ is thus equal to

$$\mathrm{E}\left[\Big(\frac{n-1}{n(n+1)} x_{n-1} + \frac{y_n}{n(n+1)} + \frac{\Delta}{n+1} - \frac{y_{n+1}}{n+1}\Big)^2\right] \, . \tag{4}$$

Noting that $\mathrm{E}\left[x_{n-1}\right] = \mathrm{E}\left[y_n\right] = \mathrm{E}\left[y_{n+1}\right] = q^{true}$, we have that the coefficient of Δ in the expansion of (4) is

$$\frac{q^{true}}{(n+1)^2}\Big[\frac{n-1}{n} + \frac{1}{n} - 1\Big] = 0 \, .$$

Thus, (4) depends on Δ only via $\Delta^2/(n+1)^2$. Since x_{n-1}, y_n, and y_{n-1} are mutually uncorrelated, we obtain from (4):

$$\mathrm{E}\left[(x_n - x_{n+1})^2\right] = \frac{v}{n(n+1)} + \frac{\Delta^2}{(n+1)^2} \, ,$$

where v is the variance of one of the individual estimates y_i. Thus, user n incurs minimum loss when $\Delta = 0$, showing that deviating from truthful behavior reduces the review bonus. Intuitively, any variation from the truth by one user only partially influences later users, raising the loss of the deviating user. □

4.3 Discussion

Truthfulness. In the results developed in this section, it is assumed that users have a private estimate of the quality of the paper, which can contain errors but is unbiased in expectation, and we show that it is a strict Nash equilibrium for users to vote such private estimate (possibly refined by looking at previous grades). Thus, truthfulness corresponds to voting what is the reviewer's belief about the *future consensus* valuation of the paper, and in particular, *other reviewers' future consensus* on the paper. Consider the situation of a reviewer who, after reading a paper, thinks that her own appreciation of the paper is y'_i, but who thinks that other reviewers will think of it y_i, with $y_i = y'_i + \Delta$, $\Delta \neq 0$. Then, the truthful strategy calls for voting y_i, and not y'_i. The estimate y'_i is best understood as the consensus estimate y_i plus a personal bias Δ. From the point of view of the incentives, it is of no consequence whether the personal bias Δ is due to high-minded reasons (such as a particular appreciation for the paper topic that the reviewer thinks others are unlikely to share), or to maliciousness (such as an attempt to cause a paper by friends to be ranker higher). The incentive scheme asks of reviewers, in both cases, to enter their most accurate estimate of the future consensus on the paper.

Grade aggregation. The rules (1)–(3) affect *bonus* computation, but they do not constrain *grade aggregation:* when presenting the papers to user who visit the publication venues, we are free to choose the mechanism used to aggregate the grades provided by the users into a single grade for the paper. One obvious aggregation mechanism consists simply in computing the average of the grades proposed by the reviewers, optionally excluding the first default grade. If reviewers refine their grades by looking at previous reviews, however, this may be suboptimal. Indeed, in the Bayesian scheme discussed in connection with Theorem 2, the private estimate y_i of user i would carry weight $\frac{1}{n} \cdot \sum_{k=i}^{n} \frac{1}{k}$ in an average of grades $\text{avg}\{x_1, \ldots, x_n\}$. This makes (unweighted) average a sub-optimal aggregator: the optimal aggregator would weigh each y_i in reverse proportion to its variance. In particular, averaging the grades over-weighs the private estimates of the first reviewers, which act as influencers on those who follow.

If reviewers look at the work of others before entering their reviews, as confirmed in experimental studies such as [Bha14], it may be better to use an average

that weighs more heavily the more recent reviews. Such a "recent-heavy" mechanism would also be better to cope with papers whose evaluation changes in time, for instance, due to the discovery of errors, or the re-evaluation of results.

5 Simulations

We have shown in the previous section that the accuracy part of the incentive ensures that, once a reviewer has chosen a paper to evaluate, it is in her best interest to evaluate it honestly. It remains to show that the informativeness term of the incentive encourages users to choose papers in a way that benefits the overall quality of the ranking. We provide evidence in this direction through a set of simulations in which a population of 1000 users evaluates a collection of 1000 papers.

We assume that each paper has an intrinsic quality q^{true} which represents our ground truth. At any given time, the system attributes a current rating 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 the bonus (3). 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.

5.1 User Models

We stipulate that each user is interested in a random sample of 100 papers out of the total 1000. On each of those papers, the user initially holds the following beliefs: the paper quality z and the corresponding expected error σ. One can think of $\frac{1}{\sigma}$ as the *competence* of the user on that paper, of which the user is self-aware. 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^{true} is sampled for each paper out of a normal distribution. Then, each user is attributed a typical error σ^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 σ is sampled from a distribution with mean σ^t. Finally, the perceived paper quality z is sampled out of a normal distribution with mean q^{true} and standard deviation σ (denoted by $\mathcal{N}(q^{true}, \sigma)$). 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 grade according to their belief, without taking into account previous reviews. In the second user model, users revise their belief to take into account the grades in the previous reviews and their supposed accuracy. In both models, each user starts with the belief (z, σ) described above, for each paper in which she is interested.

First user model. In the first user model, reviewers believe their quality estimate z to be the best estimate for the future consensus grade assigned by the system to that paper. Accordingly, accuracy is simply estimated as σ^2 and informativeness as $(q^{past} - z)^2$, where q^{past} is the average of the previous grades, leading to the reputation boost estimate

$$(q^{past} - z)^2 \cdot f^S_{\alpha,M}(\sigma^2).$$

The above estimate is used to choose which paper to review. Once a given paper is chosen to be reviewed, it will receive grade z. This user model is consistent with the assumptions of Theorem 1.

Second user model. In our second 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^{past}. Since a user does not hold a specific belief on the competence of previous reviewers, she assumes they all share the same error $\bar{\sigma}$. Then, in this 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^{past} with likelihood $\mathcal{N}(q^{true}, \frac{\bar{\sigma}}{\sqrt{n}})$. The likelihood follows from assuming that previous reviewers adopted a similar Bayesian inference procedure, starting from statistically independent private beliefs. Accordingly, the accuracy term of the incentive is estimated as $\hat{\sigma}^2$ and informativeness as $(q^{past} - \hat{q})^2$, leading to the reputation boost estimate

$$(q^{past} - \hat{q})^2 \cdot f^S_{\alpha,M}(\hat{\sigma}^2).$$

Once the user chooses a paper, he will enter the grade $\hat{q}$ for it. This user model is consistent with the assumptions of Theorem 2, except that here each user is aware of its own variance (σ) and assumes that all previous reviewers have the same

variance $\bar{\sigma}$. The same-variance assumption for previous reviewers is motivated by the fact that we envision reviews to appear anonymous, so that users cannot infer the variance of a review from the identity of its author.

5.2 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.
- **Selfish:** the user chooses the paper that maximizes the estimated reputation boost, as described in the user model, with a varying value of α.
- **Accuracy:** the user chooses the paper that minimizes the estimated accuracy loss.
- **Informativeness:** the user chooses the paper that maximizes the estimated informativeness loss.
- **Optimal:** the user chooses the paper that maximizes global loss decrease, assuming that she knows the real quality of all papers, but still grades according to her beliefs.

The "random" criterion is used to measure the performance of our incentive system compared to a system where users, lacking incentives, pick the paper they wish to review uniformly at random. The "accuracy" and "informativeness" criteria are used to show that an appropriate combination of these two (a.k.a. selfish choice) is more effective than either of them separately. The "optimal" criterion is deliberately irrealistic and meant to serve as a reference for the fastest possible global loss decrease compatible with user beliefs about paper quality.

5.3 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 current paper quality estimate and the paper intrinsic quality q^{true}.

To illustrate how more expert reviewers receive more reputation (total review bonus points), we also report in Figures 6 and 7 the Pearson and Spearman correlations between user competence (the user-typical error σ^t discussed above) and

the reputation at the end of the experiment. Additionally, the fourth column in Figures 6 and 7 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.

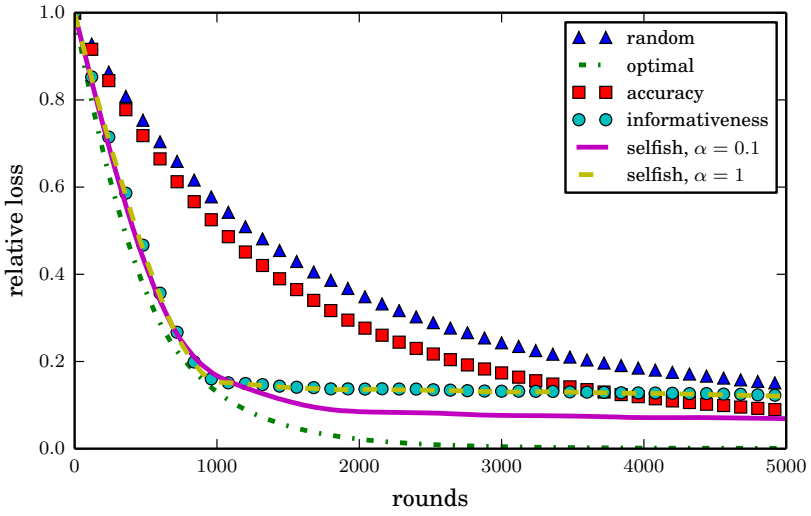

Figure 2: Relative global loss in the first user model.

5.4 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.

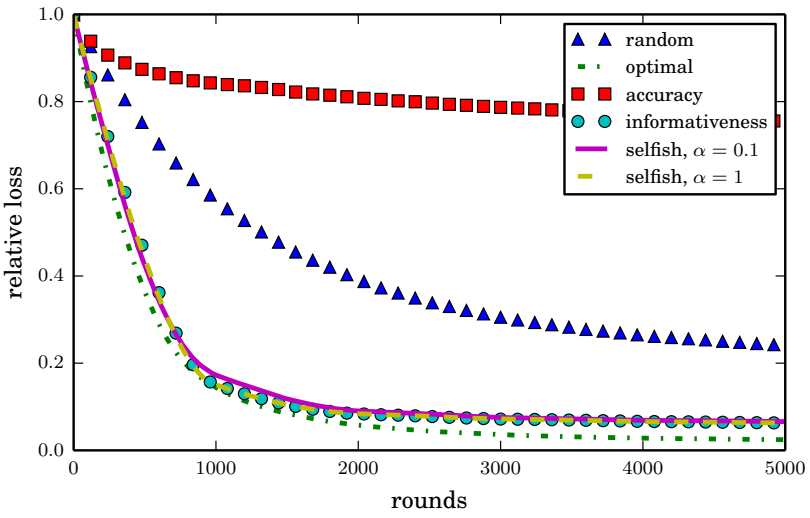

Figure 3: Relative global loss in the second user model.

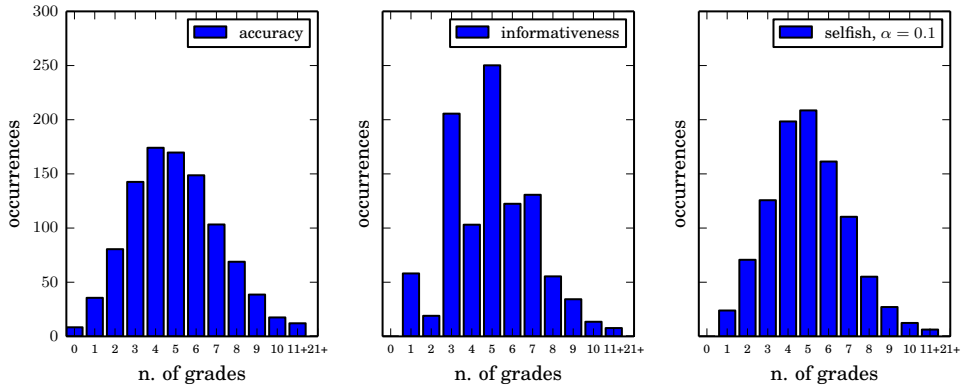

Figure 4: Distribution of the number of grades per paper in the first user model. The labels 11+ and 21+ stand for the intervals $[11, 20]$, $[21, \infty)$, respectively.

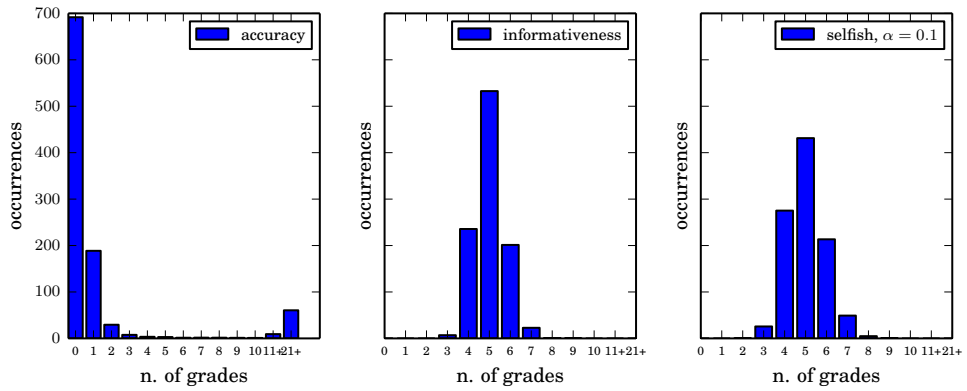

Figure 5: Distribution of the number of grades per paper in the second user model. The labels 11+ and 21+ stand for the intervals $[11, 20]$, $[21, \infty)$, respectively.

choice criterion	loss	Pearson	Spearman	rel. error	rel. error std. dev.
random	0.24	-0.005	-0.006	0.999	0.003
accuracy	0.17	0.011	0.011	0.536	0.003
informativeness	0.13	0.069	0.063	1.049	0.003
optimal	0.00	0.015	0.014	0.972	0.003
selfish, $\alpha = 0.1$	0.08	0.162	0.161	0.719	0.002
selfish, $\alpha = 1$	0.13	0.070	0.070	1.030	0.004

Figure 6: 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.

choice criterion	loss	Pearson	Spearman	rel. error	rel. error std. dev.
random	0.30	-0.002	-0.004	1.002	0.004
accuracy	0.79	0.323	0.228	0.944	0.003
informativeness	0.07	0.020	0.020	0.769	0.002
optimal	0.04	-0.028	-0.030	0.806	0.004
selfish, $\alpha = 0.1$	0.08	0.157	0.155	0.740	0.004
selfish, $\alpha = 1$	0.07	0.070	0.067	0.782	0.004

Figure 7: Summary data for the second user model. The columns have the same interpretation as Figure 6.

5.4.1 Results For First User Model

Figure 2 shows the value of the global loss, relative to the initial global loss, when paper choice is performed according to the five criteria from Section 5.2. The first column of the table in Figure 6 reports the loss after 3000 reviews. The other columns contain the Pearson and Spearman correlations between user competence and reputation, and the user propensity for reviewing papers on which they are most proficient.

The relative global loss curve shows that the selfish choice, for $\alpha \in \{0.1, 1\}$, performs very well and close to the optimal choice of papers, especially in the first 1000 rounds of the experiments. In fact, a closer inspection reveals that, when papers have the default starting score of 0 and no reviews yet, users simply choose papers with high perceived quality and no reviews, in order to reap a large informativeness bonus. Hence, at the beginning many papers go from the default score of 0 to approximately $\frac{M}{2}$, justifying the initial steep decline in global loss [5].

Then, consider the curve for the "informativeness" choice criterion. After the first 1000 rounds, when the above phenomenon leads to a near-optimal performance, the curve is essentially flat. Indeed, when users are only incentivized to provide informative grades, they will preferably select papers for which they have a very extreme opinion (very low or very high), leading to oscillation of paper scores, rather than convergence to the true value.

Notice that the global loss curve for $\alpha = 1$ is very close to the one for the "informativeness" choice criterion. This is due to the fact that the sigmoid $f^S_{1,M}(x)$ stays very close to 1 for relatively large values of x. Roughly speaking, the sigmoid "forgives" large accuracy errors. Hence, even if a user expects a significant accuracy loss, she can count on a reward almost equal to the expected informativeness bonus. Specifically, in our experiments we have $M = 10$ (grades between 0 and 10) and users have a maximum standard deviation of 5. So, their estimate for the accuracy loss is in the range $[0, 25]$, which corresponds to very limited reward penalties ($f^S_{1,10}(25) \approxeq 0.94$).

On the other hand, when $\alpha = 0.1$ even a small accuracy loss incurs a significant penalty on the reward, so the two components of the incentive are properly balanced. The global loss curve is initially steep and competitively positioned w.r.t. both the "optimal" curve and the curve based on accuracy alone. This is confirmed by Figure 6, reporting the relative error 0.719 for this case and a moderate correlation of 0.161 between competence and final reputation, higher than all other cases. Summarizing, data from this user model suggests that a choice of α close to 0.1 might be appropriate to the parameters of our populations.

[5] Moreover, as each paper is "known" to approximately 100 users, it is likely that at least one of them considers it to be of high quality, even if its quality is in fact low.

5.4.2 Results For Second User Model

Figures 3 and 7 show the relative global loss and the other performance measures for this model.

It may appear surprising that the choice based on accuracy alone performs even worse than the random choice. Indeed, when accuracy is the only incentive, users tend to focus on papers that have already received many reviews, because their quality can be more accurately predicted on the basis of the previous ratings. This creates a perverse incentive, in which the papers whose quality is best known draw the most evaluations. Figure 5 confirms that in that case more than 50 papers receive a very large amount of ratings, whereas 700 papers are completely neglected. The distribution of the number of grades per paper becomes much more balanced with the selfish choice and $\alpha = 0.1$, when the informativeness term mitigates the above issue.

Similarly to the other user model, our incentives with $\alpha = 0.1$ display the best overall performance, with 8% global loss after 3000 reviews, positive correlation between competence and reputation, and low relative error of 0.74, proving a clear bias for choosing papers on which the user is particularly competent.

In practice, this set of experiments suggests that the proposed incentive scheme may provide strong advantages, compared to rewarding accuracy alone, once it has been properly tuned to the characteristics of the user and paper populations.

Comparing Figures 2 and 3, we note that even in the optimal case, the global loss decreases faster for the first user model than for the second one. This can be explained by noting that in the first user model, users grade papers according to their individually-formed opinion, without access to other user's reviews. If n users provide grades for a paper, and the grades are then averaged, the individual opinions of each user account for $1/n$ of the average, which is optimal lacking information on the accuracy of individual users. In the second user model, instead, users use Bayesian inference to improve the accuracy of their estimate on the basis of reviews of previous users. As a consequence, the individual estimate y_i of the i-th user accounts only for $1/i$ of the grade provided by user i (assuming constant variances), and for $\frac{1}{n} \cdot \sum_{k=i}^{n} \frac{1}{k}$ of the complete average grade. This non-uniform weighing of the individually formed opinions is not optimal, and slows loss decrease.

6 Conclusions

In this white paper, we advocate a shift from pre to post-publication peer review for scientific papers. The chief benefit of post-publication peer review is the more timely circulation of scientific ideas, which can be shared as soon as the authors

decide to publish them. The key to a successful process of post-publication peer review consists in creating venues where authors are willing to post their papers for review, and where reviewers are incentivized to do useful and fair review work.

To facilitate this, we are proposing to create a tool, TrueReview, in support of post-publication peer review. TrueReview will allow people to set up new venues where papers can be submitted (for example, corresponding to conferences or special topics), as well as venues that index papers appearing in open-access venues such as arXiv. To encourage useful and accurate reviews, TrueReview will list with similar prominence both papers and reviewers: the papers will be ranked according to their quality, as assessed by the reviewers, and the reviewers will be ranked in order of the total *review bonus* they have accrued. The review bonus thus works as an incentive for reviewers.

We propose to award review bonus according to a combination of review *accuracy* and *informativeness.*

The accuracy measures the precision of a review's evaluation, in light of future evaluations. Judging a review only in view of future ones is instrumental in creating a truthful incentive for reviewers, where expressing their own best judgement on the paper's quality is an optimal strategy. Furthermore, measuring the accuracy of a review by comparing it with future reviews only rewards people who discover significant facts about papers, explain them in their review, and thereby influence future reviews.

The informativeness of a review is a measure of how much the future evaluation of a paper differs from the current one. Awarding a bonus for informativeness thus creates an incentive for reviewers to select papers who have received no or few reviews, or whose reviews are grossly imprecise. As the informativeness of a review is unrelated to the rating expressed in the review itself, including informativeness in the bonus does not alter the truthful nature of the incentive schemes.

We combine the accuracy and informativeness schemes in a multiplicative fashion, such that reviewers need to be *both* accurate and informative in order to obtain a bonus. This prevents lazy review strategies, such as picking papers with a large number of reviews and simply restating the consensus opinion on these papers (accurate but not informative), or picking new papers and just entering a random review (informative but not accurate).

We have experimented with two users models: one in which users base their review on their opinion only, and another in which they examine and account for previous ratings, before forming their opinion of the paper's quality. For both user models, our experiments show that the review bonus that combines both informativeness and accuracy is superior to considering either informativeness or accuracy alone, and is superior also to offering no specific bonus, and resorting on simpler methods such as simply counting how many reviews each user has provided.

References

[Bal15] P. Ball. Leading mathematician launches arXiv 'overlay' journal. *Nature*, 526:146, October 2015.

[BG07] Rajat Bhattacharjee and Ashish Goel. Algorithms and incentives for robust ranking. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 425–433. Society for Industrial and Applied Mathematics, 2007.

[Bha14] Malay Bhattacharyya. Viability of Crowd-Volunteered Open Research Reviews. In *Second AAAI Conference on Human Computation and Crowdsourcing*, 2014.

[dS13] Jaime A. Teixeira da Silva. The need for post-publication peer review in plant science publishing. *Frontiers in plant science*, 4, 2013.

[DSdA11] Atish Das Sarma and Luca de Alfaro. Crowdsourcing Scientific Paper Review. Technical Report UCSC-SOE-15-21, UC Santa Cruz, June 2011.

[Geh00] Edward F. Gehringer. Strategies and mechanisms for electronic peer review. In *Frontiers in Education Conference, 2000. FIE 2000. 30th Annual*, volume 1, pages F1B–2. IEEE, 2000.

[Geh01] Edward F. Gehringer. Electronic peer review and peer grading in computer-science courses. *ACM SIGCSE Bulletin*, 33(1):139–143, 2001.

[Gib16] E. Gibney. Open journals that piggyback on arXiv gather momentum. *Nature*, 530:117–118, February 2016.

[Gro10] Trish Groves. Is open peer review the fairest system? Yes. *Bmj*, 341:c6424, 2010.

[Her12] Daniel M. Herron. Is expert peer review obsolete? A model suggests that post-publication reader review may exceed the accuracy of traditional peer review. *Surgical endoscopy*, 26(8):2275–2280, 2012.

[Hun12] Jane Hunter. Post-publication peer review: opening up scientific conversation. *Frontiers in computational neuroscience*, 6, 2012.

[Kha10] Karim Khan. Is open peer review the fairest system? No. *Bmj*, 341:c6425, 2010.

[Kri12] N. Kriegeskorte. Open evaluation: a vision for entirely transparent post-publication peer review and rating for science. *Frontiers in Computational Neuroscience*, 6, 2012.

[Mar14] J. Marlow. Incentivizing peer review: The last obstacle for open access science. *Wired*, 7 2014.

[MS09] Michael R. Merrifield and Donald G. Saari. Telescope time without tears: a distributed approach to peer review. *Astronomy & Geophysics*, 50(4):4–16, 2009.

[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.

[OR94] Martin J. Osborne and Ariel Rubinstein. *A course in game theory*. MIT press, 1994.

[PHC+13] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579*, 2013.

[Rob01] Ralph Robinson. Calibrated Peer Review: an application to increase student reading & writing skills. *The American Biology Teacher*, 63(7):474–480, 2001.

[Spi02] Ray Spier. The history of the peer-review process. *Trends in Biotechnology*, 20(8):357–358, August 2002.

[STOA13] John Sadauskas, David Tinapple, Loren Olson, and Robert Atkinson. CritViz: A Network Peer Critique Structure for Large Classrooms. In *EdMedia: World Conference on Educational Media and Technology*, volume 2013, pages 1437–1445, 2013.

[TT12] George Tziralis and Ilias Tatsiopoulos. Prediction markets: An extended literature review. *The journal of prediction markets*, 1(1):75–91, 2012.

[vRDE10] Susan van Rooyen, Tony Delamothe, and Stephen JW Evans. Effect on peer review of telling reviewers that their signed reviews might be posted on the web: randomised controlled trial. *BMJ*, 341:c5729, 2010.

[VRGE+99] Susan Van Rooyen, Fiona Godlee, Stephen Evans, Nick Black, and Richard Smith. Effect of open peer review on quality of reviews and on reviewers' recommendations: a randomised trial. *Bmj*, 318(7175):23–27, 1999.

[WZ04] Justin Wolfers and Eric Zitzewitz. Prediction markets. Technical report, National Bureau of Economic Research, 2004.