IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 10, 2021, accepted July 28, 2021. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2021.3103837

Computer Simulations of Scientific Peer Reviewing

SYLVAIN HALLÉ, (Senior Member, IEEE)
Department of Computer Science and Mathematics, Université du Québec à Chicoutimi, Saguenay, QC G7H 2B1, Canada
e-mail: shalle@acm.org

ABSTRACT A simple mathematical model of the scientific peer reviewing process is developed. Papers and reviewers are modeled as numerical vectors, respectively representing the paper's value among multiple quality dimensions, and the importance given to these dimensions by a given reviewer. Computer simulations show that the model can reproduce various characteristics of a real-world paper decision process, and in particular its propensity to act as an ''arbitrary'' decision procedure for a range of submissions. A key finding of this study is that the appearance of randomness can be explained by a mismatch between high quality dimensions of a paper, and those valued by the reviewers it is assigned to. As a consequence, a program committee may exhibit arbitrariness even with a set of completely reliable reviewers. Various factors contributing to this arbitrariness are then examined, and alternate selection models are studied that could help reduce arbitrariness and reviewer effort.

INDEX TERMS Computer simulation, peer reviewing.

I. INTRODUCTION

Peer reviewing is the cornerstone of research publication, and the favored means by which scientific output is evaluated and curated. An important part of every researcher's time is either spent writing and submitting papers to peer review, or standing on the other side of the fence and reviewing papers submitted by others. Putting a new scientific result under the scrutiny of a number of knowledgeable experts of a field is widely believed to ensure the quality and soundness of the manuscripts that are deemed suitable for publication [26]. Even when papers are rejected by the members of a program committee, the feedback coming from peer review can also often result in improvements over the original submissions, that may then be re-submitted to another peer reviewing trial. To paraphrase Richard Feynman, peer reviewing is a system that has been put in place so that we, as scientists, do not fool ourselves.

Yet, peer reviewing is not without its critics. Its slow turnaround time, especially for journals, has been criticized as being an impediment to the quick dissemination of important results [3]. Peer reviewing has also been criticized for a perceived lack of reliability and fairness [4]. The anonymity of reviewers can also lead to abuse; some scientific communities, such as Computer Science, have already pointed out that ''nasty reviewing'' is more common than one would like [19], and that overall, some members of a program committee may resort to adversarial tactics to reject submissions on dubious grounds [8]. Even without these factors, scientific publication is an unforgiving activity, where many journals and conferences boast acceptance rates lower than 20%, and where facing repeated rejection is the daily bread of most authors.

Given such a harsh process, it is not surprising that many authors sometimes feel they are unfairly deprived of an opportunity to publish; after all, the rejection messages they often receive candidly admit that ''many good submissions had to be rejected''. This leads one to ponder to what extent decisions on papers are, as John Langford mentioned in a post on the subject [17], *arbitrary* —or, as some authors even suggested, a ''crapshoot'' [7]. This state of things is stoically tolerated by most authors, due to the persistent belief that high rejection is a sign of high quality. After all, opinions to the effect that acceptance rates should be *increased* [9], although they might have some traction, still represent a small minority.

The field of Computer Science in particular does not display much introspection regarding the way its members evaluate each other's work. Very few publication venues examine their reviewing process in any rigorous way, and attempt to measure whether the decisions they take are fair, systematic and reproducible. The single exception we know

The associate editor coordinating the review of this manuscript and approving it for publication was Yassine Maleh.

of is the Conference on Neural Information Processing Systems (NIPS), which performed an experiment that measured variability in reviewer decisions by submitting a sample of its submissions to two different sets of reviewers [18]. The somewhat cheerless results (the two sets agreed on only 22% of accepted papers) give credence to the arbitrariness hypothesis, which in turn reveals the need for a principled study of the peer reviewing process.

This paper makes its contribution to the question by describing a mathematical model of peer reviewing. It starts from the simple principle that a paper is evaluated along a number of numerical dimensions that are valued unevenly by a set of randomly-picked reviewers, which is explained in Section II. From this, the action of a reviewer on a paper can be abstracted into a function involving a random variable that is given more or less weight in the reviewer's appreciation, as will be shown in Section III. This simple principle is sufficient to exhibit arbitrariness in a committee's decisions, whose consequences will be discussed in Section IV. In particular, given appropriate parameters, the model can reproduce various characteristics of arbitrariness that have been observed in reality, such as the aforementioned NIPS experiment. Through computer simulations, Section V studies the impact of various factors on the tendency of a program committee to act as a random variable, and Section VI explores the relative merits of alternate ways of selecting papers. Section VIII situates this work with respect to existing literature on the study of peer reviewing, and Section VII discusses the limitations of the approach. Finally, Section IX concludes with a few additional remarks and suggestions for further study.

II. PAPERS AND REVIEWERS

The high-level reviewing process we consider is illustrated in Figure 1. A paper is dispatched to a number of reviewers, who first perform an evaluation of the paper resulting in an appreciation of the submission. This appreciation is typically turned in the final review as a position on a discrete scoring scale. The scores provided by each reviewer of the paper are then aggregated into a value reflecting its overall appreciation. Finally, the values of all papers are collected, and a decision (typically accept or reject) is then issued for each paper, based on its value relative to the value of other papers evaluated during the same process. In this section and the next one, we define a simple mathematical model reflecting this flowchart.

Let $\mathcal{C} = [-1, 1]$ be a continuous real-valued scale. A research paper is a vector $\vec{p}$ in the d-dimensional hypercube $[-1, 1]^d$. Each component of the vector represents an aspect of the paper, and the number at the corresponding position indicates the intrinsic value of the paper along this aspect. Note that high and low values of the scale do not necessarily translate as ''good'' and ''bad''. For example, one dimension could determine whether a paper is theoretical $(+1)$ or applied (-1), with each end of the scale not being preferable *in itself*.

A reviewer r is a pair $(\vec{v}, \gamma)$, where $\vec{v}$ is another d-dimensional vector in $[-1, 1]^d$, and $\gamma \in [-1, 1]$ is a numerical constant. Each component of $\vec{v}$ corresponds to the same quality aspects as for a paper, and each value represents the importance or ''weight'' this reviewer gives to this aspect when evaluating a paper. Contrary to a paper, in a reviewer's vector, $+1$ means good and -1 means bad. That is, when a component has a high value (i.e. close to 1), it indicates that the reviewer gives high importance to this aspect. However, when a component has a low value (i.e. close to -1), it indicates that the reviewer has a strong negative opinion of papers that score high on this aspect. For example, on the applied/theoretical scale given as an example above, a reviewer that prefers applied papers and strongly dislikes theoretical papers would have a value close to -1 to represent this fact. A value close to 0 for a component indicates that the reviewer gives no importance to this aspect in a paper.

The appreciation of a reviewer for a paper is defined as:

$$\alpha_r(\vec{p}) = \tau\left(\frac{1}{d}\vec{p}\cdot\vec{v} + \gamma\right)$$

The dot product $\vec{p}\cdot\vec{v}$ is simply the sum of each paper's components, weighted by the importance the reviewer gives to this component. This product is normalized so that it lies in the interval $[-1, 1]$. Again, high values indicate that the reviewer has a strong positive appreciation of the paper, and the opposite for low values. This is illustrated in Figure 2. From this, the *match* between a paper and a reviewer can be quantified as the angle θ between their respective vectors; it is expressed in radians, and can be computed by:

$$\theta = \arccos\left(\frac{\vec{p}\cdot\vec{v}}{|\vec{p}||\vec{v}|}\right)$$

The constant γ, specific to each reviewer, shifts the original score by some amount, either towards a positive appreciation (when $\gamma > 0$), or towards a negative one (when $\gamma < 0$). It is present to model the fact that reviewers may have a globally more positive or more negative attitude towards the papers they review (but the same for all papers). We call this constant the reviewer's *grumpiness*. Since the presence of this constant may shift the original score outside of the interval $[-1, 1]$, we apply the function τ, defined as $\tau(x) \triangleq \min(\max(x, -1), 1)$, which truncates any extreme values back into the bounds.

In addition to grumpiness, the modulus of $\vec{v}$ is called the reviewer's *loudness*. Intuitively, a ''loud'' reviewer gives scores across a wider range, which indicates stronger positive or negative appreciations of papers. In contrast, a quiet reviewer is such that $|\vec{v}|$ is close to zero, and has a more or less equal (and neutral) view of each submission.

It is important to observe that, for a reviewer to give a meaningful score to a paper, the two must have non-zero values at matching positions in their respective vectors; when this is the case, we say that a paper *matches* a reviewer's appreciation vector. Papers that are a poor match have a dot product close to zero, meaning that the reviewer makes almost

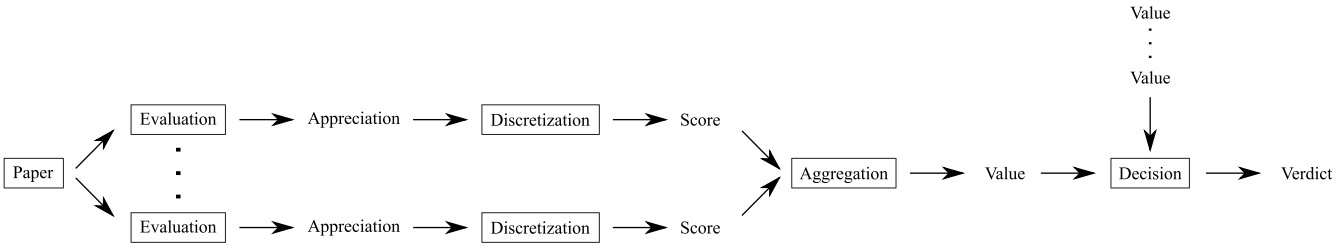


FIGURE 1. An overview of the paper reviewing process.

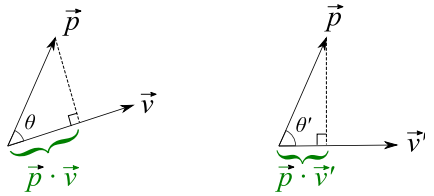


FIGURE 2. The assessment of a paper by a reviewer is modeled as the dot product of their respective quality vectors $\vec{p}$ and $\vec{v}$ (left). A different reviewer vector $\vec{v}'$ results in a different assessment (right).

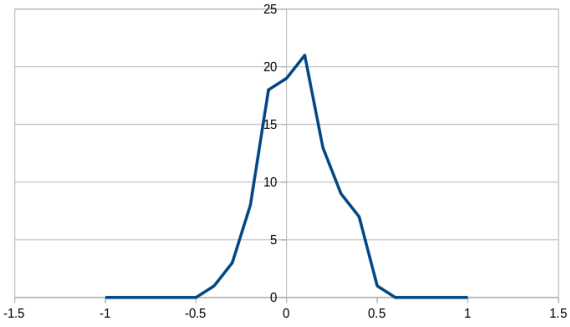


FIGURE 3. A plot of x vs. $P(\vec{v} \cdot \vec{p}) = dx$. The probability is expressed in percentage.

no distinction between any of them. When viewing papers and reviewers as vectors, this corresponds to the situation where $\vec{v}$ and $\vec{p}$ are orthogonal: what the paper *is* about is completely different from what the reviewer *cares* about.

To illustrate the importance of a good match, suppose that both $\vec{v}$ and $\vec{p}$ are uniformly chosen at random in $[-1, 1]^d$. Figure 3 shows the probability of obtaining a given dot product of these two vectors. One can see (and also demonstrate) that this product behaves like a random variable that follows a distribution centered on 0. In other words, if no guarantee can be made on the match between papers and reviewers, the resulting appreciations end up being randomly selected scores. Note that this model does not suppose that reviewers are "unreliable" or "inconsistent" in any way: each of them, taken separately, evaluates all papers in a completely deterministic fashion and ranks papers consistently. Randomness only appears through the odds that a paper be assigned to a reviewer with a matching vector.

III. PROGRAM COMMITTEES

A program committee (PC) is a pool of reviewers that has the task of choosing papers to be published from a set of submissions. In our model, for each paper, the PC picks N reviewers and asks for their respective appreciations of the paper, noted $a_1, \ldots, a_N$. Each appreciation is then discretized by applying the function δ, which turns each value in $[-1, 1]$ into a discrete value in the set $\{-b, -b + 1, \ldots, 0, 1, \ldots, b\}$. We hence make the distinction between the reviewer's real-valued *appreciation*, and the resulting discrete *score*. The latter corresponds to the input that is asked from reviewers in most PCs, and that is typically labeled with names such as *strong accept*, *weak accept*, *weak reject*, and so on. A possible discretization is to split the interval $[-1, 1]$ into $2b$ or $2b + 1$ bins of equal width (depending on whether the discrete scale allows a "neutral" score); this is the function we shall retain in the remainder of this paper.

Once the discrete scores $a'_1, \ldots, a'_N$ of each reviewer are obtained, the program committee computes their normalized average, i.e. $\overline{a} = \sum a'_i / bN$. Note that this value again lies in the interval $[-1, 1]$. The PC then decides whether a paper is accepted or rejected by checking if the average is above a fixed and predefined threshold t.

Reviewers in the PC may have arbitrary appreciation vectors. However, we can split the dimensions of a paper into two sets: those that all reviewers mostly agree on, and the others where they disagree. From this, we can extract a simplified representation of each reviewer's appreciation. Let $\vec{v}$ be a vector made of all the $d' \leq d$ dimensions where the reviewers in the PC give similar weights. Each paper can then be assigned a number called its *quality*, computed as $q(\vec{p}) = \frac{1}{d'} \vec{v}' \cdot \vec{p}'$, where $\vec{p}'$ is the projection of $\vec{p}$ on the d' agreed-upon dimensions. Quality is a value between -1 and 1, and represents a synthesis of the dimensions of the paper that make consensus among all reviewers: each paper where q is high is seen as good by all reviewers, and each paper where q is low is seen as bad by all reviewers.

We know from an earlier observation that for the remaining dimensions, where weights for each reviewer vary, the appreciation of a paper behaves like a random variable (let us arbitrarily assume it follows a normal distribution). Therefore, a reviewer's appreciation of a paper can be approximated as a function of its quality:

$$\hat{\alpha}_r(\vec{p}) = \tau\left(\lambda(\xi q(\vec{p}) + (1 - \xi)\mathcal{N}(0, \sigma)) + \gamma\right)$$

The evaluation is split into three components. The first term is the part of the appreciation that is based on the paper's quality; the second term is the part that behaves like a normal distribution with standard deviation σ; the third term is the

reviewer's grumpiness. Parameter $\xi \in [0, 1]$ indicates the fraction of the appreciation that correlates with quality — which, in the context, could be understood as the number of dimensions that reviewers agree on; a higher value of ξ indicates a higher correlation between a reviewer's appreciation and the paper's quality. Those two terms are scaled by a factor $\lambda \in (0, 1]$, which symbolizes the reviewer's loudness. The reviewer's grumpiness is then added, and the whole expression is again limited to the interval $[-1, 1]$ by function τ. We shall stress again that even though $\hat{\alpha}_r$ involves a random variable, it does not necessarily imply that reviewers themselves are random. Rather, this is used to represent disagreement between individually consistent reviewers on papers of the same quality.

A feature of this model is that it flattens multi-dimensional objects (papers and reviewers) into functions of a single dimension, the paper's quality. Equipped with such a representation, we can now explore the behavior of various program committees by varying some of the parameters. Consider the evaluation of a paper as a trial, where N PC members are picked at random, and a decision is made on the paper. Repeating this trial for the same paper, we can compute its success rate, which corresponds to the fraction of times the paper is submitted and accepted by the PC. This simple model has been implemented as a computer program, which makes it possible to simulate a reviewing process by generating a large number of "fake" papers and reviews, and studying the behavior of the resulting system according to various combinations of parameters. The source code of all simulations in this paper is available online [14] in the form of a LabPal experimental package [13]. A simulated pool of papers of uniformly distributed quality is given to a program committee of 100 members with randomly selected parameters. The process of assigning papers to reviewers and making an accept/reject decision on each is done 1,000 times.

First, let us model the "perfect" PC, where all reviewers have zero grumpiness, and all agree on all the dimensions of a paper –in other words, a paper's quality encompasses all its dimensions, and appreciations are entirely based on quality (i.e. $\xi = 1$). We can plot the success rate of a paper as a function of its quality, which results in the purple line in Figure 4. Unsurprisingly, this PC acts as a discrete quality gate: all papers below a certain quality threshold q_t are rejected all the time, and all papers whose quality lies above the threshold are accepted all the time. We argue that this is the behavior any real-world PC should tend towards, for reasons that will be discussed later.

However, the presence of disagreement, materialized by a value of ξ lower than 1, has for effect of turning this discrete gate into a continuous function, as is shown in the green line of Figure 4. In this model, papers with very high quality are still almost always accepted, and papers with very low quality are still almost always rejected. However, there exists a middle zone where papers are sometimes accepted, sometimes rejected by the PC, depending on the reviewer assignment they are given. This set of papers is somewhat

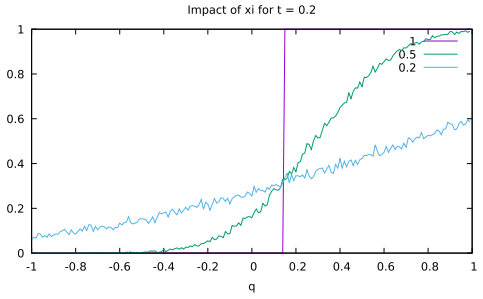


FIGURE 4. A plot of success rate in function of quality, for various values of parameter ξ.

similar to what Eric Price, in a blog post about the NIPS consistency experiment, called the *messy middle* [21]. In the center of this zone, success rate is around 50%, meaning that decisions on such papers amount to a coin flip. On either side of this "danger zone" are papers that are accepted or rejected with a higher proportion, but whose decision is still subject to a non-negligible amount of noise. Decreasing ξ even further results in the blue line of Figure 4, which is almost linear. At the very end of the quality spectrum, papers of extremely high quality only have a probability of about $3/5$ of being accepted.

One may ponder to what extent such a situation occurs in reality. Very few conferences that we know of collate (let alone divulge) statistics about their reviewing process, apart from acceptance rate. We may however turn to the famed NIPS experiment for some basis for comparison. Assuming a normal distribution of paper quality centered on $q = 0$ and standard deviation $1/2$, and after some parameter fiddling,[1] it is possible to come up with a mock PC committee which, through computer simulations, accepts 31% of submissions. This corresponds to an acceptance rate typical of many conferences in Computer Science (including NIPS), give or take a few percentage points.

The resulting PC has a success ratio that behaves exactly as the green line of Figure 4. We can measure that 21% of all submitted papers have a success rate in the interval $[1/3, 2/3]$. The simulation also shares another similarity with the NIPS experiment, in that 24% of all submitted papers get a different decision if reviewed a second time (this was 22% in the NIPS experiment); we call this measure *disagreement*.[2] Finally, we observe that 38% of accepted papers receive a rejection decision when they are evaluated a second time; this is what a post by John Langford has called *arbitrariness* [17]. Arbitrariness was even higher at NIPS 2014, with a reported figure of 60%. Therefore, from the scant data that is available to us, it seems that the simple hypotheses we put forward in this paper have the potential to simulate a PC whose behavior is similar to what has been observed in practice. As a matter of fact, the same S-curve has been obtained through statistical

[1] For the plot in Figure 4, $t = 1/5$, $\xi = 1/2$, $\sigma = 1/2$, $N = 3$, and $b = 3$.

[2] An author of one of the papers submitted to the NIPS experiment commented on his experience: his paper was clearly rejected by one committee and cheerfully accepted by the other [11].

analysis of acceptance rate based on average reviewer score in an actual journal [29].

Note that arbitrariness is mostly unrelated to a program committee's "prestige": it is simply a statistical indicator of its tendency to act as a random process. Therefore, conferences and journals with very low acceptance rates (sometimes viewed as "selective" and therefore higher ranked) are no less immune to arbitrariness than any other. Arbitrariness is indeed lower for venues with extremely low acceptance rates, for the simple reason that rejection becomes the default and firm decision for almost all papers (a reverse reasoning applies to high acceptance rates).

IV. CONSEQUENCES OF ARBITRARINESS

The presence of arbitrariness in a program committee, for whatever reasons, has several negative consequences for the reviewing process. First, it introduces inconsistency: papers that "should" be accepted are sometimes rejected despite the paper's perceived quality. From a strictly human standpoint, this obviously results in frustration and decreased confidence in the reviewing process from the part of authors. Case in point, in the instance of our model that produces data consistent with the NIPS experiment, the value of ξ is $1/2$. This means that only half of a reviewer's score is based on the paper's quality, while the largest part of that score comes from a process which, from the submitter's point of view, behaves at random.

However, it should be noted that this phenomenon also has quantitative impacts, the first being reduced quality. One way of seeing the effect of arbitrariness is that it swaps papers across the quality threshold line: rejected papers of higher quality are being replaced by accepted papers of lower quality. It follows that the selection of papers that are to be published is, on average, increasingly lower as the arbitrariness of the PC increases.

There is another consequence of arbitrariness that has been less studied, which is the phenomenon of *paper bouncing*. If a paper is rejected by a perfect PC, the only possible way for the authors to get it accepted is to increase its quality. With an arbitrary PC, a second course of action is possible: merely re-submitting (i.e. "bouncing") it, either as is or with trivial modifications.[3] Indeed, a paper whose quality lies in the "danger zone" is decided more or less on a coin flip; it therefore seems reasonable to simply flip the coin again in hopes of receiving a more favorable reviewer assignment. This second course of action becomes increasingly appealing as the S-shape of Figure 4 widens.

As an example, Figure 5 superimposes on the same plot the probability that a paper gets accepted after one trial, and the probability that a paper gets accepted when allowed to be submitted one more time if rejected. One can see that bouncing has a positive effect on success rate equivalent to an increase in the paper's quality. In our example, a paper with a quality of $q(\vec{p}) = 0.1$, when resubmitted, increases its odds of being accepted in the same way as if its quality were 0.3. Additional bouncing increases this probability with diminishing returns, yet one can see that the quality level at which papers get accepted steadily decreases: if the number of re-submissions tends to infinity, every paper ultimately gets accepted. In other words, paper bouncing can be viewed as a mechanism that progressively turns an arbitrary PC into a process that accepts everything given enough time.

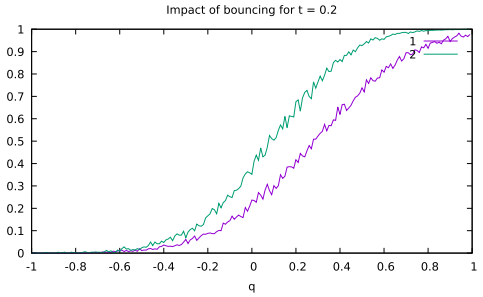


FIGURE 5. A plot of success rate in function of quality, after one (right) and two submissions (left).

Paper bouncing in itself is detrimental to the reviewing process in many ways. First, it obviously increases the load on reviewers, since the same paper gets submitted multiple times. At some point, either it gets accepted, in which case the previous rounds of reviewing that resulted in rejection have been a waste of time; or the authors give up, in which case *all* the rounds of reviewing have been useless since the paper might as well not have been submitted at all. Second, it also delays publication of results, by imposing on some papers with a reasonable quality a few unfortunate rejections before finally allowing them to be published as is. Third, the global quality of published papers is also impacted negatively. Our previous observation showed that re-submitting a paper has an effect on its success rate similar to an increase in quality. However, one should not forget that the paper itself is left unchanged —and so is its quality.

Case in point, we ran a simulation where our pool of papers was submitted to two scenarios. In the first, papers are submitted to an arbitrary PC ($\xi = 1/2$), and are repeatedly bounced when rejected. In the second scenario, papers are submitted to a less arbitrary PC ($\xi = 4/5$), which creates a strong incentive for authors to increase a paper's quality before resubmitting: the quality of each paper is raised by a constant $k = 1/5$ after each rejection. In both scenarios, papers are allowed to be bounced three times. The end result confirms our arguments: average time to publication decreases from 2.72 to 2.46 rounds of reviews, average quality of published papers increases from 0.3 to 0.38, publication rate increases from 56% to 80%, and the total number of reviews performed by the committee decreases by 9%.

However, even attempts at improving a paper's quality may end up having the same effect as bouncing. We recall that in our model, quality is based on a paper's dimensions whose appreciation makes consensus across reviewers; however, which of all the dimensions these actually are is not necessarily known, neither by the authors nor by the reviewers

[3]This point of view is also discussed in a post by Tim Vines [29].

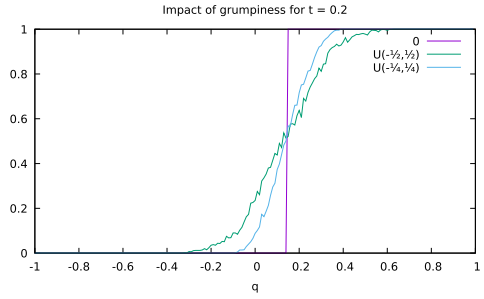


FIGURE 6. The impact of grumpiness on success rate in function of quality.

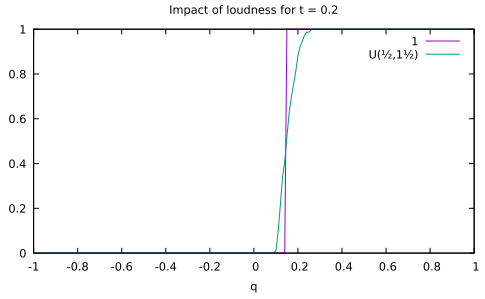


FIGURE 7. The impact of loudness on success rate in function of quality.

themselves. Consider a paper that is rejected by a first set of reviewers R whose definition of quality is based on a set of dimensions D. Suppose that the paper is modified by the authors and re-submitted to a new set of reviewers R', whose notion of quality is based on another set of consensual dimensions D'. It is possible that the modifications on the paper affect its score over dimensions in $D \setminus D'$: this corresponds to aspects of quality valued by members of R, but that members of R' don't care about (i.e. over dimensions they weigh close to zero). Hence, what is a net quality improvement for R is indistinguishable from bouncing for members of R'.[4]

V. REDUCING ARBITRARINESS

We now turn our attention to means by which arbitrariness can be reduced in program committees, by studying the impact of various parameters and modifications to our original model. Obviously, one can trace the core of arbitrariness to the presence of the term $(1 - \xi)\mathcal{N}(x, \sigma)$ in the equation stating the appreciation of each reviewer for a paper. As we explained, this term models the reviewers' disagreement over the dimensions of a paper that should be valued (and whether these elements should be valued positively or negatively). Reducing the impact of this disagreement (i.e. increasing the value of ξ) is obviously a key factor that helps a PC tend towards a perfect quality gate. However, we shall see in the following that arbitrariness, and the S-shape that comes from it, can arise for other factors.

A. REVIEWER-SPECIFIC PARAMETERS

Let us start with reviewer grumpiness, which is the constant bias given to the appreciation of a paper by each reviewer. Figure 6 shows the impact on the quality gate for a committee where reviewers have zero grumpiness, and for committees where reviewer grumpiness is uniformly distributed in a small ($[-1/4, 1/4]$) and a large interval ($[-1/2, 1/2]$). All reviewers are identical except for this parameter; in particular, the amount of randomness in their decision is null. One can see that increasing grumpiness results in an increasingly wider S-shape.

[4]Our model also makes possible a situation where work is made on dimensions that R and R' value in opposite directions. The authors improving the paper based on R's feedback will then receive even worse reviews when submitting it to R'. This is another consequence of re-submission: one always makes changes for the last PC, not the next one.

It is worthy of mention that, in this scenario, all reviewers rank the papers in exactly the same way. That is, they all precisely agree on which paper is better than which.[5] The only difference is in their appreciation of where the acceptable threshold for acceptance lies. Therefore, even when all reviewers value the same elements in a paper in the same way, the mere uneven location of their "quality bar" suffices for a perfect PC to turn into an arbitrary one. Indeed, although grumpiness is a constant for each reviewer, its variability across reviewers makes it act as a random variable when the paper is submitted to a program committee.

Another distinguishing parameter of reviewers is their loudness; we recall that loudness is the modulus of a reviewer's appreciation vector, which translates in our simplified model as a multiplicative constant λ that has for effect of expanding or compressing the appreciation range. Figure 7 shows what happens to a perfect PC when the reviewers' loudness is allowed to vary, where λ is uniformly picked in the interval $[1/2, 3/2]$. This represents a situation where some reviewers are quieter than they should, while some others are louder than they should. As one can see, variation in loudness also introduces arbitrariness, although of a different shape as for grumpiness. Again, variation in loudness has no effect on the relative ordering of papers made by each reviewer —that is, all reviewers still rank all papers in the same order.

B. COMMITTEE-SPECIFIC PARAMETERS

The parameters we studied so far were concerned with variability between reviewers. Other parameters determine how a program committee collects and synthesizes appreciations from reviewers, and makes a decision on acceptance or rejection.

One first obvious parameter is the number N of reviewers that are asked to give their appreciation. All simulations up to this point have been run with $N = 3$, which is the typical number of reviews that most papers receive in conferences and journals. Figure 8 shows the impact on arbitrariness that a higher or lower number of reviews for each paper may have. As expected, an increase in the number of reviews brings the function closer to its perfect square shape, while a decrease in the number of reviews has the opposite effect;

[5]With the exception of reviewers whose appreciation vector is the null vector, which we assume never happens.

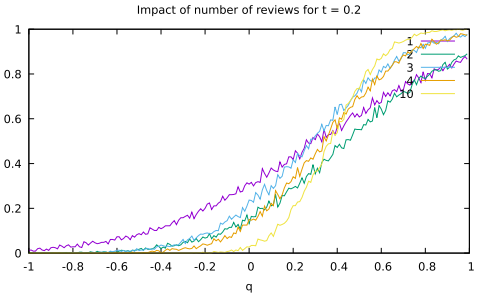


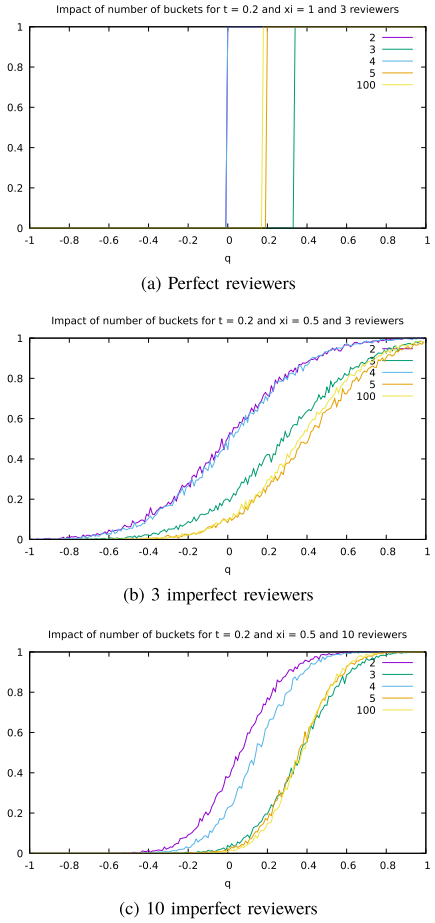
FIGURE 8. The impact of number of reviewers on success rate in function of quality.

at $N = 1$, the function is close to linear for a large part of the quality spectrum. The positive impact of an increase in number of reviews can be seen as a manifestation of the "wisdom of crowds" [27], which, in this case, is such that individual reviewer variations on each paper's appreciation progressively cancel out to reveal the paper's intrinsic quality. Note however that increasing the number of reviewers produces diminishing returns: we can measure that arbitrariness is at 60% with a single review, 45% with three reviews, 40% with four, and 30% with ten.

Another interesting question is whether decisions on papers become more precise when reviewers are allowed a wider range of scores. Although intuition may hint that the answer is positive, it is nevertheless interesting to examine how a change in the scoring scale can alter the global decisions of a PC. Let us first examine the impact of scoring granularity for perfect reviewers ($\xi = 1$, $\lambda = 1$, $\gamma = 1$ for everybody). Figure 9a shows the success function in such a scenario, for 2 up to 100 scoring levels. With two levels, reviewers are only allowed to register a pass/fail verdict; with 100 levels, the scale is getting closer to the continuous appreciation function from which the score is extracted.

One can see that, in the absence of other sources of noise, each PC still acts as a hard quality gate: they only differ in the quality cutoff threshold between certain acceptance and certain rejection. A somewhat more surprising element is how each scale applies a systematic positive or negative bias to papers lying in a specific quality interval. For example, the quality cutoff for the scales with 2 and 4 levels lies exactly at $q = 0$, which means that papers rejected as per the perfect quality threshold ($1/5$) are actually accepted by the "coarse-grained" PC. The reverse effect can be observed for the scale with 3 levels: this time, some accepted papers in the perfect PC become rejections in the coarse-grained one. This trend goes against the notion that what could be viewed as *quantization noise* cancels out across all reviewers.

This systematic bias globally lessens as the number of levels increase, although not steadily. This is explained by the fact that this bias is caused by a mismatch between the PC's quality cutoff threshold, and the locations of the discrete score jumps across the continuous appreciation interval. Case in point, the scoring scale with 5 levels has a category boundary at $q = 1/5$, and we can see from the plot that it makes decisions on papers with the same precision as the scale with 100 levels.



(a) Perfect reviewers

(b) 3 imperfect reviewers

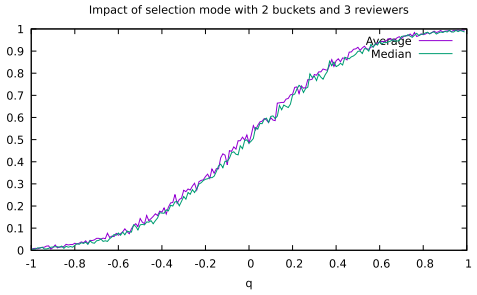
(c) 10 imperfect reviewers

FIGURE 9. The impact of scoring granularity on success rate in function of quality.

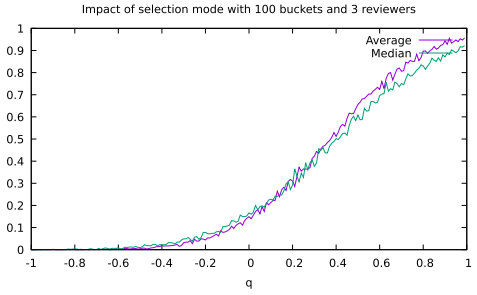
From this observation, one can conclude that the use of a finer-grained scoring scale is not desirable *per se*, but only because a scale with more levels lessens the probability that the PC's cutoff threshold lies far from a discrete category boundary.

Therefore, scoring granularity is one parameter of a program committee that does not appear to influence its arbitrariness, but only its effective acceptance threshold t, shifting it higher or lower depending on the scale and the value of t. An additional example can be seen in Figure 9b, where γ, λ and ξ are restored to their original distributions. It can be observed that the S-shape of the success function is identical for all scales, and is merely shifted left or right on the quality axis. This tends to indicate that a single pass/fail verdict from each reviewer could be sufficient, provided that it is corrected for bias. One possibility is adding more reviewers; Figure 9c shows the same scales, but where each paper is reviewed by 10 people instead of 3.

Finally, one may ask whether the threshold on average score could be replaced by another method for selecting papers. It is often argued that the median is a central trend measure that is more robust to the presence of extreme values. Figure 10a shows the impact of evaluating a threshold on the median score given by reviewers, instead of the average score.



(a) Binary score



(b) 100-valued score

FIGURE 10. The impact of selection mode on success rate in function of quality.

Surprisingly, there is no discernible difference; the situation remains the same even when increasing the granularity of the scoring scale to 100 levels, as is shown in Figure 10b. In fact, the use of the median does reduce variability, but in the range of scores that each trial of the same paper receives; however, the probability that this score lies over or under the cutoff line remains unchanged. In other words, the interval of scores received by each paper is compressed, but not is shifted left or right with respect to the committee's quality threshold. Therefore, it looks like replacing the average by the median is equivalent, in terms of the committee's arbitrariness on the long run.

VI. ALTERNATE COMMITTEES

The proposed computational model makes it possible to simulate and study other, more drastic changes to the way papers are being selected by a committee. We describe a few such models in the following.

A. RANKING COMMITTEE

In this alternate committee, each reviewer is given a pool of papers to review. The appreciation for each paper is done in the same fashion as before, by applying the equation given in Section III on each paper, according to the reviewer's specific parameters –that is, this appreciation is still subject to loudness, grumpiness and arbitrariness. However, instead of turning these continuous appreciation values into a location on the discrete scoring scale, reviewers are merely asked to provide the ordering of the papers, from the one they consider best down to the paper they consider worst. This technique, called *ordinal rating*, has been advocated in other fields as a means to avoid various biases and calibration issues

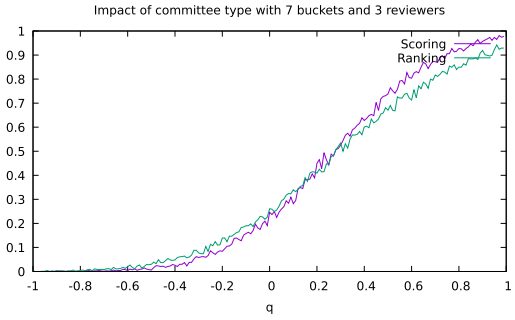


FIGURE 11. Comparison between a scoring and a ranking committee.

that otherwise arise in the traditional scoring system (called *cardinal rating*) [15], [20].

A numerical value is still associated to each reviewed paper, this time by taking the fraction of papers in the reviewer's pool it ranks above (excluding itself). Thus, the first paper in the list is given a score of 1, while the last is given 0. Thus, every paper evaluated by a single reviewer is assigned a normalized value in the interval $[0, 1]$. This direct translation of cardinal scores into ordinal ones can be supported by the empirical measurement, in real-world program committees, of a high level of agreement of the ordinal rankings with the cardinal scores when both are asked of the reviewers [25].

The rest of the operation proceeds in the same way as before: scores for each paper are aggregated (for example, by taking the average), and papers above a predetermined threshold are accepted. The intuition behind this model is that, as observed in Section III, part of a committee's arbitrariness results in the presence of parameters λ and γ –yet, these two parameters have no effect on the way each reviewer ranks the papers relative to each other. Asking reviewers to merely rank the papers, without requiring them to locate them on an absolute scoring scale, should therefore cancel the effects of λ and γ, leaving ξ as the sole source of randomness from a reviewer's standpoint.

Figure 11 compares the behavior of two committees with the same reviewers: the first operates using a discrete scoring scale, while the second uses the ranking method described above. The $[0, 1]$ threshold for the ranking committee is adjusted so that it accepts the same number of papers as the corresponding scoring committee, so their acceptance rate is identical. The plots show that they behave in a strikingly similar manner, which seems to contradict the aforementioned claims that ordinal ranking avoids various forms of bias and miscalibration. However, although the ranking committee lessens the impact of grumpiness and loudness on the reviewer's decisions, in counterpart, it is more sensitive to the assignment of papers to reviewers. After all, a reviewer that is given an exceptionally good (or bad) patch of papers to evaluate has no way of making them stand out by giving all of them high or low scores. It seems, from the results of these simulations, that these two effects balance each other more or less. Miscalibration bias would indeed vanish, but only if all reviewers were assigned all papers.

B. VARIABLE-PRECISION COMMITTEE

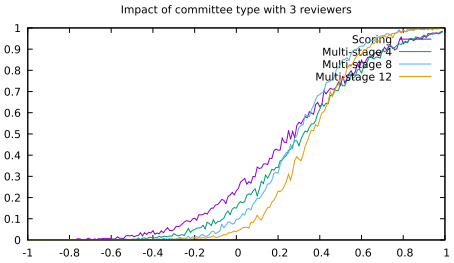
The previous committee resulted in the same number of reviews per paper as for the classical scoring committee, and hence corresponds to a consumption of the same amount of "reviewer effort". Yet, we have seen in Figure 8 that putting more reviewer effort does sharpen a committee's verdicts, but requires an impractical increase to show noticeable benefits. This is in part due to the fact that every paper is given the same heightened scrutiny. A smarter management of the limited reviewing resources would, in contrast, do away with submissions that reach a clear positive or negative consensus, and direct more effort towards submissions lying in the messy middle.

To this end, we study an alternate committee model that operates in n stages. Papers are rated by reviewers using the same discrete scoring scale as before, and a predefined threshold score t is used as a guideline for acceptance or rejection; however, each stage of the reviewing process is also associated with an interval width w. In the first stage, each paper is assigned to a single reviewer; any paper whose discrete score lies above $t + w_1$ is immediately accepted, while any paper whose score lies below $t - w_1$ is immediately rejected. Papers in the interval $[t - w_1, t + w_1]$ are deemed not to be decisively assessed, and move on to the second stage.

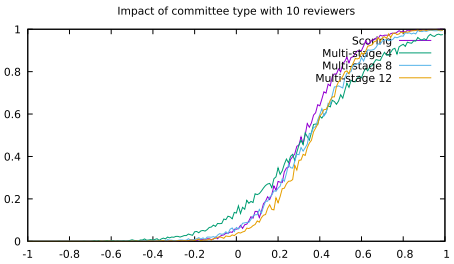
This second stage operates in the same way as the first: submissions are assigned a second reviewer, and the aggregate score (e.g. average) of both the first and the second review is considered. The interval width is decreased to a value $w_2 < w_1$; any papers outside the interval $[t - w_2, t + w_2]$ are either accepted or rejected, and the remaining ones move to the next stage, with yet one more reviewer and a smaller interval. Once the last stage is over, multiple courses of action are possible for the papers that have still not been decided. They can all be accepted, rejected, picked as in a classical scoring committee based on whether their final aggregate score lies over the threshold t, or even be chosen on a coin flip.

A particular feature of this model is that it spends the most reviewing resources on papers that consistently straddle the threshold line, while papers that converge more rapidly towards a decision are expelled from the process; thus each paper is given a variable amount of attention. Figure 12 shows the results of such a committee, with the same pool of reviewers and papers, compared with variants of the multi-stage committee. For this particular experiment, $b = 3$, $w_1 = 2$, and the interval width is reduced at each stage by setting $w_{i+1} = w_i \cdot 3/4$. Figure 12a plots the multi-stage committees against a scoring committee with a fixed number of 3 reviews per paper, and Figure 12b with 10 reviews per paper. As for the ranking committee, the threshold value t of the multi-stage committee has been set so that it produces an acceptance rate similar to that of the single-pass committee.[6]

[6]The only effect of altering t in both models is the lateral translation of each function; it has no effect on the shape of the curve, which is the point of the discussion.



(a) 3 reviews per paper



(b) 10 reviews per paper

FIGURE 12. Comparison between single- and multi-pass scoring committees.

The results show that multi-stage committees achieve a steeper quality-success function than the classical single-stage scoring committee. With $n = 8$, the multi-stage committee produces a curve similar to what a single-stage committee achieves with 10 reviews per paper. However, the latter does so at the cost of a total of more than 2 million reviews, while the multi-stage approach requires less than half that number (about 840,000). This shows the potential of such a committee to automatically tune the precision required for each paper.

Also worthy of mention is the fact that the fate reserved to papers that make it to the final stage is more or less irrelevant in terms of arbitrariness. Case in point, Figure 13 shows a plot for multi-stage committees that vary only in that final step, for each of the four alternatives discussed above. Each curve is shifted left or right, indicating a slight tendency towards acceptance or rejection; however, the shape of each function shows no noticeable difference. In particular, the coin-flip decision is no more arbitrary than a decision based on the paper's aggregate score. This relatively surprising result indicates that the committee is ultimately left with a set of papers whose quality cannot be meaningfully assessed by its reviewers.

Some criticism can be addressed towards such a committee model. First, it requires time: since each stage depends on the results of the previous one, none can be executed in parallel and every reviewer evaluates a single paper at a time. This makes it ill-suited to fixed-deadline venues such as conferences, but could prove less of an issue for continuous reviewing processes such as journals. Second, although the good faith of each participant is taken for granted, a single malicious reviewer could decide on a paper by giving an artificially high or low score at the first stage of the process.

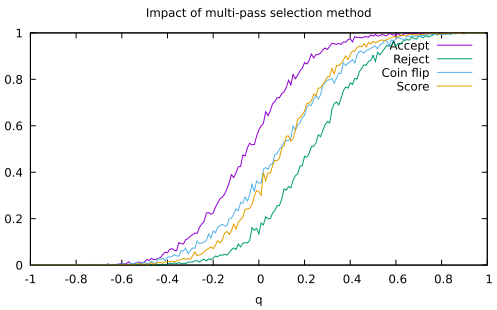


FIGURE 13. The impact of multi-pass selection mode on success rate in function of quality.

The feasibility of such an attack decreases as the stages pass. Finally, the model makes it difficult to adjust its parameters once they have been set: modifying the value of t, w_1, or the recurrence relation between w_i and w_{i+1} cannot realistically be done without restarting the process from the beginning. This is a departure from existing, single-phase scoring committees where thresholds can be decided once all reviews have been received. Whether this is a benefit or a drawback of the model is left to discussion.

C. JOURNAL-STYLE COMMITTEE

Our original model is geared towards a conference-style program committee, where papers are submitted all at once, and a decision is made on all of them at once after a single round of review. Journals typically work in a drastically different way, as they accept papers continuously, and allow multiple reviewing cycles. It can be seen as the exact opposite of the bouncing strategy: while bouncing sends a paper without modification to a different set of reviewers, the journal process has the same reviewers re-assess the paper after mandatory modifications have been made to it.

As for a conference model, we assume that a re-submission with corrections increases the paper's intrinsic quality. However, the journal model introduces a second effect. Since the reviewers assessing the revised version are generally the same as the original submission, one can expect that the revision of the paper, conducted based on reviewers' comments, increases the paper's quality for the dimensions that are valued by the reviewer. Expressed in terms of papers and reviewers as vectors, a revision in the journal model therefore increases the modulus of the paper's vector, but also reduces the angle between this vector and that of the reviewer, as is shown in Figure 14. In our simplified unidimensional model, this amounts to an increase of both parameters q and ξ when the paper is reviewed a second time. Of all models studied so far, this is the only one whose hypotheses contribute to a direct increase on parameter ξ.

It may therefore be relevant to measure whether the journal workflow improves the arbitrariness of a program committee. The effect can be visualized by a simulation illustrated in Figure 15. It compares the acceptance rate in function of quality between a (scoring) conference-style committee,

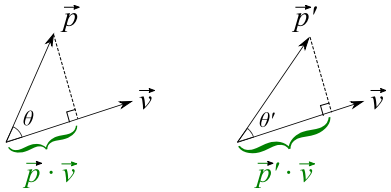


FIGURE 14. When a paper $\vec{p}$ is revised and re-assessed by the same reviewer $\vec{v}$, one expects that $\theta' < \theta$.

and a journal-style committee as discussed above, with the same parameters (see footnote 1). In both cases, papers can be submitted up to five times. Each time a paper is rejected, its quality q is increased by 1/5. In the case of the conference committee, the paper is then re-submitted to a new randomly selected set of reviewers, and re-evaluated. In the case of the journal committee, the paper is re-submitted to the same set of reviewers who evaluated it the first time. To account for the fact that paper revisions, in this case, tend to decrease the angle θ between the paper and the reviewer's vectors, the value of ξ is also increased by a small amount (here 1/20).

As one can see, even a modest realignment of a submission towards quality dimensions valued by the reviewers can produce a sharp reduction of the arbitrariness, all other things considered equal. In other words, the "moving target" phenomenon induced by the memoryless succession of program committees can be seen as a contributing factor for arbitrariness. Note that this is not achieved at the expense of increased reviewer effort, since the total number of reviews in both simulations remains within a margin of 3%.

VII. LIMITATIONS AND THREATS TO VALIDITY

The proposed model and the theoretical reflections that ensued should be taken for what they are: an obvious simplification of reality that necessarily cuts a few corners. We mention a few such corners in the following.

Our model of a program committee supposes that every paper with the same discrete scores receives the same accept/reject verdict —that is, selection is based solely on quality assessment and is independent of the quality of other papers submitted to the same committee. Yet, many PCs operate under different constraints, such as a maximum number of papers to accept, or a target acceptance rate not to exceed regardless of the submissions' global merit. Even though quality is involved in the process (for example, by ranking papers in order of appreciation based on reviewer feedback), further arbitrariness can be introduced by the need to draw a line through a patch of papers that look all alike, quality-wise. Reviewers themselves may apply the same kind of reasoning, by downgrading (or upgrading) the score of a paper relative to the quality of other submissions they are asked to review in the same committee. Such a selection model could also be studied by starting from the same principles.

Each dimension of a paper's appreciation is also assumed to be evaluated in an independent and systematic way by a reviewer. This eschews known psychological factors that are

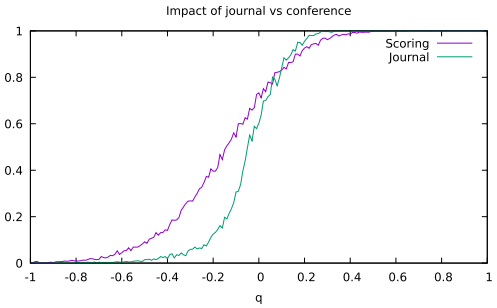


FIGURE 15. Comparison of a conference-style (scoring) committee, with a journal-style committee.

involved in reviewing; for example, a badly-written paper will often score less on technical merit than a well-written one, despite being of an equivalent technical value. That is, good or bad quality on a dimension can create ''crosstalk'' over other dimensions of the paper —a form of confirmation bias. In addition, our model dismisses any interaction between reviewers that could make some of them change their score, and possible rebuttals by authors that could have the same effect (although the effect of rebuttals has been discussed earlier and appears to be marginal).

Our empirical measurements have also assumed a uniform distribution in the intrinsic quality of papers. This parameter could be challenged in its own right; for example, other studies have rather used a skewed distribution of quality to run simulations [28]. A non-uniform distribution with a higher proportion of papers in the ''messy middle'' would amplify the effects of arbitrariness observed in the empirical measurements conducted in this paper.

Finally, any doubts that can be raised regarding the realism of our proposed model can be deflected towards the broader issue of the scarcity of hard data about peer reviewing. As we already mentioned, the only reliable source of information that can be extracted from conferences and journals is their acceptance rate, which paints an arguably very fragmentary portrait of the state of a program committee. Submitting a paper to a specific venue is the result of a choice by the authors, based on a number of conscious and unconscious criteria. It would be reasonable that one such criterion be the odds that a given venue evaluates the submission in a systematic and reproducible fashion, instead of deciding its fate on what amounts to a biased coin flip. One could even argue that a strong indicator of a paper's value should be how decisive is its acceptance by a program committee across multiple trials —something that is hinted by a venue's low arbitrariness, not its one-shot acceptance rate, as low as it may be.

VIII. RELATED WORK

Most works that studied the peer reviewing process have concentrated on an empirical analysis of trends observed in actual publication venues. In the field of Computer Science, an important large-scale study by Ragone *et al.* has involved the compilation of statistics for more than 9,000 reviews of 2,800 submitted contributions [22]. The paper formally defines the concept of peer review *validity*, which is the capability of a process to identify submissions that are scientifically correct, and which are likely to have an impact or be of interest to the publication venue's audience.

Some elements of the study are not addressed by our simple mathematical model, such as the correlation between reviewing scores and eventual impact in terms of citations (which the authors empirically measured as being weak). However, some other observations find an echo in the notions we discussed in the last few pages. For instance, the authors empirically observed the presence of reviewers that ''consistently give higher (or lower) marks than the others independently from the quality of the specific contribution they have to assess''; this is precisely accounted for in our proposed model by the grumpiness parameter γ, which they call *rating bias*. Similarly, the study observes reviewers giving ''marks that are always very close to the threshold for a given criteria'' (such as 3 on a scale from 1 to 5); this time, this directly corresponds to the loudness parameter λ, which Ragone *et al.* call *threshold bias*.

The paper also observed greater agreement between reviewers for papers at both extremes of the scale (very good and very bad): this also matches the behavior of program committees simulated by our model. Therefore, it seems that our proposed model reproduces features that are indeed observed empirically. Finally, the paper proposes mechanisms to improve the reviewing process in order to decrease reviewing effort. One of them is a multi-phase model similar to the variable-precision committee simulated in Section VI-B. Here again, observations from our simulations coincide with the conclusions of the paper.

Schultz [24] conducted an empirical study assessing the impact on the number of reviewers in journal submissions, by analyzing the fate of 500 manuscripts submitted to a single journal. It observed that rejection rates were not significantly different whether two or three reviewers were used. It should be noted, however, that rejection rate is a different concept from arbitrariness, which is the focus of the present paper.

Among other works on the empirical evaluation of existing processes, we already cited the well-known NIPS experiment [17], [21], from which stems the idea of studying program committees in terms of their arbitrariness. The 2016 edition of the conference performed a *post hoc* analysis of its reviewing process [25], and observed, among other findings, ''significant miscalibration with respect to the rating scale''. An empirical study of comparative peer reviewing (similar to the ranking committee simulated in Section VI-A) yielded positive results in that reviewers produced comments that were both longer, and better-rated by an external group of experts [6]. Note however that this study does not measure the arbitrariness of the committee, and moreover was not focused on the reviewing of scientific research papers.

It shall be noted that our proposed formal model does not incorporate features present in many program committees,

such as the possibility for authors of responding to reviews. However, it can be argued that such features have little impact on the simulation. For example, in the field of Natural Language Processing, Gao *et al.* studied the impact that rebuttals have on the final decision given to a paper [12]. They revealed that marginal (and statistically significant) influence on the final scores (especially for borderline papers), but that a reviewers' decision is largely determined by their initial score and the distance to the other reviewers' initial scores. Results from NIPS 2016 also observed little impact of these rebuttals on paper scores [25].

Fewer studies have attempted to define and simulate models of peer reviewing. Among these, Bentley defines a model of peer reviewing for grant proposals using genetic algorithms [2], where a proposal is modeled as a gene, with reviewers assessing the quality of a proposal across five dimensions. Day also uses simulations to study the effect of bias on the attribution of grant among between a ''preferred class'' and a ''non-preferred class'' of principal investigator [10].

On his side, Allesina uses an agent-based model of reviewers in a journal-style process [1]. Each paper is modeled as a three-dimensional vector $\vec{p} \in [0, 1]^3$, with the three dimensions respectively representing the paper's fit with respect to the journal's topic (T), its technical quality (Q), and its novelty (N). This model differs from the one proposed here in that it takes into account the assignment of reviewers to papers based on their expertise with respect to the submission. Each reviewer of a paper is asked to produce an estimate of its true three vector parameters, which carries an error inversely proportional to the familiarity of the reviewer with the topic of the manuscript. The editor aggregates these estimates and computes the product of the cumulative distribution functions for all three parameters. This distribution is used to derive a probability p_a; the manuscript is accepted based on a biased coin flip with probability p_a of being accepted. The author uses this model to explore the consequences of an alternate model where the editor is allowed to reject without review papers that will realistically result in rejection.

Tan *et al.* also provide a simulation model of peer reviewing, with a focus on the measurement of perceived quality decrease in journal standards [28]. Of particular interest is the fact that this work also studies the impact of paper resubmissions on overall quality, and results in similar conclusions regarding the practice of paper bouncing (although not named as such).

Finally, Kovanis *et al.* also use an agent-based model of peer reviewing, in order to study a model where past reviews of a paper are shared when this paper is re-submitted [16]; this goes in line with our assessment of Section VI-C, which tends to favor a system where papers are re-submitted to the same reviewers (thus providing a stateful reviewing process). As with our current model, the proposed one also models a paper by an intrinsic quality score, which is represented as the flattening of multiple quality dimensions. Evaluation of a paper by a reviewer is subject to a scoring error assuming a given probability distribution. However, this work differs from our proposed contribution in two respects. First, it focuses on the total effort and time to publication, and not on the arbitrariness of the decisions made by a program committee. Second, the presence of randomness in a reviewer's decision is taken as a design *hypothesis*, whereas in our proposed model, it is merely a *consequence* of the mismatch between a paper and a reviewer's vectors. As a matter of fact, to the best of our knowledge, the model presented in this paper is the first where randomness is not taken for granted, and is rather explained from higher principles.

IX. CONCLUSION

In this paper, we developed a mathematical model of the scientific peer reviewing process. This model is grounded on the simple principle that a research paper can be modeled as a numerical vector representing multiple independent dimensions of its intrinsic quality. Reviewers are also modeled as vectors, where each dimension corresponds to the value they assign to each quality dimension. The assessment of a paper by a reviewer becomes nothing but the dot product of their respective vectors. This model is sufficient to explain the presence of arbitrariness in the reviewing process, which occurs when the same paper receives inconsistent accept/reject decisions depending on the set of reviewers it is assigned to.

In this context, arbitrariness occurs in the presence of a mismatch between non-null entries of the paper's quality components, and the reviewers' value given to each component. We have shown how a greater mismatch makes the decision on a paper behave increasingly like a random process, leaving an ever-smaller fraction of the total ''score'' corresponding to an actual assessment of the paper's quality. An important take-home point is the observation that the appearance of randomness from the author's standpoint does not imply that reviewers themselves are random. As we stressed earlier, it is possible for all reviewers to rank all papers in the same order, and still end up with a reviewing process that produces arbitrariness.

Based on these principles, a simplified equation representing a paper's assessment was derived, which contains two main terms: the fraction ξ of the paper's score dependent on its quality q, and the fraction $1 - \xi$ of the score behaving at random. This function can then be scaled by a constant factor λ and shifted by another constant γ, which we respectively called the ''loudness'' and ''grumpiness'' of a given reviewer.

An interest of this model is its relative simplicity, which makes it easy to run a large number of fake reviewing committees on a population of papers of randomly-generated quality. This made it possible to study the impact of various parameters on the perceived arbitrariness of a committee. To this end, we presented the results of computer experiments varying the values of variables λ, γ and ξ mentioned earlier, and also explored the effect of other elements of the reviewing process, such as the number of reviewers and the granularity of the scale by which continuous scores are discretized.

Despite the various negative consequences that arbitrariness can bring, our numerical simulations have shown that the options to reduce it are few and far between. In particular, decreasing variance in reviewer loudness, changing the scoring granularity or replacing average threshold by median threshold all have little to no impact on a committee's arbitrariness. Even increasing the number of evaluations per paper would require an impractical amount of reviews to iron out inter-reviewer variability and restore the perfect quality gate that would be expected of a PC. In our proposed model, the only two parameters that have any meaningful impact on the S-shape of a committee's success function are ξ and γ. That is, the surest way to decrease arbitrariness is not only to make reviewers agree on the largest possible number of dimensions of a paper (ξ is close to 1), but also ensure they have equal severity (γ is close to 0, or at least similar for all reviewers).

This is at the same time obvious, and also easier said than done, but recent initiatives have been started to reverse this trend. Worthy of mention are the ACM SIGSOFT Empirical Standards, which stem from the observation that "constant rejection is rooted in dissensus within scientific communities regarding how research should be conducted" [23]. By striving to provide a systematic and agreed-upon set of evaluation criteria for empirical papers in Software Engineering, this standard has the potential to contribute to reducing the part of a paper's appreciation left to a reviewer's personal taste —which, in our proposed model, translates into a decrease of parameter ξ. Some scholars have argued that upcoming scientific revolutions may occur because of a change in the way publications are evaluated [5]. It is to be hoped that similar endeavors in other fields of Computer Science will, in time, help reinstate predictability and confidence in the peer reviewing process.

REFERENCES

[1] S. Allesina, "Modeling peer review: An agent-based approach," *Ideas Ecol. Evol.*, vol. 5, no. 2, pp. 27–35, 2012.

[2] P. J. Bentley, "The game of funding: Modelling peer review for research grants," in *Proc. Genet. Evol. Comput. Conf. (GECCO)*, F. Rothlauf, Ed., Montreal, QC, Canada, Jul. 2009, pp. 2597–2602.

[3] B. Bilalli, R. F. Munir, and A. Abelló, "A framework for assessing the peer review duration of journals: Case study in computer science," *Scientometrics*, vol. 126, no. 1, pp. 545–563, Jan. 2021.

[4] L. Bornmann and H.-D. Daniel, "Selection of research fellowship recipients by committee peer review. Reliability, fairness and predictive validity of board of trustees' decisions," *Scientometrics*, vol. 63, no. 2, pp. 297–320, Apr. 2005.

[5] M. Buchanan, "Come the revolution," *Nature Phys.*, vol. 6, no. 2, p. 2, 2010.

[6] J. Cambre, S. Klemmer, and C. Kulkarni, "Juxtapeer: Comparative peer review yields higher quality feedback and promotes deeper reflection," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, R. L. Mandryk, M. Hancock, M. Perry, and A. L. Cox, Eds., Montreal, QC, Canada, Apr. 2018, p. 294.

[7] A. Campos-Arceiz, R. B. Primack, and L. P. Koh, "Reviewer recommendations and editors' decisions for a conservation journal: Is it just a crapshoot? And do Chinese authors get a fair shot?" *Biol. Conservation*, vol. 186, pp. 22–27, Jun. 2015.

[8] G. Cormode, "How NOT to review a paper: The tools and techniques of the adversarial reviewer," *ACM SIGMOD Rec.*, vol. 37, no. 4, pp. 100–104, Mar. 2009.

[9] G. Cormode, A. Czumaj, and S. Muthukrishnan. (2004). *How to Increase the Acceptance Ratios of Top Conferences?* Accessed: Apr. 20, 2021. [Online]. Available: https://www.cs.rutgers.edu/ muthu/ccmfun.pdf

[10] T. E. Day, "The big consequences of small biases: A simulation of peer review," *Res. Policy*, vol. 44, no. 6, pp. 1266–1270, Jul. 2015.

[11] A. Defazio. (Oct. 2014). *The NIPS Consistency Experiment*. [Online]. Available: https://www.aarondefazio.com/tangentially/?p=21

[12] Y. Gao, S. Eger, I. Kuznetsov, I. Gurevych, and Y. Miyao, "Does my rebuttal matter? Insights from a major NLP conference," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 1274–1290.

[13] S. Halle, R. Khoury, and M. Awesso, "Streamlining the inclusion of computer experiments in a research paper," *Computer*, vol. 51, no. 11, pp. 78–89, Nov. 2018.

[14] S. Hallé, "Computer simulations of scientific peer reviewing (LabPal package)," Zenodo, Tech. Rep. 2021, doi: 10.5281/zenodo.5168782.

[15] A.-W. Harzing *et al.*, "Rating versus ranking: What is the best way to reduce response and language bias in cross-national research?" *Int. Bus. Rev.*, vol. 18, no. 4, pp. 417–432, 2009.

[16] M. Kovanis, L. Trinquart, P. Ravaud, and R. Porcher, "Evaluating alternative systems of peer review: A large-scale agent-based modelling approach to scientific publication," *Scientometrics*, vol. 113, no. 1, pp. 651–671, 2017.

[17] J. Langford. (Jan. 2015). *The NIPS Experiment*. [Online]. Available: https://cacm.acm.org/blogs/blog-cacm/181996-the-nips-experiment/fulltext

[18] N. Lawrence. (Dec. 2014). *The NIPS Consistency Experiment*. [Online]. Available: http://inverseprobability.com/2014/12/16/the-nips-experiment

[19] B. Meyer. (Aug. 2011). *The Nastiness Problem in Computer Science*. [Online]. Available: https://cacm.acm.org/blogs/blog-cacm/123611-the-nastiness-problem-in-computer-science/fulltext

[20] N. Stewart, G. D. A. Brown, and N. Chater, "Absolute identification by relative judgment," *Psychol. Rev.*, vol. 112, no. 4, p. 881, 2005.

[21] E. Price. (Dec. 2014). *The NIPS Experiment*. [Online]. Available: https://blog.mrtz.org/2014/12/15/the-nips-experiment.html

[22] A. Ragone, K. Mirylenka, F. Casati, and M. Marchese, "On peer review in computer science: Analysis of its effectiveness and suggestions for improvement," *Scientometrics*, vol. 97, no. 2, pp. 317–356, Nov. 2013.

[23] P. Ralph *et al.*, "ACM SIGSOFT empirical standards," 2020, *arXiv:2010.03525*. [Online]. Available: https://export.arxiv.org/abs/2010.03525

[24] D. M. Schultz, "Are three heads better than two? How the number of reviewers and editor behavior affect the rejection rate," *Scientometrics*, vol. 84, no. 2, pp. 277–292, Aug. 2010.

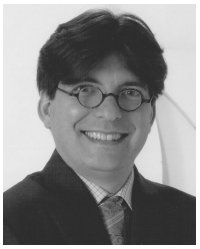
[25] N. B. Shah, B. Tabibian, K. Muandet, I. Guyon, and U. von Luxburg, "Design and analysis of the NIPS 2016 review process," *J. Mach. Learn. Res.*, vol. 19, pp. 49:1–49:34, Sep. 2018.

[26] P. Spyns and M.-E. Vidal, *Scientific Peer Reviewing: Practical Hints and Best Practices*. Cham, Switzerland: Springer, 2016.

[27] J. Surowiecki, *The Wisdom of Crowds*. New York, NY, USA: Doubleday, 2004.

[28] Z. Tan, N. Cai, J. Zhou, and S. Zhang, "On performance of peer review for academic journals: Analysis based on distributed parallel system," *IEEE Access*, vol. 7, pp. 19024–19032, 2019.

[29] T. Vines. (Dec. 2011). *Is Peer Review a Coin Toss?* [Online]. Available: https://scholarlykitchen.sspnet.org/2011/12/08/is-peer-review-a-coin-toss/



SYLVAIN HALLÉ (Senior Member, IEEE) received the Ph.D. degree from the Université du Québec à Montréal. After completing his Ph.D. degree, he started working at UQAC, in 2010, after working as a Postdoctoral Research with the University of California Santa Barbara. He is currently the Canada Research Chair in software specification, testing, and verification and a Full Professor of computer science with the Université du Québec à Chicoutimi, Canada. His research interests include software testing, formal verification, and computer networks and communications. He has earned several best paper awards for his work on the application of formal methods to various types of software systems.