Advances in Complex Systems

A simulation of disagreement for control of rational cheating in peer review

FRANCISCO GRIMALDO

Departament d'Informàtica, Universitat de València
Av. de la Universitat, s/n, Burjassot, 46100, Spain,
francisco.grimaldo@uv.es

MARIO PAOLUCCI

Institute of Cognitive Sciences and Technologies
Italian National Research Council
Via Palestro 32, Roma, 00185, Italy
mario.paolucci@istc.cnr.it

Received (received date)
Revised (revised date)

Understanding the peer review process could help research and shed light on the mechanisms that underlie crowdsourcing. In this paper, we present an agent-based model of peer review built on three entities - the paper, the scientist and the conference. The system is implemented on a BDI platform (Jason) that allows to define a rich model of scoring, evaluating and selecting papers for conferences. Then, we propose a programme committee update mechanism based on disagreement control that is able to remove reviewers applying a strategy aimed to prevent papers better than their own to be accepted ("rational cheating"). We analyze a homogeneous scenario, where all conferences aim to the same level of quality, and a heterogeneous scenario, in which conferences request different qualities, showing how this affects the update mechanism proposed. We also present a first step towards an empirical validation of our model that compares the amount of disagreements found in real conferences with that obtained in our simulations.

Keywords: Artificial social systems; Peer Review; Agent-based simulation; Trust reliability and reputation

1. Introduction

Large scale collaboration endeavors amongst humans are making the headlines of scientific magazines and attracting the attention of the research community. The cases of Wikipedia and Amazon's Mechanical Turk are striking examples; some consider these ICT-mediated collaborations to be the first step in a transition towards collective intelligence [8, 38], a transition not devoid of risks as averaging effects [6] and isolation [26]. To understand if and how this transition is happening and what its consequences might be, we need to carefully examine the already existing social and cultural structures that anticipate, in part or in whole, this kind of collabo-

ration. The most important of these structures - a social artefact in itself - is the complex social institution known as *peer review*.

Peer review, the process that scrutinizes scientific contributions before they are made available to the community, lies at the core of the social organization of science. Curiously, while the measurement of scientific production, that is, the process that concerns the *citation* of papers - scientometrics - has been an extremely hot research issue in the last years, we can't say the same for what concerns the process of *selection* of papers, although some attention has been focused on its shortcomings. Indeed, the actual effectiveness of peer review in ensuring quality has yet to be fully investigated. In [24], the review process is found to include a strong "lottery" component, independent of editor and referee integrity. While the heterogeneous review approach to a decision between two options is supported by Condorcet's jury theorem[a], if we move beyond simple accept/reject decisions the simplicity of the solution disappears. A more sophisticated and precise outlook on peer review that considers scoring, ranking, and reputation would tell a different story; in fact, scoring has been shown to have non trivial effects on the reviewers' choice (see the marks distributions in [28]), rankings for citations have been shown to diverge from rankings resulting from peer review [12], and theory and practice of reputation systems [15, 17] have been proposed as potentially transformative approaches for traditional peer review. All these ideas could in turn help to detect kinds of potential failures that are not waived by Condorcet's theorem.

These issues are particularly relevant because peer review should take advantage of the new information publishing approach and technologies created by Web 2.0 and beyond. At the same time, diffuse dissatisfaction of scientists towards the current mechanisms of peer review is perceived - anecdotally, as list of famous papers that were initially rejected and striking fraudulent cases are published, and statistically, as numerical evidence on the failures of peer review [28] is starting to appear.

Peer review is an open social system, that is made complex by the interactions between its components and by role superposition - as an example, consider the feedback that can be activated by the same people acting as authors and reviewers. This complexity needs to be matched by a suitable modeling approach; peer review appears to be amenable to study by numerical, agent-based models [25] , that could be validated both on the micro and the macro level, and on which what-if analysis could be performed, thus testing "in silico" proposed innovations. Solutions should be searched through a federation of models, in a pluralistic modeling approach [21]; in this paper, we propose one possible agent-based model of peer review and, inspired by the introduction of rational cheaters in [39], we test how a simple mechanism based on disagreement control could help controlling this kind of cheating.

[a]The theorem states how, roughly speaking, if independent voters on a decision can do even slightly better than random, the more of them the better - thus supporting democratic approaches to decision (even if the hypothesis of independence is rather unrealistic).

The rest of the paper is organized as follows: the next section reviews the literature on simulation of peer review. We then outline a general model of peer review endowed with a reviewer disagreement control mechanism, with a few implementation details. In the results section, we show how the mechanism works under two different conditions. In the last section, we present our conclusions and draw the path for future work.

2. Related work

The idea of studying science using scientific methods is at the core of scientometrics [32]. From the standpoint and for the purposes of scientometrics, science can be defined as a social network of researchers that generate and validate a network of knowledge. Hence, many scientometric studies have described the structure and evolution of science, while a few others have aimed to replicate and predict the structure and dynamics of science. It is the latter group in which we locate this paper, that can be classified as a quantitative predictive domain-specific computational model.

Mathematical models have been proposed not only to explain statistical regularities [16], but also to model the spreading of ideas [19] and the competition between scientific paradigms [36] and fields [7]. Furthermore, they have been used to model the relation between publishing, referencing, and the emergence of new topics [18], as well as the co-evolution of co-author and paper-citation networks [9]. The model classes used for the mathematical modeling of science dynamics cover stochastic and statistical models, system-dynamics approaches, agent-based simulations, game-theoretic models, and complex-network models.

Many different conceptualizations of science are possible [10, 30] depending on the goal and type of modeling performed. Models that conceptualize science as a social activity will use researchers, teams, and invisible colleges as key social terms. Models that simulate science as a knowledge network have to define knowledge terms such as documents and journals. Models that place a central role on the bibliographic data used in model validation require a definition of bibliographic terms. Models that conceptualize science as an evolving system of co-author, paper-citation, and other networks will need to define network terms. Given the importance of textual documents in the practice of science, our model focuses on the quality assessment of papers produced by scholars.

Peer review is the principal mechanism for quality control in most scientific disciplines, as it determines not only what research results are published but also what scientific research receives funding or what fellowships are granted, thus clearly influencing scientific career [4]. For many years the peer review process has been a target for criticism in relation to traditional research criteria of: poor reliability, as reviewers rarely agree on their recommendations; low fairness, as reviewer's recommendations are frequently biased, that is, judgments are not based solely on scientific merit, but are also influenced by personal attributes of the authors, appli-

cants, or the reviewers themselves; and lack of predictive validity, because there is little or no relationship between the reviewers' judgments and the subsequent usefulness of the work to the scientific community. This paper deals with the problem of reliability, also coined in other works as "the luck of the reviewer draw" [13], as it faces the control of disagreements resulting from referees that follow a biased review behavior.

Scientific merit is multifaceted and up to nine areas have been identified [5] for the assessment of manuscripts and fellowships: relevance of contribution, writing/presentation, design/conception, method/statistics, discussion of results, reference to the literature and documentation, theory, author's reputation/institutional affiliation, and ethics. Although work dealing with the predictive validity of peer review has questioned the validity of judgements prior to publication [34], is has also pointed out the need for future peer review data to be analysed using multilevel models (e.g. referee ratings for the quality of the proposals) with either categorical or continuous response variables [22]. Following this latter point of view, our research focuses on the ex-ante evaluation of the potential impacts of a paper, as opposed to the ex-post process of counting citations for papers. In this scenario, some research [1] has used beta distributions to describing the quality of a paper regarding aspects such as: topic, technical quality and novelty. In our researh, we use an aggregated quality value for each paper that allows us to model bias in the peer review, present when factors that are independent of the quality of a submission correlate statistically with the judgement of reviewers.

The relation of reviewers' overall ratings and selection commitees' final decisions has been studied and different decision-making strategies have been analysed such as: rejection when all reviewers recommend rejection, rejection when any reviewer recommends rejection, and rejection when a majority of reviewers recommend rejection [33]. Many aspects of the peer review process vary case by case and this variation largely depends on the type of application. For example: reviewers and persons reviewed may or may not be anonymous (double-blind vs. single blind); reviewers may be assigned permanently or temporaly, as well as they may represent one scientific discipline or a variety of disciplines; a single reviewer or a commitee may provide a peer review; etc. Accordingly, the peer review process should be examined with regard to so-called interaction effects, because attributes of the authors or applicants and attributes of the reviewers are potential sources of bias in peer review.

Nicola Payette's [27] main premise is that science is some sort of distributed cognitive system that can be reverse engineered (and hopefully optimized) using agent-based models (ABMs). ABMs should then become part of the policy-makers toolbox, as they enable us to challenge some idealizations and to capture a kind of complexity that is not easily tackled using analytical models [32]. A striking example of the possibilities of agent-based modelling of science is Gilbert's model. Gilbert started out with a simple agent-based model of a candidate mechanism for simulatin Lotka's law [23] pattern for the distribution of papers per author. While scientists

played only a very small role in his first model, other researchers [37] have followed a cognitive approach in which authors were not merely passive placeholders, but cognitively capable individuals whose success or failure depends on their ability to learn in the scientific world.

Recent work has shown that there is a quantitative, model-based way to select among candidate peer-review systems [1]. It uses agent-based modelling to quantitatively study the effects of different alternatives on speed publication, quality control, reviewers' effort and authors' impact. As a proof-of-concept, it contrasts an implementation of the classical peer review system adopted by most journals, in which authors decide the journal for their submissions, with a variation in which editors can reject manuscripts without review and with a radically different system in which journals bid on manuscripts for publication. Then, it shows that even small modifications to the system can have large effects on these metrics, thus clearly demonstrating that peer review is a very complex system that cannot be fully described using simple models.

The work presented in this paper is inspired to the ideas in [39], where the authors focus on an optimizing view of the reviewer for his or her own advantage. To this purpose, they define a submission/review process that can be exploited by a *rational cheater* [11] strategy in which the cheaters, acting as reviewers, reject papers whose quality would be better than their own. In that model, the score range for review is very limited (accept or reject) and in case of disagreement (not unlikely because they allow only two reviewers per paper), the result is completely random. They find out that a small number of rational cheaters quickly reduces the process to random selection. The same model is expanded in [31], focusing not on peer review of papers, but of funding requests. Only a limited amount of funding is available, and the main focus is to find conditions in which a flooding strategy is ineffective. The number of cheaters, differently from this study and from [39], is not explored as an independent variable. However, similarly to the present work, the strong dependence of results from the mechanism chosen (number of reviews, unanimity) is evidenced.

In [20], the authors introduce a larger set of scores and use three reviewers for paper; they analyze the effect of several left-skewed distributions of reviewing skill on the quality of the review process. They also use a disagreement control method for programme commitee (PC) update in order to improve the quality of papers as resulting from the review process.

None of the models introduced above consider the reviewer effort as an important factor. Instead, in [35] the authors study effort and its impact on referee reliability, and in turn, on the quality and efficiency of the process. Their results emphasize the importance of homogeneity of the scientific community and equal distribution of the reviewing effort.

In this work, we will use the score range and programme commitee update defined in [20], and we will apply it to control the effect of rational cheaters as

presented in [39], adding also, partly inspired by [35], two different scenarios: homogeneous and heterogeneous conferences.

3. The Peer Review Model

In this section we define the entities involved in the peer review process, we propose a new model to reproduce its functioning and we present an agent-based implementation of this model.

3.1. *Peer review entities*

The key entities we identify within the peer review process are: the *paper*, the *scientist* and the *conference*. We define them as follows:

- The *paper* entity is the basic unit of evaluation and it refers to any item subject to evaluation through a peer review process, including papers but also, for example, project proposals. We assume that the actual value of a paper is difficult to ascertain and that it can only accessible through a procedure implying the possibility of mistakes.
- *Scientists* write papers, submit them to conferences and review papers written by others. Regarding paper creation, the value of a paper will depend on the writing skills of the authors. The submission decision must consider aspects such as the characteristics of the conference (e.g. acceptance rate), those of the authors (e.g. risk taking), etc. Scientists will also be characterized by their reviewing skills, that represent the chance they actually understand the paper they review, thus being the primary cause of reviewing noise. The evaluation process might involve other strategic behaviors possibly adopted by scientists, such as the competitor eliminating strategy used by rational cheaters in [39].
- The *conference* entity refers to any evaluation process using a peer review approach. Hence, it covers most journal or conference selection processes as well as the project evaluations conducted by funding agencies. Every *paper* submitted to a conference is evaluated by a certain number of *scientists* that are part of the programme committee (PC) of the conference. Thus, the conference is where all the process comes together and a number of questions arise. For example, since the number of evaluations a paper receives are just a few (three being a typical case): can the review-conference system ensure quality in the face of variable reviewing skills or strategic behaviors, thanks to some selection process of PC composition that leans on disagreement control? The peer review model presented below is meant to tackle this kind of questions by concretising the different issues introduced for the general entities presented above.

3.2. *Proposed model*

The proposed model represents the peer review problem by a tuple $\langle S, C, P \rangle$, where S is the set of *scientists* playing both the role of authors that write papers and

the role of reviewers that participate in the PC of a set of *conferences* C. *Papers* P produced by scientists have an associated value representing their intrinsic value, and receive a review value from each reviewer. These values are expressed as integers in an N-values ordered scale, from strong reject (value 1) to strong accept scores (value N).

Every scientist $s \in S$ is represented by a tuple of the form of Eq. 1.

$$s = \langle ap, aq, as, rs, rt, rd \rangle \tag{1}$$

Regarding paper production, each scientist has an associated author productivity ap, the number of papers uniformly written per year. Papers are of the form $p = \langle a, iv \rangle$, being $a \in S$ the author of the paper and $iv \in \{1, .., N\}$ the intrinsic value (quality) of the paper. This intrinsic value is calculated considering the author quality $aq \in \{1, .., N\}$ and the author skill value $as \in [0, 1]$. Whereas aq represents the standard author quality, as represents the production reliability of the same. Hence, scientists as authors write papers of value aq with probability as, and of random value with probability $(1 - as)$ in order to produce, occasionally, some paper with different quality with respect to their standard. Similarly, as a reviewer, each scientist has an associated reviewer skill value $rs \in [0, 1]$ as well as a reviewing type $rt \in \{\texttt{normal}, \texttt{rational}\}$. Finally, the rd value measures the risk propensity of the scientist, i.e., the inclination to send papers to conferences whose acceptance values differ from their evaluation of their own papers.

In algorithm 1 we show the pseudocode carried out by scientists to review papers. The **if** statement in line 1 models the noisy evaluation of papers, where the result of reviewing is accurate with probability rs, and completely random with probability $(1 - rs)$. Here, `Random` is a function providing a random float number in the range $[0, 1]$ whereas `RandomInt` returns a random integer in $\{1, .., N\}$. Furthermore, in line 7 we have incorporated the rational cheating strategy introduced in [39]. Hence, *rational* cheaters punish those papers whose intrinsic value is greater than his own author quality, thus trying to clear the way for his papers - preventing better papers to appear and, for example, collect more citations than one's own. It is worth mentioning that the intrinsic value of a paper is not available to *rational* cheaters (what could be seen as an un-realistic feature), but only an estimated value that depends on their skill as reviewers.

Conferences $c \in C$ are represented by a tuple of the form of Eq. 2.

$$c = \langle m, PC, rp, pr, av, I, dt, pu \rangle \tag{2}$$

Each conference is celebrated every year in a certain month m, in which it issues a call for papers. In algorithm 2 we show the pseudocode executed by scientists when deciding whether to submit a paper to a conference after having received its call for papers. Note how the noisy evaluation of papers also occurs when evaluating one's own papers in lines 2 - 6. Scientists decide whether to submit papers or

Algorithm 1 Pseudocode to review papers

Input: Paper Intrinsic Value (iv), Reviewer Skill (rs), Reviewer's Author Quality (aq), Reviewing Type (rt)
Output: Review Value for the paper ($reviewValue$)

```
1: if rs > Random() then
2:    estimatedValue ← iv
3: else
4:    estimatedValue ← RandomInt(1, N)
5: end if
6: if rt = rational then
7:    if estimatedValue < aq then
8:       reviewValue ← estimatedValue
9:    else
10:      reviewValue ← 1
11:   end if
12: else
13:   reviewValue ← estimatedValue
14: end if
```

not in accordance with their risk propensity, which is expressed through the integer parameter rd. Hence, the submission happens when the distance between the estimated paper value and the conference acceptance value av is less than or equal to rd (see line 7).

Algorithm 2 Pseudocode to submit papers

Input: Available Papers (AP), Reviewer Skill (rs), Risk Degree (rd), Conference (c), Conference Acceptance Value (av)

```
1: for all p such that p ∈ AP do
2:    if rs > Random() then
3:       estimatedValue ← iv
4:    else
5:       estimatedValue ← RandomInt(1, N)
6:    end if
7:    if |estimatedValue − av| ≤ rd then
8:       Submit(p, c)
9:    end if
10: end for
```

Conferences employ a subset of scientists $PC \subseteq S$ as their programme committee, whose size depends on the number of reviews requested per paper rp and the number of reviews done per PC member pr. Then, they accept those papers whose

average review value is greater than the acceptance value av.

Conferences also keep track of disagreements between reviewers, as they might be a signal of low reviewer skill or cheating. One disagreement event is not enough to find out which of the disagreeing parts is to blame. Thus, conferences maintain an image $i \in I$ of each scientist that has ever been a PC member, accounting for the number of disagreements with the other reviewers. Images are of the form $i = \langle s, nd, nr \rangle$, where s is the scientist, nd is the accumulated number of disagreements and nr is the total number of reviews carried out. Disagreements are calculated on a paper basis as the difference between the review value given by the reviewer and the average review value for that paper. When this difference gets higher than a disagreement threshold dt, the reviewer disagreement count grows by one. The dt parameter could also be fine-tuned for the detection of more sophisticated cheating approaches.

Reviewer images are used to update the PC by discarding the pu percentage of reviewers with the highest ratio nd/nr and selecting new ones from S. This way, conferences perform a selection process which selects reviewers who provide similar evaluations. Given our choice for reviewers' mistakes (i.e. if they don't understand the paper, the evaluation is random), this mechanism should also select good reviewers.

In algorithm 3 we show the pseudocode executed when celebrating a new edition of a conference. Firstly, function `CallForPapers` in line 3 broadcasts the conference call for papers and receives papers submitted during a fixed period of time (currently, two months). Secondly, function `UpdatePC` in line 4 adjusts the PC to the number of papers received as well as discards the $pu\%$ of reviewers with the worst image. New members for the PC are selected randomly from the set of scientist S. Thirdly, the **for** statement starting in line 5 is in charge of the evaluation process: function `AskForReviews` returns the reviews from rp reviewers, different to the author and randomly chosen from the PC, in the form of pairs $[s, rValue]$, where s is the reviewer and $rValue$ is the grade given to the paper; function `ComputeAvgReview` computes the average review value for the paper; lines 12 - 16 accept those papers over the acceptance value; and functions `GetImage` and `UpdateImage` in lines 17 - 24 retrieve and update the image of the reviewers after checking for disagreements. Finally, accept and reject notifications are sent to the authors by functions `NotifyAccepts` and `NotifyRejects`.

3.3. *Agent-based implementation*

Some general characteristics of agent-based models (ABMs) make them well suited to the modelling of the scientific process [32]. Heterogeneity states that agents can differ from one another in as many ways as the parameter range for each of their individual properties will allow. While this is something that would be very hard to track with traditional analytical models, the computer makes it possible to deal with a number of heterogeneous agents. Autonomy refers to the absence of

Algorithm 3 Pseudocode to celebrate a conference

Input: Celebration Year ($year$), Conference Acceptance Value (av), Current Programme Committee (PC), Current Scientists' Images (I), Percentage of PC update (pu), Scientists (S), Reviews Per Paper (rp), Papers Per Reviewer (pr), Disagreement Threshold (dt)

Output: New Programme Committee (PC), New Scientists' Images (I)

$AccPapers \leftarrow \phi$
$RejPapers \leftarrow \phi$
$RcvPapers \leftarrow CallForPapers(year, av)$
$PC \leftarrow UpdatePC(PC, S, I, pu, \lceil |RcvPapers| * rp/pr \rceil)$
for all p such that $p \in RcvPapers$ **do**
  $Reviews \leftarrow AskForReviews(p, rp, PC)$
  $sumOfReviews \leftarrow 0$
  for all $r = [s, rValue]$ such that $r \in Reviews$ **do**
    $sumOfReviews \leftarrow sumOfReviews + rValue$
  end for
  $avgReviewValue \leftarrow sumOfReviews/|Reviews|$
  if $avgReviewValue \geq av$ **then**
    $AccPapers \leftarrow AccPapers \cup \{[p, avgReviewValue]\}$
  else
    $RejPapers \leftarrow RejPapers \cup \{[p, avgReviewValue]\}$
  end if
  for all $r = [s, rValue]$ such that $r \in Reviews$ **do**
    $[nd, nr] \leftarrow GetImage(I, s)$
    if $|avgReviewValue - rValue| > dt$ **then**
      $I \leftarrow UpdateImage(I, s, nd + 1, nr + 1)$
    else
      $I \leftarrow UpdateImage(I, s, nd, nr + 1)$
    end if
  end for
end for
$NotifyAccepts(AccPapers)$
$NotifyRejects(RejPapers)$

central control. In the context of social simulation, this can be likened to a form of methodological individualism: while institutions (and other macro-structures) can set policies (rules, values, etc.) that will influence an agents behaviour, they are not directly coordinating the agents or moving them around. At each time step in a simulation, agents make their own decisions in order to achieve their individual goals, possibly including some sort of individual and collective agent learning and qualitative change, things that definitely happen in the peer-review world. Hence, ABMs are concerned with the micro-level processes that give rise to observable,

higher-level patterns. If an ABM can generate some macro-phenomenon of interest, then it can at least be considered a candidate explanation for it.

In this paper we apply agent-based simulation as the modelling technique [2] to represent the peer review process. With respect to statistical techniques employed for example in [12] or [24], the agent-based or individual-based approach allows us to model the process explicitly. In addition, it helps focusing on agents, their interaction, and possibly also their special roles - consider for example the proposal in [24] of increasing pre-screening of editors or editorial boards. Such a change is based on trust in the fair performance of a few individuals who take up the editors role. Thus, these individuals deserve detailed modeling, that could allow us to reason on their goals and motivations [14].

The proposed peer review model has been implemented as a MAS (Multi-Agent System) over Jason [3], which allows the definition of BDI agents using an extended version of AgentSpeak(L) [29]. As depicted in figure 1, this MAS represents both scientists and conferences as agents interacting in a common environment. The environment handles the clock system and maintains the agents' belief base. As every agent lives in its own thread, the system runs in a (simulated) continuous time. Thus, agents can concurrently react to the passage of time by triggering different plans such as that of writing new papers or celebrating a new edition of a conference. Communication between conferences and scientists take place within these celebrations: conferences broadcast their call for papers, which cause scientist to decide whether to submit their available papers; reviewers part of the PC are asked for reviews of papers; and authors are notified about the acceptance or rejection of candidate papers.

The implemented MAS is highly configurable; the number and characteristics of both conferences and scientists can be independently set, following different statistical distributions (e.g. uniform, normal, beta...). Thus, the MAS can be configured to run different simulations and evaluate the effects of the parameters in the proposed peer review model.

4. Scenarios

In this section, we present the results of a set of simulations involving 1000 scientists and 10 conferences across 50 years. Each scientist writes two papers per year ($ap = 2$), so that the overall production amounts to 2000 papers uniformly distributed over the year.

Paper intrinsic values (quality) and review values are expressed as integers in a 10-values ordered scale, from one to ten. Author qualities ($aq \in \{1, .., 10\}$) follow a (discretized) Beta distribution[b] with $\alpha = \beta = 5$. We choose this shape, a bell shaped curve with mean 5.5 and symmetrically distributed between one and ten, in

[b]The beta distribution is the obvious choice for a statistic in a fixed interval as the one we are using - the alternative being a normal distribution with cut tails, but that is just an approximation, and much less flexible, for example, in terms of central value.

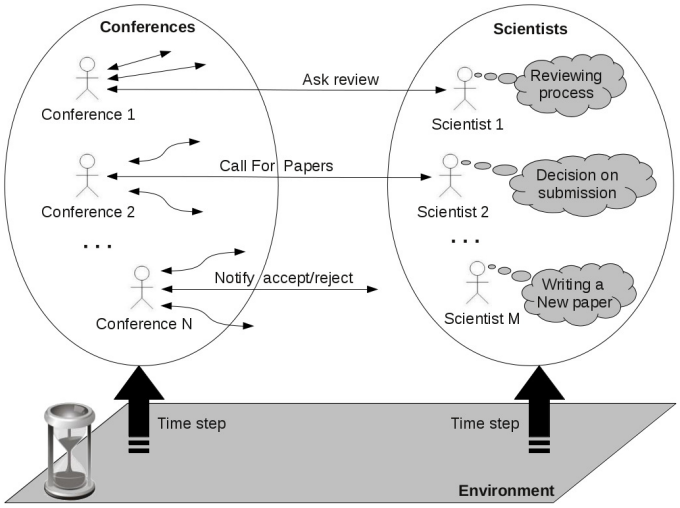

Fig. 1. Overview of the MAS implementation.

the hypothesis that average papers are more common than either excellent or bogus papers. Author skills (as) and reviewers skills (rs) follow a uniform distribution in [0.5,1], that we consider a moderate level of noise in the production and evaluation of papers. With respect to the reviewing type (rt), we show results with a mix of regular reviewers and rational cheaters; in most of the cases, up to 30% of the latter. We have performed simulations up to 90% of rational cheaters but, when those become majority, the probability of having two over three cheating reviews grows enough to turn the system upside down - PCs get filled with rational cheaters and the whole system collapses, often ending up with no papers accepted at all.

Conference parameters have been set to reproduce two different experimental scenarios that we call *homogeneous condition* and *heterogenous condition*. These scenarios are a first step to understand the emergence of quality specialization in the structure of workshops, conferences and papers. To this purpose, we compare a system without specialization with one in which conferences differ in the quality they request from a paper.

In the *homogeneous condition* (Hom) all the conferences act in the same way, as they aim to accept papers whose quality is just above the average score ($av = 5.5$). Scientists are then configured to submit papers to the first conference available after the moment of production (their risk propensity being set to ten, $rd = 10$). In the *heterogeneous condition* (Het) we have one conference for each acceptance value from 1 to 10. In this way, we distinguish high-quality from low-quality conferences. Scientists submit papers to a conference whose av differs, at most, one score from the estimated paper value ($rd = 1$). For instance, a conference with $av = 7$ would

only receive papers of estimated quality from six to eight. Conferences are scheduled along the year so as to avoid conferences of similar acceptance value to appear next to each other and reduce contention for the papers. We did this by selecting a permutation p that maximizes Eq. 3.

$$\sum_{i=1}^{10} \sum_{j=i+1}^{10} \frac{|p_i - p_j|}{min(|i-j|, |i-j+N|)} \tag{3}$$

Conferences in both the *homogeneous condition* and the *heterogenous condition* ask for three reviews per paper ($rp = 3$) and each PC member carries out a maximum number of three reviews ($pr = 3$). The disagreement threshold is set to four ($dt = 4$) and the percentage of PC members that are updated each year is ten percent ($pu = 10$).

4.1. *Results*

Our research hypothesis is that the PC update mechanism proposed will effectively find out and expel the rational cheater scientists. The argument that rational cheaters will find themselves in disagreement with others every time they act strategically makes sense and, in fact, in figure 2 we can observe how rationals decrease substantially in the conditions where they are more abundant, up to an initial value of 30%. The PC update mechanism results significantly more effective in the homogeneous condition than in the heterogeneous one (two-sided t test with p-value of 0.036 in 2050).

Note that for the homogeneous condition, averaging over conferences removes little information, while in the heterogeneous one, where conferences differ in their acceptance value, this averaging could hide information. We address heterogeneous conferences individually in section 4.2.

Let us now focus on indicators showing the effectiveness of the rational cheating strategy. The purpose of adopting a rational cheating strategy is to remove potential competition from better authors and papers. Thus, the effect of rational cheaters should be seen as an increase in the number of papers that should be accepted, but end up being rejected. We call these "good papers rejected" (GPR). The opposite, that is, the papers that should end up rejected but do not, are named as "bad papers accepted" (BPA). Note that, although the definition is the same, the details differ between the scenarios defined above. For example, a paper with quality seven that gets a rejection is automatically a GPR in the homogeneous scenario whereas, in the heterogeneous case, this depends on the acceptance value of the conference. That is, if the conference has an acceptance value of nine the rejection is due and the same paper would not count as a GPR.

Figures 3) and 4) respectively show the number of GPR and of BPA for the scenarios considered in this paper. For the simulations starting with more rational cheaters (Hom-30 and Het-30 in Figure 3), the decrease in the number of GPR,

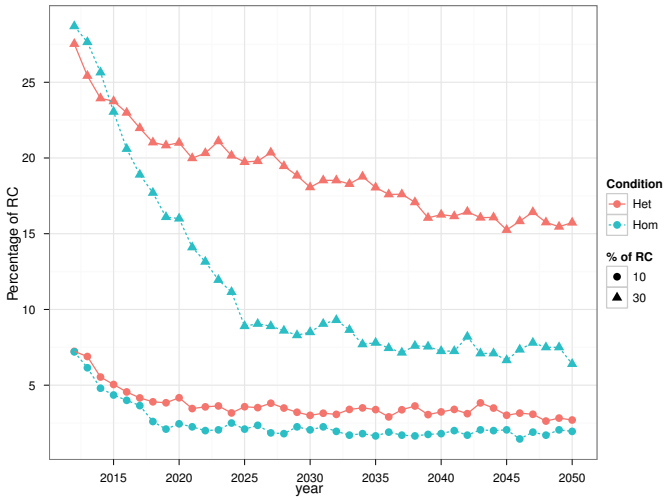

Fig. 2. Percentage of rational cheaters (RC) under homogeneous (Hom) and heterogeneous (Het) conditions with initial percentages of ten and 30%, averaged over 10 conferences. The presence of rational cheaters decreases in the first ten years, with the Hom scenario being more effective.

following the removal of rational cheaters from the PC, is already significant after a few years (p-value of 0.02 between 2011 and 2015). However, notwithstanding the very low quantity of rational cheaters at the end of the simulation (consider for example the case of Hom-30), the total number of GPRs remains rather high.

With respect to the number of bad papers accepted, they remain rather stable (see Figure 4), even though being at a lower absolute quantity with respect to the GPR. Only in the Hom-0 condition they seem to decrease in time. But what is more interesting is that the number of BPA at the onset of the simulation and during the first years is inversely proportional to the quantity of rational cheaters at the start. Thus, no rational cheaters bring more BPA than a 30% of rational cheaters, and this is true for both scenarios.

In figure 5 we show the number of accepted papers, that grows in time for the conditions with rational cheaters. As they are expelled from the PCs, the number of accepted papers grows to approach that of conditions without rational cheaters. This is likely to be happening also because of the reduction in the GPR (i.e. less good papers rejected means more papers accepted).

What about quality? Is the removal of rational cheaters from the programme committees going to make a difference in the quality of accepted papers? Surprisingly, in figure 6, we can see that the removal of rational cheaters does not contribute to higher average quality of papers. Only the Hom-30 condition shows an initial increase in quality (two-sided t-test between 2011 and 2025 gives a p-value of 0.003).

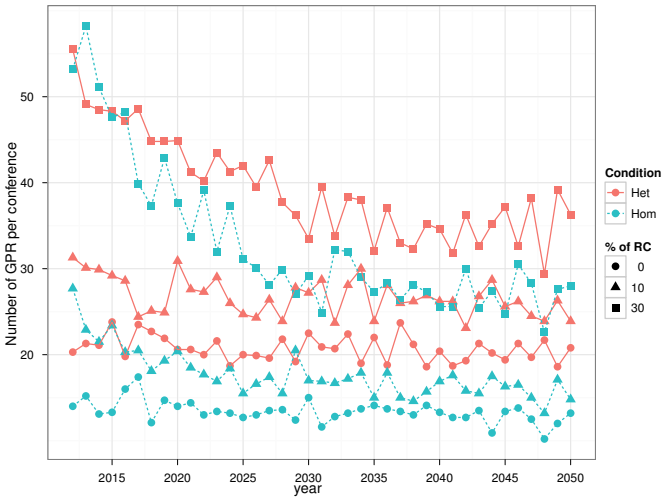

Fig. 3. Number of Good Papers Rejected (GPR) for the homogeneous and the heterogeneous scenarios, with initial percentages of rational cheaters from none to 30%. GPRs decrease significantly for both conditions with 30% of rational cheaters. Compare with the removal of rational cheaters from the PC in figure 2.

4.2. *Looking at heterogeneous conferences*

We now open up the box of heterogeneous conferences to see how they contribute to the averages shown previously. From Figure 7 (left column), where we show the percentage of rational cheaters for each individual conference (characterised by an acceptance value), we see immediately how the PC update mechanism fails in moving rational cheaters away from the PC when the quality of the conference is low. If the acceptance value reaches 4 or lower, there is no decrease at all. This happens due to the paper quality being too near to the lowest possible value used by rational cheaters to prevent publication of competitive papers. Consider, for example, a rational cheater with author quality six. Within a conference of quality eight, it will act as a rational in all cases. But if that same agent ends in a PC for a conference with acceptance value four, it will never act as a rational because rationals give fair reviews to papers under their author quality. Thus, that conference feels no need to drive it away from the PC.

This is also reflected in the quantity of good papers rejected (see the right column of Figure 7). While low-level conferences reject very few papers, better conferences let more GPR to slip away. Though, in these higher-quality conferences, there is a decreasing trend in this kind of mistakes, slower for the acceptance values from four to six, and faster for the better ones.

Finally, we examine the number of accepted papers per conference. As it was foreseeable, more papers are accepted by mid-quality conferences, simply because

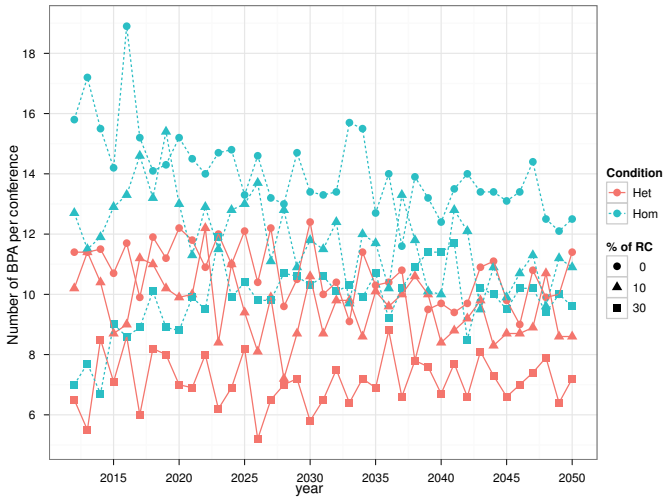

Fig. 4. Number of Bad Papers Accepted (BPA) for the homogeneous and the heterogeneous scenarios, with initial percentages of rational cheaters from none to 30%. The number of BPA is inversely proportional to the rate of rational cheaters at the start (for example, the difference between Hom-30 and Hom-0 in 2011 is significant with p-value of $1 * 10^{-5}$).

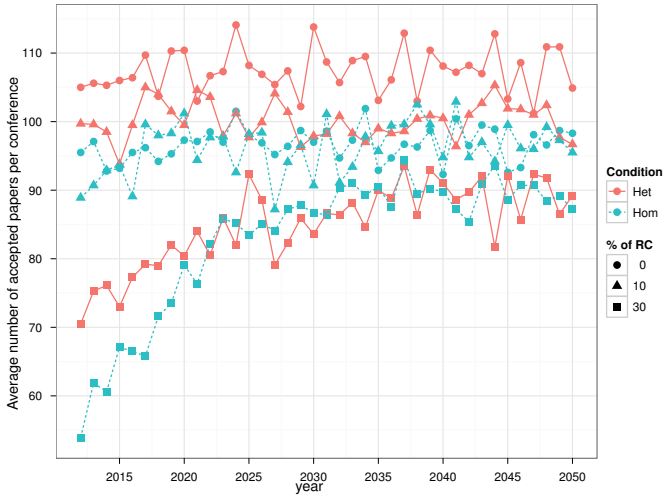

Fig. 5. Number of Accepted Papers for the homogeneous and the heterogeneous scenarios, with initial percentages of rational cheaters from none to 30%, averaged over ten runs. Conferences in the heterogeneous scenario systematically accept more papers than in the homogeneous one.

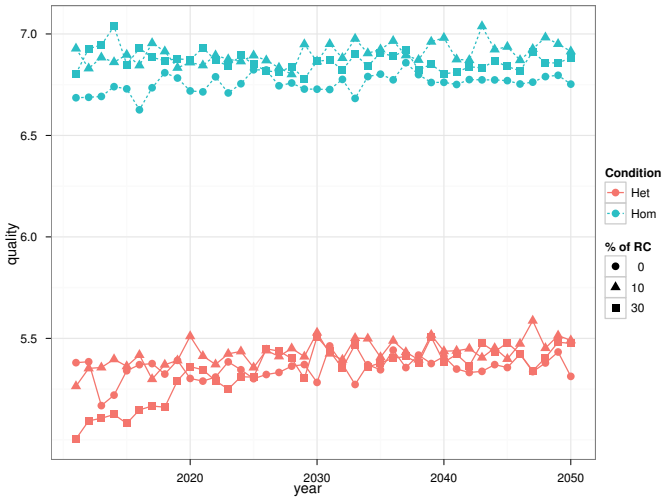

Fig. 6. Average paper quality for the homogeneous and the heterogeneous scenarios, with initial percentages or rational cheaters from none to 30%. The quality remains constant notwithstanding the removal of rational cheaters (as seen in 2). Only the Hom-30 condition shows an initial increase in quality (two-sided t-test between 2011 and 2025 gives a p-value of 0.003).

our distribution of quality is chosen so that more papers of this kind are available. The interesting part of figure 8 is the increasing trend that is distinguishable for conferences with acceptance value greater or equal to five. The cause here, in accordance with the ratio of rational cheaters seen in figure 7, is the improvement of PC quality thanks to the removal of rational scientist, that increases the number of papers accepted, mainly through the decrease of unfair good papers rejected.

4.3. *A step towards empirical validation*

How much do reviewers disagree in the real world? In our idealized model, a relatively long time span, in the order of ten to twenty years (see Fig. 2, is needed to find out rational cheaters and to drive them off PCs[c]. To perform a meaningful validation of our proposed mechanism, we would need data of comparable length for a large enough number of conferences.

Regrettably, obtaining any kind of systematic peer review data has been a hurdle, not only for the authors of this paper, but even for financed EU projects in the field[d]. We aimed for much less - that is, a qualitative validation obtained from conferences that have been made accessible through personal connections of the au-

[c]It should be considered that one of our "years" is just an instantiation of all conferences. In a more active field, the cycle could be as short as one real time month

[d]Jordi Sabater, LiquidPub project, personal communication.

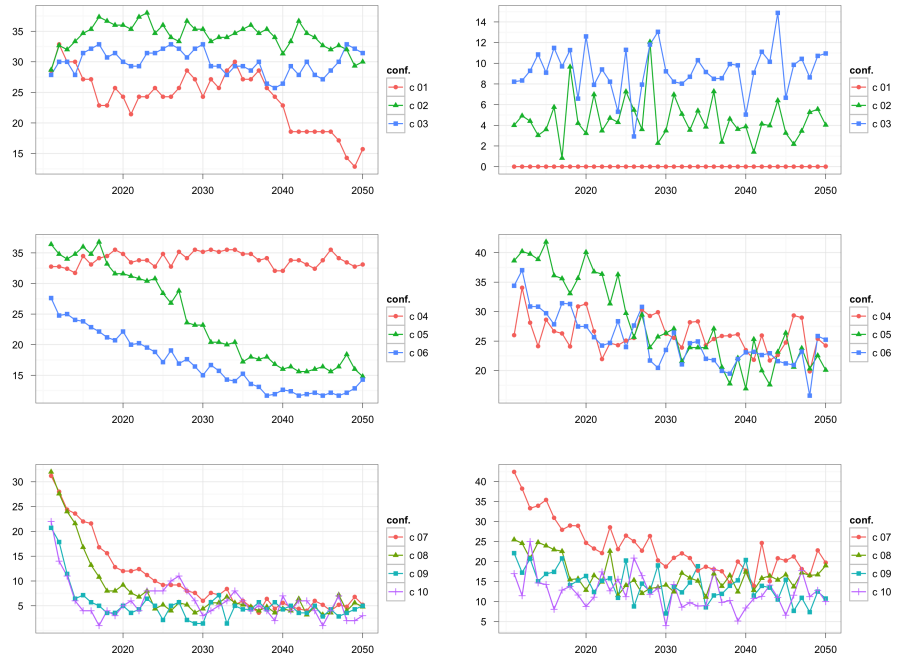

Fig. 7. Left: Percentage of rational cheaters in time, condition (Het-30) with ten conferences with acceptance values from one (c01) to ten (c10). Conferences with higher acceptance values push rational cheaters away faster. Right: Number of Good Papers Rejected (GPR) in time, condition (Het-30) with ten conferences with acceptance values from one (c01) to ten (c10). Mid-quality and high-quality conferences reduce GPR as they push rational cheaters away.

thors, whose names we hide for privacy reasons. While this set can not be considered statistically representative, it adds realism to our work.

They amount to 13 small conferences or workshops for a total of 308 papers. Disagreements, for the observed data, have been calculated rescaling the disagreement threshold $dt = 4$, used in the simulation on a scale of $N = 10$ values, to the scale of (in most cases, seven) values used by the conference (i.e. from strong reject to strong accept). The number of disagreements per paper, over the whole set, is about 0.11.

More in detail, in Table 1 we present a comparison between the amount of disagreements per paper in actual conferences and in our simulated conferences. While some of the conferences (those ranked C and B in the Computer Research & Education (CORE) conference ranking available at http://core.edu.au/) seem to occur in a high-agreement phase, other less prestigious ones show a disagreement rate between 5% and 18%. For the first set of conferences, we could hypothesize that the PC had been grown through processes like the one we have modelled, so that the initial rate of rational cheaters has been eliminated. Of course, we can not

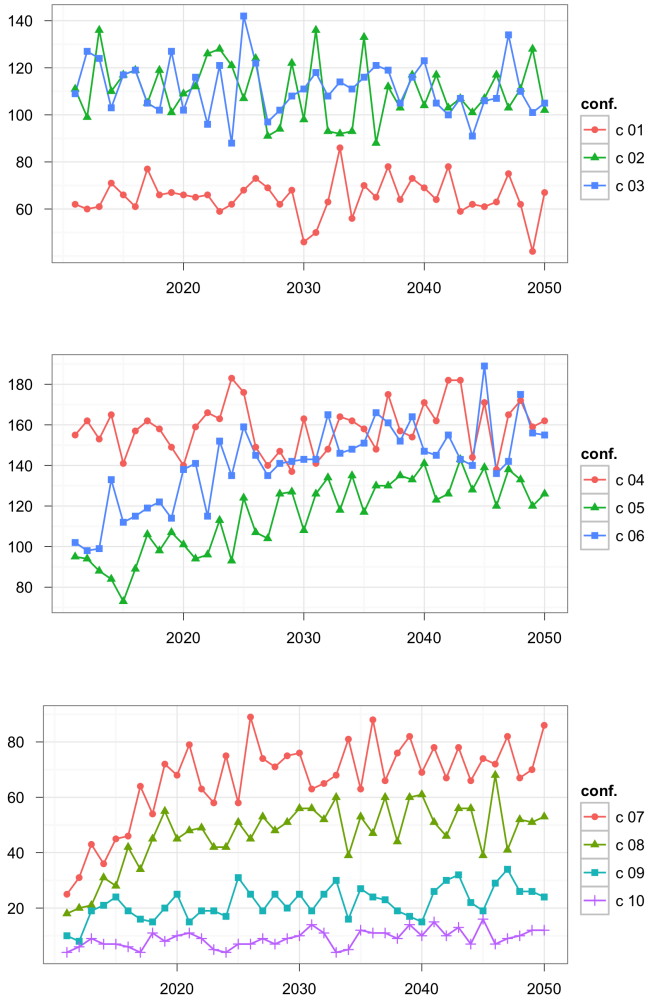

Fig. 8. Number of accepted papers in time, condition (Het-30) with ten conferences with acceptance values from one (c01) to ten (c10). Conferences with acceptance values over five increase the number of papers accepted as a result of the expulsion of rational cheaters.

discern between this case and some consensus obtained through other mechanisms (for example, lenient evaluations, as it has been shown in some reputation systems).

The second set of conferences (i.e. International, Summer School and National) shows a disagreement ratio that can be placed neatly between the values produced by our simulation. For space reasons, we only point out how the rate of disagreement for summer schools is comparable to the initial simulated values for a 30%

Table 1. Comparison of the percentage of disagreement found in real conferences and that resulting from running the simulation scenarios defined in section 4

Conference	% Initial disagreements	% Final disagreements	% Disagreement reduction
National	18.25	-	-
Summer School	10.71	-	-
International	5.41	-	-
Intl. Core C	5.0	-	-
Intl. Core B	0.0	-	-
Hom-0%RC	4.3	2.9	32.6
Hom-10%RC	6.1	4.5	26.2
Hom-30%RC	11.9	5.6	52.9
Het-0%RC-LQ	4.7	3.6	23.4
Het-0%RC-MQ	3.4	1.7	50.0
Het-0%RC-HQ	4.2	3.8	9.5
Het-10%RC-LQ	9.4	4.2	55.3
Het-10%RC-MQ	8.6	5.5	36.1
Het-10%RC-HQ	5.2	2.4	53.9
Het-30%RC-LQ	46.0	11.8	74.4
Het-30%RC-MQ	16.0	6.1	61.9
Het-30%RC-HQ	3.9	2.8	28.2

Note: Hom-10%RC stands for homogenous condition with an initial 10% of rational cheaters in the PC. Het-10%RC-LQ stands for Heterogeneous condition with an initial 10% of rational cheaters in the PC. Low-quality (LQ) conferences are those with acceptance values below four, mid-quality (MQ) conferences have an acceptance value between four and six and high-quality (HQ) conferences have acceptance values greater than six.

of rational cheaters, while that of international conferences compares to the results obtained for 10% of rational cheaters, or to those with 30% of rational cheaters after the application of the reducing mechanism (that is, the percentage of final disagreements). These results show how our model fairly reproduces the amount of disagreements per paper found in actual conferences even if the simulation is completely agnostic of the level of disagreement present in the list of reviews.

Finally, in Table 1 we also show how the proposed programme commitee update mechanism is able to reduce the number of disagreements of the simulated conferences in the course of the years. Reduction of disagreements is substantial in the homogeneous case, ranging from 26.2% to 52.9%. Figure 9 gives a general idea of what is happening in the simulation of this scenario. As the number of rational cheaters in the PC decreases during the first ten years (see also Figure 2) so does the total number of disagreements. It is worth mentioning that this amount is reduced even though there are no cheaters in the PC (i.e. 0% of RC), since the proposed programme commitee update mechanism also expels scientists with low reviewer

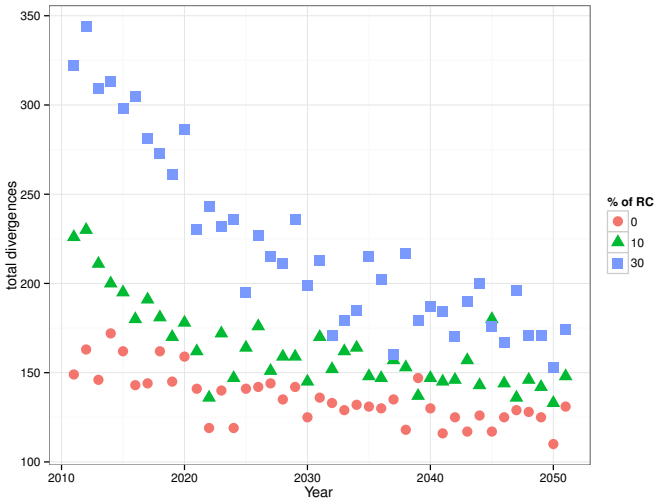

Fig. 9. Total number of disagreements under the homogeneous condition with initial percentages of ten and 30%, averaged over 10 conferences. The number of disagreements highly decreases in the first ten years.

skills, thus selecting the best candidates. Under the heterogeneous condition, this disagreement reduction ranges from a marginal 9.5% for the high quality branch of heterogeneous conferences (since there are no cheaters in Het-0%RC-HQ, this percentage is again due to the selection of scientists with a higher reviewer skill), to the substantial 74% of the low quality branch of heterogeneous conferences starting with a 30% of rational cheaters; thus showing how the simple PC update mechanism proposed could be used to reduce the number of disagreements found in some real conferences.

5. Conclusions and future work

This work highlights the importance of adopting more transparent and adaptive policies for conference programme committees. Whereas PC formation is currently more influenced by issues such as path dependency, inertia or self-selection, the application of objective and independent criteria may be beneficial to the quality of science.

Our results show how the mechanism introduced to control disagreement in the PCs is also effective in removing most of the rational cheaters from the process. The benefits can be measured in terms of the growing number of accepted papers and of the decrease in the number of mistakes (good papers rejected).

When the quality of the conferences is homogeneous, rational cheaters are reduced but at the expenses of the number of accepted papers. It is important to note that neither the homogeneity nor the heterogeneity of conferences determined the

sharp transition to random selection shown in [39]. We hypothesise that this is due to the fact that our model is based on a larger score range and three, instead than two, reviewers.

A next step in this research would be to ground our model against data extracted from higher-quality conferences as well as journals with an impact factor. However, this data has proven surprisingly difficult to obtain. Not only our queries to the owners of those systems went unanswered, but we knew that other researchers had the same situation (none of [35, 39] managed to ground their assumption either). The difference between the immediate availability of publication and citation data is especially striking.

Acknowledgments

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n 315874, GLODERS Project; it has been supported by the Spanish MICINN, Consolider Programme and Plan E funds, European Commission FEDER funds and Universitat de València funds, under Grants CSD2006-00046, TIN2009-14475-C04-04 and UV-INV-AE11-40990.62, as well as under the FuturICT coordination action. We gratefully acknowledge the supportive contribution of the anonymous reviewers. The presence of the section on validation is due to their requests. For validation, we are also very grateful towards the chairs of conferences that allowed us to use their data.

References

[1] Allesina, S., Modeling peer review: an agent-based approach, *Ideas in Ecology and Evolution* **5** (2012).

[2] Bonabeau, E., Agent-based modeling: methods and techniques for simulating human systems., *Proceedings of the National Academy of Sciences of the United States of America* **99** (2002) 7280–7287.

[3] Bordini, R. H., Hübner, J. F., and Wooldridge, M., *Programming multi-agent systems in AgentSpeak using Jason* (John Wiley & Sons, 2007).

[4] Bornmann, L., *Annual Review of Information Science and Technology* (2011) 199–245.

[5] Bornmann, L., Nast, I., and Daniel, H.-D., Do editors and referees look for signs of scientific misconduct when reviewing manuscripts? A quantitative content analysis of studies that examined review criteria and reasons for accepting and rejecting manuscripts for publication, *Scientometrics* **77** (2008) 415–432.

[6] Brabazon, T., The google effect: Googling, blogging, wikis and the flattening of expertise, *Libri* **56** (2006) 157–167.

[7] Bruckner, E., Ebeling, W., and Scharnhorst, A., The application of evolution models in scientometrics, *Scientometrics* **18** (1990) 21–41.

[8] Buecheler, T., Sieg, J. H., Füchslin, R. M., and Pfeifer, R., *Crowdsourcing, Open Innovation and Collective Intelligence in the Scientific Method: A Research Agenda and Operational Framework* (MIT Press, Cambridge, Mass, 2011), pp. 679–686.

[9] Brner, K., *Atlas of Science: Visualizing What We Know* (MIT Press, Cambridge, Mass., 2010).
[10] Brner, K. and Scharnhorst, A., Visual conceptualizations and models of science, *Journal of Informetrics* **3** (2009) 161 – 172.
[11] Callahan, D., Rational Cheating: Everyone's Doing It, *Journal of Forensic Accounting* (2004) 575+.
[12] Casati, F., Marchese, M., Ragone, A., and Turrini, M., Is peer review any good? A quantitative analysis of peer review, Technical report, Ingegneria e Scienza dell'Informazione, University of Trento (2009), `http://eprints.biblio.unitn.it/archive/00001654/`.
[13] Cole, S., Cole, J. R., and Simon, G. A., Chance and consensus in peer review., *Science* **214** (1981) 881–886.
[14] Conte, R. and Castelfranchi, C., *Cognitive Social Action* (London: UCL Press, 1995).
[15] Conte, R., Paolucci, M., and Sabater Mir, J., Reputation for Innovating Social Networks, *Advances in Complex Systems* **11(2)** (2008) 303–320.
[16] Egghe, L. and Rousseau, R., *Introduction to Informetrics* (1990).
[17] Frishauf, P., Reputation Systems: A New Vision for Publishing and Peer Review, *Journal of Participatory Medicine* **1** (2009) e13a+.
[18] Gilbert, N., A simulation of the structure of academic science, *Sociological Research* **2** (1997) 1–25.
[19] Goffman, W., Mathematical approach to the spread of scientific ideas - the history of mast cell research, *Nature* **212** (1966) 449–452.
[20] Grimaldo Moreno, F., Paolucci, M., and Conte, R., A Proposal for Agent Simulation of Peer Review, *Social Science Research Network Working Paper Series* (2010).
[21] Helbing, D., Pluralistic Modeling of Complex Systems, *Science and Culture* **76** (2010) 315–329.
[22] Jayasinghe, U. W., Marsh, H. W., and Bond, N., A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings, *Journal of the Royal Statistical Society - Series A - Statistics in Society* **166** (2003) 279–300.
[23] Lotka, A. J., The frequency distribution of scientific productivity, *J Washington Acad Sci* **16** (1926) 317–324.
[24] Neff, B. D. and Olden, J. D., Is Peer Review a Game of Chance?, *BioScience* **56** (2006) 333–340.
[25] Newman, M. E. J., Complex Systems: A Survey, *American Journal of Physics* **79** (2011) 800–810.
[26] Pariser, E., *The Filter Bubble: What the Internet Is Hiding from You* (Penguin Press HC, The, 2011).
[27] Payette, N., For an integrated approach to agent-based modeling of science, *Journal of Artificial Societies and Social Simulation* **14** (2011) 9.
[28] Ragone, A., Mirylenka, K., Casati, F., and Marchese, M., A quantitative analysis of peer review (2011).
[29] Rao, A. S., AgentSpeak(L): BDI agents speak out in a logical computable language, in *Proc. of MAAMAW'96*, ed. Verlag, S., number 1038 in LNAI (1996), pp. 42–55.
[30] Rodriguez, M. A., Bollen, J., and Van de Sompel, H., A practical ontology for the large-scale modeling of scholarly artifacts and their usage, in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, JCDL '07 (ACM, New York, NY, USA, 2007), ISBN 978-1-59593-644-8, pp. 278–287, doi:10.1145/1255175.1255229, `http://doi.acm.org/10.1145/1255175.1255229`.
[31] Roebber, P. J. and Schultz, D. M., Peer Review, Program Officers and Science Fund-

ing, *PLoS ONE* **6** (2011) e18680+.

[32] Scharnhorst, A., Börner, K., and Besselaar, P. v. d. (eds.), *Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences* (Springer, Berlin, 2012).

[33] Schultz, D. M., Are three heads better than two? how the number of reviewers and editor behavior affect the rejection rate, *Scientometrics* **84** (2010) 277–292.

[34] Smith, R., Peer review: a flawed process at the heart of science and journals, *JRSM* **99** (2006) 178–182.

[35] Squazzoni, F. and Gandelli, C., Saint Matthew strikes again: An agent-based model of peer review and the scientific community structure, *Journal of Informetrics* **6** (2012) 265–275.

[36] Sterman, J. D., The growth of knowledge: Testing a theory of scientific revolutions with a formal model, *Technological Forecasting and Social Change* **28** (1985) 93 – 122.

[37] Sun, R. and Naveh, I., Cognitive simulation of academic science, in *International Joint Conference on Neural Networks, IJCNN 2009, Atlanta, Georgia, USA, 14-19 June 2009* (IEEE, 2009), pp. 3011–3017, doi:http://dx.doi.org/10.1109/IJCNN.2009.5178638.

[38] Surowiecki, J., *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations* (Doubleday, 2004).

[39] Thurner, S. and Hanel, R., Peer-review in a world with rational scientists: Toward selection of the average, *European Physical Journal B-Condensed Matter* **84** (2011) 707.