Simulation-based evaluation of a crowdsourced expert peer review system

Imre Lendak[1,2]
[1]Faculty of Technical Sciences, University of Novi Sad, Novi Sad, Serbia
[2]Faculty of Informatics, Eötvös Loránd University, Budapest, Hungary
lendak@uns.ac.rs

Abstract— **The primary goal of this paper is to propose and simulate crowdsourcing-based solutions which might optimize the scientific peer review system. More specifically, a global reviewer database and gamification techniques will be proposed with the goal of obtaining more high-quality reviews for papers received by journals. The proposed modifications were assessed in a multi-agent simulation environment, in which the members of the reviewer crowd were modeled as agents. Our simulation-based evaluations implemented in the MASON multi-agent environment showed that the introduction of the above improvements would allow editors to find the most suitable and responsive reviewers, as well as to lower the number of scientific papers which do not receive enough reviews.**

***Keywords**— peer review system, crowdsourcing, multi-agent simulation, gamification*

I. INTRODUCTION

Most researchers are aware of the pressing problems in the present-day scientific peer review system, in which authors write papers, submit them to conferences or journals, editors or conference organizers send them out for review, volunteer reviewers review the papers and provide feedback, based on which editors/organizers make decisions about the fate of the scientific contributions (i.e. accept, accept with modifications or reject). The manifold problems faced by the actors in this system can be classified into the problem classes faced by authors, reviewers, editors and the scientific community in general. Authors usually wait for excessively long periods between paper submission and the receipt of reviews or decisions. The reviews are sometimes performed by non-experts who fail to provide high quality feedback and criticism due to a lack of complete understanding of the scientific work. Reviewers are contacted by journals and conferences to review papers which often fall outside their main areas of expertise and are pressured by editors to complete the reviews as soon as possible. Reviewers are usually not paid for their work. Editors might not be entirely aware of the specific expertise of the reviewers whom they have in their contact lists or reviewer databases they can access. Reviewer databases are fragmented and maintained in different journal or conference management systems. There are even worse cases, in which the reviewer databases are kept in Excel (or similar) documents. Editors often need to send out many email invitations to reviewers and multiple reminders to acquire a minimum number of reviews, which is a considerable overhead. When they do not receive sufficient numbers of high quality reviews, editors might err due to a lack of information, or decide to reject papers which did not receive sufficient numbers of reviews. Those errors can be classified into false positives (lower quality papers accepted) and false negatives (high quality papers rejected). The cases of false positives and negatives affect the entire scientific community, by failing to ensure that only the papers with the highest merit get accepted and published. For example, papers with radical new ideas might have significant difficulties before being published in the current system.

Some argue that the scientific community should switch to a non-voluntary review system, in which reviewers would be financially rewarded for their work. That solution would be similar to the project proposal review systems implemented by the European Union (and other funding bodies), in which the reviewers receive a small financial reward for each (project) proposal reviewed. Such a scheme is implemented to a certain degree by scientific venues which require authors to pay a fee for publishing papers. Although it can be quite efficient, it poses great risks to the scientific community, because the wider acceptance of such models might lead to a situation in which the rich publish and the poor perish. In that scenario high quality research results might not be published at all if they fail to acquire funds for publishing.

The goal of this paper is to propose a different solution to the above listed problems. We propose to implement a more transparent, crowdsourced expert peer review system, which would consist of a global reviewer database (i.e. the crowd of reviewers) and additional tools at the disposal of editors. Instead of financial rewards, the reviewers could be motivated in a game-like fashion, via the introduction of reviewer points, levels, leaderboards, badges and other accomplishments received after completing a certain number or type of reviews. Both the traditional peer review system and its augmented variants were modeled as games played by the crowds of authors, editor and reviewers. Those models are then in turn investigated in a multi-agent simulation environment, in which all actors are agents aiming to maximize their utilities, e.g. authors to maximize the probability of paper acceptance, reviewers to review only high-quality papers in their specific areas of expertise and editors to receive a maximum number of high quality reviews and thereby be able to accept and publish the best papers.

978-1-5386-9151-9/19/$31.00 ©2019 IEEE

II. Related Works

There is a long history of papers analyzing the issues present in the scientific peer review system (PRS). Somewhat surprisingly, most papers dealing with this topic fall into the domain of medicine. In reference [4] the authors claim that although the review system is central to science, science has little to say about it, i.e. it is not widely analyzed and optimized by researchers. The authors of reference [1] discuss the system's merits and weaknesses, as well as some proposed changes to mitigate them. One such change proposed is to create more open systems, e.g. publishing all reviews and comments received alongside the papers. Such transparency allows readers to assess both the paper and its reviews. It also allows authors to post answers to the reviews. The authors also claim that reviewers need to be trained and accredited, thereby allowing them to provide adequate feedback and detect fraud. Walsh et al in reference [13] analyze anonymous peer review and evaluate the feasibility of an open PRS.

The authors of reference [6] claim that the peer review system is biased. That thought is further examined by others, who claim that the system is especially biased because reviewers give better reviews to authors belonging to their social groups based on an analysis of peer reviews of postdoctoral fellowship applications [14]. Unethical practices (e.g. unsupported findings, plagiarism) is a significant issue in the peer review process as well [10].

The introduction of digital services assisting the various actors of the state-of-the-art PRS introduced numerous benefits and were assessed as early as 1996 [5]. The authors of reference [12] analyzed the role of bibliometrics in automatic peer review in the same year. There were significant new developments in the 20+ years since 1996. Today, editors can usually rely on a journal or conference management systems' reviewer database. Those databases store reviewer expertise information, i.e. list the key domains in which they are highly skilled. The editorial systems can be configured to automatically remind reviewers who failed to respond. Some leading journals give strict deadlines (e.g. 30 days to complete a review) to reviewers and remove reviewers who fail to meet those deadlines. Modern review systems might include algorithmic support for computing the accuracy of reviewers, e.g. by comparing them to the contributions of other reviewers [2]. Such systems usually do not utilize techniques developed in gamification to further motivate reviewers, although the effectiveness of such schemes was assessed in other domains [9] and in crowdsourcing as well [10]. The authors of reference [7] assessed the potential benefits of applied gamification on the throughput of the peer review system measured in review length and editor load.

Additional developments are necessary to truly harness the capabilities of modern digital systems and services. The goal of this paper is to explore the viability of those possibilities via simulation-based assessments.

III. Present-day Review System

Most current review systems rely on anonymous reviews, i.e. they do not disclose their reviewers' identities neither to authors nor to readers (of accepted papers). They rely on an event and review management system to store submitted papers, review results and acceptance decisions. These systems might store additional information about reviewers, e.g. domains of expertise, past performance, e.g. how many days they usually need to complete a review, percentage of completed vs not completed reviews.

As discussed in the previous sections these systems tend to be slow, the quality of reviews is often lower and some of the best papers (especially those presenting radically new ideas) might get rejected due to some form of bias (e.g. reviewers' bias towards the well-known and commonly accepted). Some of the key causes of these problems are the insufficient number of high-quality reviews, the lack of globally accessible reviewer databases, and the inherent non-transparent nature of the entire review process.

IV. Proposed Future Review System

In this subsection we outline a novel scientific peer review process addressing the above identified problems. More specifically we will describe the potential benefits brought forward by a global reviewer database and the introduction of gamification techniques in to the (scientific) peer review system.

A. Global reviewer database

As a key measure to counter the **low number of high quality reviews** in the current expert peer review system, we propose to introduce a global reviewer database, which would contain sufficient information about the crowd of reviewers whom editors might contact with reviews. The minimum information about each reviewer would consist of basic information (full name, contact information) extended with the list of very specific scientific domains of high levels of expertise. For example, instead of 'computer science', the reviewers might be dubbed as experts in 'crowdsourcing', 'distributed algorithms' or similarly specific domains. Additionally, the reviewer database would contain information about every reviewer's past reviewing performance consisting of at least the number of completed/not completed reviews, average number of days to complete a review and review alignment with other reviewers' findings, e.g. if there were four reviews for a single paper then their similarity would be measured and outlying reviewers (who give strikingly different reviews compared to other reviewers) would be assigned (negative) points.

B. Gamification in peer review

We propose to introduce **gamification techniques** applied in other crowdsourcing domains to counter the above described possible negative effect. More specifically, we advocate that the global reviewer database should be accompanied by a point system, levels, leaderboards and accomplishments. Reviewers would receive points for each reviewing related activity, obtain levels and maintain a position on a global reviewer leaderboard within their specific scientific domains based on their reviewing activity, e.g. the number of completed reviews, the timeliness of review submissions, the accuracy of reviews (i.e. similarity to other reviews submitted). The gamification element could incorporate challenges and link them to accomplishments, e.g. a special accomplishment would be reachable via doing a certain type and number of reviews within a system-defined period.

Additionally, the inherently **non-transparent** nature of the current peer review system can be mitigated by publishing all reviews alongside the papers. Depending on the journals' assessment, the reviews published might be anonymous or attributed to the reviewers. In both cases, openly publishing reviews together with accepted papers would encourage scientist to additionally debate not just the papers, but the reviews as well. This change comes with a risk, namely it might further lower the number of received reviews, as scientists would have to invest more time and effort to be absolutely sure that their reviews are not found to be of lower than necessary quality. Additional reviewer points received for published reviews and up-voted, useful comments written about reviews published by others could be used to reward reviewers and motivate them to collaborate in such exchanges.

V. Simulation Environment

In this section we describe the simulation environment used to investigate both the traditional peer review system and its modified variants.

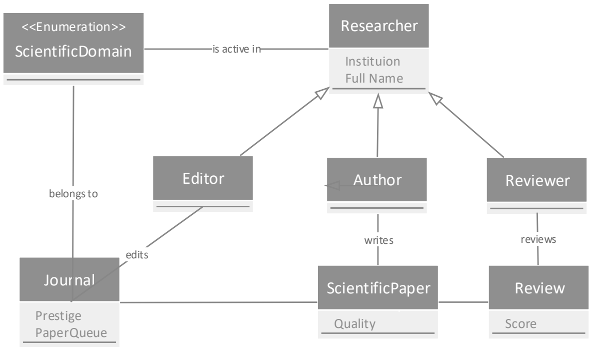

Fig. 1. Class diagram of peer review system actors

In Figure 1 we present a class diagram with the most important entities of the peer review system. There are three types of researchers listed (editor, reviewer, author), scientific papers, reviews and journals. Both researchers and journals belong to one or more scientific domains. Each researcher can have one or more of the roles discussed above, i.e. author, editor, reviewer. Theoretically, one researcher can be in all three roles at the same time. In the below sub-sections we formalize the behavior of this crowd.

A. Authors

Class *Author* represents authors as agents who assigned to one or more scientific domains. They submit papers periodically at randomly chosen times, e.g. two times a year on average. Their papers have a randomly chosen quality around an average value. The journal to which they submit their new papers is selected randomly from the set of journals in their scientific domains. For simplicity, authors give up when their papers are rejected, i.e. the re-submission process is not modeled, although it could be easily incorporated into this research. Authors are aware of the (randomly chosen) quality of their papers, but it does not affect their journal choices, i.e. they might submit strong papers to weak journals in our peer review model.

B. Editors

Class *Editor* models the behavior of real-world editors and the two decisions they need to make about each paper they manage. The first editorial decision (D_e) is whether the paper is of sufficient quality and in scope and it is made upon receiving a new paper. We model this decision-making step with equation (1), i.e. if the (perceived) paper quality is significantly lower than the journal's prestige, than the editor immediately rejects is. Value '*pq*' is the (editor's) perceived paper quality, '*jp*' is the journal's perceived prestige. Constant c_1 allows us to configure the sensitivity of the editors, e.g. lower their sensitivity. In this decision phase the editors might reject papers and do not send them out to reviewers for review.

$$D_e : pq \geq jp - c_1 \tag{1}$$

The second decision about a paper described by equation (2) and it is made when the reviews come in from the reviewers. It is based on the average review scores as a measure of the true paper quality and the journal's paper queue (*jq*), which stores received papers which are assigned to editors and/or reviewers. Constant c_1 allows us to configure the sensitivity of the editor and it might allow him/her to accept papers which are slightly below the perceived prestige of the edited journal.

$$D_r : avg(reviews) \geq jp + jq + c_2 \tag{2}$$

It is important to note, that the status of the journal's queue is a dynamic variable, i.e. it changes as time elapses. The other variables in the above formulae are time-invariant, i.e. they do not change with time. Another important aspect of the model is the editorial paper rejection when there are no reviews for papers. If the editors contact all available reviewers in a specific scientific domain, for example via a journal's reviewer database and do not manage to collect sufficient numbers of reviews, then they might reject papers. This secondary editorial rejection was an important aspect in the experimental evaluation presented in the next section.

C. Reviewers

The main task of class *Reviewer* is to model the behavior of reviewers and their two (main) decisions when assigned a paper for review: first they decide whether to accept the review and if they accept, then they do the review and assign a score to the paper. These decisions can be depicted by below formulae (3) and (4).

$$D_a: (jp \pm c_3) \leq (pq \pm c_4) - rq \tag{3}$$

Reviewers take into consideration the perceived paper quality (*pq*), the journal prestige (*jp*) and their queue of paper reviews (*rq*) as shown in formula (3). The queue is relevant, as the longer it is, the reviewer is more likely to decline to review due to his/her high workload. Constants c_3 and c_4 allow us to configure how well reviewers assess journals and papers in this phase.

We model reviewer bias in the scoring and decision-making process with formula (4). The reviewer assigns a score (*c_5*) to each reviewed paper which might significantly vary around its true quality *pq*. Constant *c_5* essentially depicts the maximum reviewer bias.

$$D_p = pq \pm c_5 \tag{4}$$

There is an edge case as well, in which the review is accepted, but never completed, e.g. due to reviewer overload or forgetfulness (i.e. the reviewer forgets to submit a review on time). Although this use case was not modeled in the work presented, it could be easily included as part of future research.

The following additional classes from Figure 1 were implemented as well:

- *Journal* class instances contain a journal prestige and maintains a list of submitted papers, editors and (known) reviewers.
- *Paper* instances encapsulate quality and references to authors, scientific domains and authors.
- *Review* instances maintain assignment and completion dates, reviewer decisions (accepted or rejected) and references to a paper and a reviewer.

The simulation environment's central class was named *PeerReview* and its task was to coordinate the simulation and to maintain lists of scientific domains, journals, researchers, papers and reviews. This class is not shown in Figure 1.

VI. Results

We implemented the above described simulation scenario in the MASON [8] simulation environment. MASON is a high performance, multi-agent simulation environment, which allows researchers to model and simulate the behavior of complex systems. It was used to model authors, reviewers and editors as agents, which perform their research-related activities periodically, e.g. on a weekly basis. The custom agent behavior code was written in the Java programming language. In our case this meant that the classes from Figure 1 were implemented in Java and the behaviors of the various actors in the scientific peer review system presented in the previous two sections were transformed into Java code. The author published a similar simulation-based study about a crowdsourced parking spot monitoring solution [3].

We configured the simulation environment with twenty scientific subdomains of computer science, ranging from CPU design to crowdsensing. Each author and reviewer was assigned three scientific domains chosen randomly. The most important general simulation settings applicable in all experiments are shown in Table I below.

TABLE I. Common Simulation Settings

Sim. length	*Step size*	*Journal count*	*Reviewer count*	*Author count*	*New paper every*
2 years	1 week	1,000	5,000	10,000	26 weeks

The simulation step was one-week-long and the total simulation lengths were two years, i.e. 104 weeks. There were 1,000 journals with separate editors, 50,000 reviewers and 10,000 authors who produced new papers every 26 weeks (half year) on average. Those new papers were assigned a randomly chosen quality based on the below formula, i.e. random number around an average quality of 65 points.

$$q_{paper} = avg_{quality} \pm 35, avg_{quality} = 65 \tag{5}$$

Reviewers assessed papers and assigned a numerical score between 0 and 100 based on formulae (3) and (4). Editors consider those scores and compare them to their journal's prestige which is calculated at simulation setup and was calculated by formula (5).

The constants c_i in formulae (1) to (4) were configured in the following manner:

- c_1 was set to -20, thereby allowing editors to consider lower quality papers and submit them for review.
- c_2 was set to 0, practically disallowing the publication of papers with quality lower than the journal's prestige.
- Both c_3 and c_4 were set to 10, thereby allowing reviewers to make randomly chosen, moderately erroneous journal and paper quality assessments by +/- 10 percentage points.
- c_5 was set to 20 percentage points, thereby modeling (potentially significant) reviewer bias and variance around the exact quality of scientific papers.

After entering the above configuration settings, three experiments were conducted to investigate the operation of the traditional peer review system, an upgraded system with a global reviewer database and a peer review system employing gamification techniques and thereby motivating reviewers.

A. Experiment 1: Traditional peer review system

In the present-day scenario editors have access to a limited pool of expert reviewers via separate reviewer databases maintained by each journal or the editors themselves. We modeled this by randomly selecting 10% of domain experts and adding them to the journals' reviewer databases. We configured each journal to belong to a single scientific domain. In this scenario the reviewers submitted anonymous reviews, which allowed them to submit significantly different decisions, which was modeled by allowing the review scores to be widely spread around the true paper quality.

TABLE II. PRESENT-DAY PEER REVIEW SIMULATION RESULTS

Accepted / declined review ratio	*Editorial rejection*	*Paper acceptance ratio*
0.23	0.10	0.19

With the above presented simulation settings we measured the values shown in TABLE II. The 10,000 authors in this simulation submitted roughly 26,000 papers to the 1,000 journals. In our experiment roughly every fifth reviewer contacted accepted to do the reviews – see column one in Table II presenting the accepted to declined paper review ratio. This number significantly depended on the value of constant c_3, which, if configured to be a negative value, allowed the reviewer agents to accept reviews even if they thought that the papers had somewhat lower quality compared to the journal's prestige (to which the paper was submitted).

We also measured the number of reject decisions by the editors. There were two reasons for making such decisions: either the received papers were of lower quality than the editor's threshold, or the editor failed to find at least three reviewers in the relevant scientific domain. In our first experiment we found that ~10% of all decisions were of this kind (see column three in Table II above). The number of these decisions depended on the accessibility of expert reviewers via the journals' reviewer databases. We also measured the general acceptance ratio of papers, which was a realistic ~19% (see column three in Table II).

B. Experiment 2: Global reviewer database

We modeled the availability of a global reviewer database (as opposed to fragmented reviewer databases maintained by journals and conferences) via assigning all reviewers (i.e. Reviewer instances) to three scientific domains and making them available to (all) editors, i.e. editors had access to the complete set of reviewers active in a certain scientific domain, as opposed to the previous experiment, in which only a limited number of randomly chosen reviewers were registered in each of the journals' reviewer database.

Similarly to the previous experiment, editors made random reviewer choices, i.e. the selection process was not based on past reviewer performance or level of perceived expertise. These aspects of the peer review system will be investigated by the authors as part of their future work

TABLE III. GLOBAL REVIEWER DATABASE SCENARIO

Accepted / declined review ratio	*Editorial rejection*	*Paper acceptance ratio*
0.21	0	0.19

The significantly increased size of the reviewer pool (~10% of all reviewers in experiment #1 compared to 100% in this scenario) had a marked positive effect on the editorial rejection rate, namely it completely cancelled it – see column two in Table III. The rest of the values were not affected significantly when compared to those in Table II.

C. Experiment #3: Gamification in peer review

We theorize that if the reviewers accessible via the global reviewer database were motivated via assigning them points, levels, a global leaderboard and various accomplishments and challenges, it would also motivate them to accept more reviews and complete them more regularly. The introduction of these additional elements can be modeled by modifying the initial reviewer decision shown in formula (3) as shown in formula (6).

$$D_a: (jp \pm c_3) \leq (pq \pm c_4) - rq + game \qquad (6)$$

We expected to see a linear increase in review acceptance on the reviewers' side when the value of variable '*game*' was set to a positive value. This hypothesis was proven experimentally as shown in the below table IV. The key takeaway of this table is in its last row, where we can see that the ratio of accepted to declined paper reviews can be higher than one if we manage to build a gamification-based motivation scheme which increases the likelihood of review acceptance by ten points (on a 0 to 100 scale). In such a scheme it would be sufficient to send out six review invitations to receive at least three reviews. Knowing that some editors need to contact tens of potential reviewers to obtain that many reviews today [7], this would be a significant improvement.

TABLE IV. GAMIFICATION-INTRODUCED BENEFITS

Gamification points ('game')	*Accepted / declined review ratio*
1	0.25
3	0.31
5	0.43
10	1.12

The above discussed experiments were run three times and the measured values were averaged to verify their correctness. A personal computer (laptop) with an Intel i7 processor and 8 GB of memory was used for testing purposes. Memory use rose up to ~2 GB during testing and the CPU loads also reached 100%. A single simulation run usually lasted less than one minute.

VII. CONCLUSION

This paper presents a what-if analysis of the traditional peer review system, analyzes some of its drawbacks and proposes to introduce a global reviewer database, gamification techniques and public reviews (i.e. publishing reviews alongside with papers). We modeled both the present-day scientific peer review system and the proposed modifications with multi-agent models and implemented them in the MASON multi-agent environment. The models were used to assess (1) the positive effects a global reviewer database and (2) the beneficial effects of implementing gamification techniques to motivate reviewers to accept paper reviews, e.g. their impact on the likelihood of reviewers accepting to review scientific papers when asked to. It was shown in the simulation environment that the introduction of a global reviewer database might bring the number of paper rejections (by editors) caused by insufficient numbers of reviews to zero. We also showed that the introduction of gamification techniques (e.g. a point system, leaderboards, challenges, accomplishments) would boost review acceptance, eventually making it more likely that a review is accepted than not when the sum of such motivations reaches a certain threshold.

As future work, the author intends to implement more complex decision models, to assess a truly transparent, open review system's benefits, in which reviews would be published alongside the scientific papers, as well as reviewer selection based on past performance.

Acknowledgment

This research was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia under grants III-42004 and TR33013.

References

[1] D.J. Benos et al, "The ups and downs of peer review," Advances in physiology education, 2007, vol 31(2), pp. 145-152.

[2] K. Cho & C.D. Schunn, "Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system", Computers & Education, 2007, vol 48(3), pp. 409-426.

[3] Farkas, K., & Lendák, I., "Simulation environment for investigating crowd-sensing based urban parking", 2015 International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS), 2015, pp. 320-327.

[4] S. Goldbeck-Wood, "Evidence on peer review--scientific quality control or smokescreen?" British Medical Journal, 1999, vol 318 (7175), pp. 44-45.

[5] S. Harnad, Implementing peer review on the net: scientific quality control in scholarly electronic journals, Scholarly publishing: the electronics frontier. MIT Press, Cambridge, MA, 1996.

[6] C.J. Lee, C.R. Sugimoto, G. Zhang & B. Cronin, "Bias in peer review," Journal of the Association for Information Science and Technology, 2012, vol 64 (1), pp. 2-17.

[7] I. Lendak & K. Lendak-Kabok, "Game theory, crowdsourcing and the peer review system," Research Evaluation in the Social Sciences and Humanities (RESSH 2017), Antwerp, Belgium, 2017, pp. 86-91.

[8] S.Luke, C. Cioffi-Revilla, L. Panait, K. Sullivan & G. Balan, "MASON: A multi-agent simulation environment", Simulation: Transactions of the Society for Modeling and Simulation International, 2005, vol 82 (7), pp. 517-527.

[9] E. Meckler, F. Brühlmann, K. Opwis & A.N. Tuch, "Do points, levels and leaderboards harm intrinsic motivation?: an empirical analysis of common gamification elements," 1st International Conference on Gameful Design, Research, and Applications, Toronto, Canada, 2013, pp 66-73.

[10] B. Morschheuser, J. Hamari & J. Koivisto, "Gamification in Crowdsourcing: A Review," 49th Annual Hawaii International Conference on System Sciences (HICSS), 2016, pp. 4375-4384.

[11] F. Napolitani, C. Petrini & S. Garattini, "Ethics of reviewing scientific publication," European Journal of Internal Medicine, 2017, vol 40, pp. 22-25.

[12] van Raan, Antony, "Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises," Scientometrics, 1996, vol 36.3, pp. 397-420.

[13] E. Walsh, M. Rooney, L. Appleby & G. Wilkinson, "Open peer review: a randomised controlled trial," The British Journal of Psychiatry, 2000, vol 176(1), pp. 47-51.

[14] C. Wenneras and A. Wold, "Nepotism and sexism in peer-review," Women, Science, and Technology, 2001, pp. 46-52.