UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE

38050 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

IS PEER REVIEW ANY GOOD?
A QUANTITATIVE ANALYSIS OF PEER REVIEW

Fabio Casati, Maurizio Marchese, Azzurra Ragone, Matteo Turrini

August 2009

Technical Report # DISI-09-045

Preliminary Draft

Is peer review any good?
A quantitative analysis of peer review
Preliminary Draft

Fabio Casati, Maurizio Marchese, Azzurra Ragone, Matteo Turrini

Universita degli studi di Trento, Trento, Italy
{casati,marchese,ragone,turrini}@disi.unitn.it

Abstract. In this paper we focus on the analysis of peer reviews and reviewers behavior in conference review processes. We report on the development, definition and rationale of a theoretical model for peer review processes to support the identification of appropriate metrics to assess the processes main properties. We then apply the proposed model and analysis framework to data sets about reviews of conference papers. We discuss in details results, implications and their eventual use toward improving the analyzed peer review processes. Conclusions and plans for future work close the paper.

1 Introduction

Review systems are primarily used both for assessing the quality of a given work and, consequently, to provide hints about the reputation of the author who crafted it. The role of this activity is fundamental, in order to increase the quality and productivity related to a particular scientific area and also in credits assignment.

In this sense, the goal of our work is to understand the characteristics of the review systems available nowadays in academia. In parallel, there is the need of highlighting weaknesses and strengths of the process and of studying aspects potentially related to two core features that should, in principle, characterize every evaluation system: fairness and efficiency. Once all those aspects have been sufficiently understood, there might be the possibility to propose new models for the evaluation, that could provide further metrics and algorithms for setting up the process more easily, also trying to get some improvements.

We start from the definition of a theoretical model for peer reviews to support the identification of appropriate metrics to assess relevant properties of the review processes, that could help, then, to improve the process itself. Specically, we aim at a peer review process that allows the selection of high quality papers and proposals, that is fair and efficient (in terms of minimizing the time spent by both authors and reviewers).

The outcome of this line of work has been the definition of metrics:

- to identify and understand correlation between different ranking criteria and assessment phases;
- to analyzes the robustness of the evaluation process to stochastic perturbation of the marks distribution; this allow us to quantify and put in the proper scale the computed values of disagreement and rating biases

– to identify and understand the average level of disagreement and rating biases present in the process and the assessment of compensation methods
– to compute efficiency related metrics, such as correlation and minimal number of reviews.

All these information are useful to better characterize the different evaluation processes, making these more open and transparent and indicating ways to improve them. Once the model and the analysis tools have been defined, we have applied them to data about reviews of conference papers. Interesting aspects of the review system have been inferred from the analysis, which could help in the understanding of such complex process. In brief, for the specific analyzed data sets:

– there is a significant degree of randomness in the review process, more marked than we initially expected; the disagreement among reviewers sometimes is pretty high and there is very little correlation between the rankings of the review process and the impact of the papers as measured by citations. Furthermore, in the majority of the studied cases rankings and acceptances are sensitive to variations in the marks given by reviewers;
– it is always possible to identify groups of reviewers that consistently give higher (or lower) marks than the others independently from the quality of the specific proposal they have to assess. Moreover, we show that our proposed unbiasing procedure can have a significant effect on the final result. This information and proposed unbiasing tool could be useful to the review chairs to improve the fairness of the review process
– Here we show that it is possible to minimize efforts without compromising quality by reducing the number of reviews on papers whose fate is clear after, for instance, only a couple of reviews.

2 Modeling and measuring peer reviews

This section describes a model for peer reviews and identifies metrics that help us assess properties of the reviews. We focus specifically on:

– modeling peer reviews of scientific papers.
– identifying metrics that help us *understand* and *improve* peer reviews. Specifically, we aim at a peer review process that allows (i) the selection of high *quality* papers and proposals , that is *fair*, and that is *efficient* (in terms of minimizing the time spent by *both* authors and reviewers).

In the subsequent sub-sections we will present and discuss appropriate definitions for what we mean by quality, fairness and efficiency. The model and metrics provided in this section are generally applicable to a variety of review processes, and can be used as a benchmark framework to evaluate peer reviews, also beyond the experimental results provided in this document.

2.1 Peer review model

This section formally defines and abstracts the peer review process. The level of abstraction and the aspects of peer reviews that are abstracted are driven by the goal of the work and the metrics we want to measure, discussed in the next section.

From a *process* perspective, peer reviews may vary from case to case, but usually they proceed along the following steps. Authors submit a set $\mathcal{C} = \{C_1, \ldots, C_n\}$ of *contributions*[1] for evaluation by a group $\mathcal{E}$ of *experts* (the peers, also called reviewers). Contributions are submitted during a certain time window, whose deadline is d_s. Each contribution is assigned to a number of reviewers and its flow through the process can be supervised by senior reviewers (a set $\mathcal{SR} \subset \mathcal{E}$ of distinguished experts that analyze reviews and help chairs take a final decision on the contribution). The assignment can be *continuous* (contributions are assigned as they come, as in the case of journals or some projects proposal) or in *batch*, and can be done in various ways (e.g., via bidding, based on topics, or based on decisions by chairs). In most processes, the number of reviewers N_R *initially* assigned to review or supervise each contribution is predefined and equal for all contributions. The typical settings for conferences is to have three reviewers and zero or one senior reviewer per paper. In the general case, each contribution may be assigned to a different number of reviewers.

The review occurs in one or more *phases*. We denote with N_P the total number of phases. In each phase p_k, contributions are assigned, marks are given, and contributions that are allowed to proceed to the next phase are selected. The next phase may or may not require authors to send a revised or incremental version of the contribution. At the end of each phase there is a discussion over the reviews (possibly involving author feedback). Some processes require the discussion to end in a "consensus" result for each of the marks. In all cases, the discussion results in a decision on whether each contribution is accepted or not. The entire process is supervised by a set $\mathcal{CH} \subset \mathcal{E}$ of *chairs*.

For example, the typical conference has a one-phase review, with discussion at the end leading to acceptance or rejection of each paper. Some conferences, such as *Sigmod*, have a 2-phase review process where in the first phase each paper is assigned to two reviewers and only papers that have at least one accept mark go to phase 2 and are then assigned to a third reviewer. This is done to minimize the time spent in reviewing. EU FET-Open proposals also follow a 2-phase process, where proposals are assigned to three reviewers in each phase and where authors send a revised and extended version of the proposal in phase 2. This is done to minimize both the reviewing effort and the proposal preparation effort (only a few groups can submit long proposals).

Given the above, we model a *phase* p_k of a peer review process as follows:

Definition 1. *A phase* $p = (\mathcal{C}, \mathcal{E}, \mathcal{M}, \pi, \gamma, \sigma, \rho, \mathcal{A})$ *of a peer review process consists of:*

- *a set* $\mathcal{C} = \{C_1, \ldots, C_n\}$ *of* contributions *submitted for evaluation;*
- *a set* $\mathcal{E}$ *of* experts*, which includes:*
  - *a set* $\mathcal{CH} \subset \mathcal{E}$ *of* chairs *that supervise the review process*

[1] Notice that in the rest of the paper we use both the terms contribution and paper with the same meaning

- *a set $\mathcal{SR} \subset \mathcal{E}$ of distinguished experts (sometimes called senior reviewers) that analyze reviews and help chairs take a final decision on the contribution*
- *a set $\mathcal{R} \subseteq \mathcal{E}$ of experts that act as reviewers of the contributions*

and s.t. $\mathcal{CH} \cup \mathcal{SR} \cup \mathcal{R} = \mathcal{E}$

- *a set $\mathcal{M}$ of mark sets, $=\{M^1, \ldots, M^q\}$, where for each mark set a total order relation $\leq$ always exists. For each mark set M^j there is a distinguished value called* acceptance threshold*, denoted by t^j.*
- *an assignment function $\pi : \mathcal{C} \longrightarrow \mathcal{P}(\mathcal{R}) \times \mathcal{P}(\mathcal{SR})$ assigning each contribution to a subset of the reviewers and a subset of the senior reviewers (an element of the respective powersets).*
- *a scoring function $\gamma : \{c, r\} \longrightarrow M^1 \times \ldots \times M^q$ such that $c \in \mathcal{C}$ and $r \in \pi(c)$. This function models the marks assigned by each reviewer.*
- *a score aggregation function $\sigma : M^1 \times \ldots \times M^q \longrightarrow \mathbb{N}$. This models the way in which in some review processes one can derive an aggregate final mark based on the individual marks.*
- *a ranking function $\rho : \{c\} \longrightarrow \mathbb{N}$*
- *a subset $\mathcal{A} \subset \mathcal{C}$ that denotes the accepted contributions.*

In the rest of the section we will refer to a simple example in order to explain the metrics used in the peer review evaluation process.

EXAMPLE 1. Consider a review process with:

- $|\mathcal{C}| = 10$ submitted contributions;
- $|\mathcal{E}| = |\mathcal{R}| = 5$ experts that act as reviewers and no chairs ($|\mathcal{CH}| = 0$) or senior reviewers ($|\mathcal{SR}| = 0$);
- a set of three criteria: $M_1 = quality$, $M_2 = implementation$, $M_3 = impact$, where for each criteria M_j the reviewer can assign a mark whose domain is the set of integer $\{1, \ldots, 10\}$;
- the number of contributions assigned to each reviewer $N_{RC} = 3$.

We next define metrics for peer review that are then computed over the review data in our possession. The purpose of such metrics is to help us understand and improve peer reviews under two dimensions: **quality** and **efficiency**. Informally, quality is related to the final result: a review process ensures quality if the best contributions are chosen. Efficiency is related to the time spent in preparing and assessing the contributions: a process is efficient if the best proposals are chosen with minimal time spent both by authors in preparing the contribution and by reviewers in performing the reviews. This is consistent with the main goals of peer reviews: selecting the best proposals with the least possible effort by the community. As we will see, both quality and efficiency are very difficult to measure. Notice that in this document we disregard other potential benefits of peer reviews, such as the actual content of the feedback and the value it holds for authors. We focus on metrics and therefore on what can be measured by looking at review data.

2.2 Quality-related metrics

In an ideal scenario, we would have an objective way to measure the quality of each contribution, to rank contributions or to, at least, select "acceptable" contributions from others. If this was the case, we could measure the quality of each peer review process execution, and identify which processes are more likely to be of high quality. Unfortunately (or fortunately, depending on how you see it) it turns out that quality is subjective, and there are no objective or widely accepted ways to measure quality. Nevertheless, we think it is possible to define metrics that are approximate indicators of quality (or lack thereof) in a review process and use them in place of the ideal "quality". An important subclass of these metrics are those related to *fairness*, which study whether the authors got a fair chance for their contribution to be accepted. In the next sub sections we explore the rationale behind a number of proposed quality-related metrics. We then detail their definitions and present some simple examples to illustrate their potential application.

Divergence from the ideal ranking In this metric we do assume that, somehow, we have the "correct" or "ideal" ranking for the contributions. We can assume or imagine that this can be conceptually measured in some way. For example, the ideal ranking could be the one each of us defines (in this case the comparison is subjective and so is the value for the metric), or we can define it as the ranking that we would have obtained if all experts reviewed all contributions (as opposed to only two or three reviewers).

In such a case we could try to assess how much the set of the actual accepted contributions differs from the set of the contributions that should have been accepted according to the ideal ranking. More formally, we can define a measure called **divergence**, in order to compute the distance between the two rankings, i.e., the *ideal* ranking and the *actual* ranking (the outcome of the review process). We next give the formal definition of divergence following Krapivin et al. [2], adapted to our scenario.

Definition 2 (phase quality divergence). *Let $\mathcal{C}$ be a set of submitted contributions, $n = |\mathcal{C}|$ the number of submissions, ρ_i and ρ_a, respectively, the ideal ranking and the actual ranking, and t the number of accepted contributions according to the actual ranking. We call divergence of the two rankings $Div_{\rho_i,\rho_a}(t, n, \mathcal{C})$ the number of elements ranked in the top t by ρ_i that are not among the top t in ρ_a.*
The normalized divergence $NDiv_{\rho_i,\rho_a}(t, n, \mathcal{C})$ is equal to $\frac{Div_{\rho_i,\rho_a}(t,n,\mathcal{C})}{t}$, and varies between 0 and 1.

Therefore, through this metric it is possible to assess how much the set of the actual accepted contributions *diverges* from the set of contributions ranked w.r.t. the ideal quality measure, and so how many contributions are "rightly" in the set of the *accepted contributions* and how many contributions are not.

One can also compute the divergence for each reviewer, considering, instead of the whole set of contributions, only those contributions rated by a particular reviewer, i n order to assess, instead of the quality of the entire process, the quality of a specific reviewer.

The definition is analogous to the above one.

Definition 3 (reviewer quality divergence). *The reviewer quality divergence* $Div_{\rho_i,\rho_a}(t_r, n_r, \mathcal{C}_r)$ *is defined analogously to the phase quality divergence with the difference that we restrict the divergence computation to the set* $\mathcal{C}_r$ *of contributions reviewed by reviewer* r *instead of the entire set of submissions* $\mathcal{C}$ *(so,* $n_r = |\mathcal{C}_r|$ *is the number of contributions reviewed by* r*, accepted and rejected ones, and ranked, and* t_r *is the number of contributions reviewed by* r *whose score is equal or greater than the contribution in the accepted set with the lowest score), as per the scoring function.*
The normalized reviewer quality divergence $NDiv_{\rho_i,\rho_a}(t_r, n_r, \mathcal{C}_r)$ *is given by* $NDiv_{\rho_i,\rho_a}(t_r, n_r, \mathcal{C}_r) = \frac{Div_{\rho_i,\rho_a}(t_r,n_r,\mathcal{C}_r)}{t_r}$.

Divergence a posteriori One may argue that quality can be measured *a posteriori* (years after the completion of the review process), by measuring the impact that the paper had, for example by counting citations. Hence, we can compare this ranking a posteriori with the one coming out of the review process. However, we can do this only for accepted papers, since for rejected ones we do not have a way to assess their impact (they have not been published, or at least not in the same version as they were submitted).

As done before, we could use the divergence measure, using as metrics the *citation-based* estimates and the *actual* ranking of contributions, but restricting the analysis to the set of accepted contributions $\mathcal{A}$ instead of $\mathcal{C}$, as only for those we have the two rankings. In this case, in the divergence formula, t would be equal to n and divergence would be always zero. We can still use divergence as an analysis mean, but only if we restrict to, say, examining the difference in the ranking in the top k contributions, with $k < t$.

For instance, following Example 1, in Figure 1 we can compute the divergence with respect to the top contributions. In Fig.1a we show the example data set. Contributions are ranked with respect to the *actual* ranking (the one resulting from the review process). For each contribution we show: (i) the overall score given by the reviewers, computed as the average over all the three criteria, (ii) the number of citations that the contribution received (say after 3 years from publication), (iii) the actual ranking, (iv) the citation-based ranking, (v) the outcome (accepted/rejected) for each paper.

In Fig.1b we compute how many elements in the top t are different between the two rankings. For instance, we compute for $t = 3$: $NDiv_{\rho_c,\rho_a}(3, 8, 8) = 0.66$, with $\rho_c =$ Citation-based ranking, $\rho_a =$ Actual ranking, and for $t = 5$: $NDiv_{\rho_c,\rho_a}(5, 8, 8) = 0.2$. Please note that by definition of the divergenge metric, we have $NDiv_{\rho_c,\rho_a}(8, 8, 8) = 0.0$. We want to underline that the result for $t = 3$ indicates a high divergence between the two rankings. It means that 66% of the times the two rankings are different and this would lead to a 66% difference in the probability of acceptance of a contribution if we would accept 3 out of 8 contributions, which corresponds to ca. 37% acceptance ratio, i.e., a typical acceptance ratio. This mean that we are doing wrong choices while selecting contributions in the 66% of the cases, which is a huge value. Probably, extending this analysis to a greater number of contributions (say, 30 out of 80), the divergence value will be lower, still, one should compute which is an *acceptable*, or, say, *ideal* divergence value that could be accepted for a peer review process, in order to consider the review process a quality review process, that is a process that gives different results

w.r.t. a simple random selection of the contributions. This is the focus of our current investigation; moreover, once such a value has been computed, one can try to estimate the effort required w.r.t. this *ideal* divergence value (see Section 2.3).

Another useful metric we can use in our evaluations is the Kendall τ distance, the typical metric for measuring a difference between two rankings [1]. The Kendall τ distance - differently from the divergence which computes only the number of elements that differs between two sets - computes the difference in the position of the elements between two sets. Therefore is more useful than the divergence measure when the two sets to compare have the same number of elements, as when the two sets to compare are *only* the accepted contributions ranked using the actual ranking and the citation-based ranking.

Definition 4 (Kendall τ distance). *Let ρ_1 and ρ_2 be two different rankings, the Kendall τ distance is measured as the number of steps needed to sort bi-ranked items so that any pair A and B in the two rankings will satisfy the condition:*

$$sign(\rho_1(A) - \rho_1(B)) = sign(\rho_2(A) - \rho_2(B))$$

In our setting, given the actual ranking ρ_a and the citation-based ranking ρ_c we can compute the Kendall τ distance for any pair of contributions c_i and c_j in the set of the accepted contributions $\mathcal{A}$, satisfying the condition:

$$sign(\rho_a(c_i) - \rho_a(c_j)) = sign(\rho_c(c_i) - \rho_c(c_j))$$

As before, while the above definition is given for the whole process, an analogous metric can be defined to analyze each single reviewer (as long as a reviewer had to rank/rate contributions that have been accepted and therefore have an associated citation count). In this case the Kendall τ distance is computed for any pair of contributions c_i and c_j in the set of the accepted contributions reviewed by r.

The Kendall τ distance is typically normalized (NK_τ) by dividing it by $\frac{n(n-1)}{2}$ (which is the maximum value, corresponding to lists ordered in the opposite way), so that the distance is in the [0,1] interval.

Referring to Example 1, we can compute the normalized Kendall τ distance NK_τ for the set of contributions (see Fig.1a), taking into account the two rankings: the actual ranking (*Rank(Avg review mark)*) and the citation-based ranking (*Rank(Citations)*). In particular, we have that:

$$NK_\tau = 0.43$$

The values of the normalized Kendall τ distance are always in $0 \leq NK_\tau \leq 1$. If $NK_\tau = 1$ this means that the ranked items are always in a different position with respect to the two rankings, while if $NK_\tau = 0$ the two rankings produce the same ordered list. In this case the normalized distance is $NK_\tau = 0.43$, meaning that the review process in the example has performed rather poorly with respect to the selected *a posteriori* quality measure based on citations.

Paper ID	Avg review mark	Citations	Rank(Avg review mark)	Rank(Citations)	Outcome
1	9.89	3	1	7	accepted
2	9.33	20	2	5	accepted
3	9.00	126	3	1	accepted
4	9.00	98	3	2	accepted
5	8.56	54	5	4	accepted
6	8.00	20	6	5	accepted
7	7.11	70	7	3	accepted
8	7.00	1	8	8	accepted
9	6.89	Not Avail.	9	Not Avail.	rejected
10	1.89	Not Avail.	10	Not Avail.	rejected
				Kendall τ distance = 0.43	
				Acc. Threshold = Avg mark >= 7	

(a) Available data set

Paper ID	Avg review mark	Citations	Rank(Avg review mark)	Rank(Citations)
1	9.89	3	1	7
2	9.33	20	2	5
3	9.00	126	3	1

$NDiv_{\rho c,\rho a}(3, 8, 8) = 0.66$

Paper ID	Avg review mark	Citations	Rank(Avg review mark)	Rank(Citations)
1	9.89	3	1	7
2	9.33	20	2	5
3	9.00	126	3	1
4	9.00	98	3	2
5	8.56	54	5	4

$NDiv_{\rho c,\rho a}(5, 8, 8) = 0.2$

Paper ID	Avg review mark	Citations	Rank(Avg review mark)	Rank(Citations)
1	9.89	3	1	7
2	9.33	20	2	5
3	9.00	126	3	1
4	9.00	98	3	2
5	8.56	54	5	4
6	8.00	20	6	5
7	7.11	70	7	3
8	7.00	1	8	8

$NDiv_{\rho c,\rho a}(8, 8, 8) = 0$

(b) Normalized divergence mesured a posteriori

Fig. 1: Normalized divergence: sample of 8 accepted contributions ($n = 8$), with respect to the top ranked ($t = 3$ and $t = 5$)

An ongoing analysis we are performing is on to try to estimate the Kendall τ trend, that is try to predict the value of K_τ for all contributions (accepted and not accepted ones), based on the value and distribution of the K_τ for the set of accepted contributions.

Another useful metric is the Kendall τ rank correlation coefficient [1] that measures the degree of correspondence between two rankings, so assessing the significance of this correspondence.

Definition 5 (Kendall τ rank correlation coefficient). *Let n_c be the number of concordant pairs and n_d the number of discordant pairs in the data set, and n the total number of contributions, we define the Kendall τ rank correlation coefficient as follow:*

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

The coefficient τ can assume the following values:

- $\tau = 1$, the two rankings are the same;
- $\tau = -1$, one ranking is exactly the reverse of the other;
- $\tau = 0$, the two rankings are completely indipendent.

Obviously, for values such that $0 \leq \tau \leq 1$, increasing values of τ imply increasing agreement between the two rankings, while for $-1 \leq \tau \leq 0$ decreasing values of τ imply increasing disagreement between the two rankings.

Disagreement between referees Through this metric we compute the similarity between the marks given by the reviewers on the same contribution. In particular, given a specific criterion j, we compute how much the marks of a reviewer i differ from the marks of the other $r-1$ reviewers (Definition 6), Then we compute for a specific criterion j the disagreement of a reviewer i with respect to the others over the whole set of contributions (Definition 7), and, finally, over all the criteria (Definition 8).

The rationale behind this metric is that in a review process we expect a high agreement between reviewers. It is natural that reviewers have different opinions on contributions. However, if the marks given by reviewers are comparable to marks given at random and have high disagreement, then the results of the review process are also random, which defeats the purpose. The reasons for having reviewers (and specifically for having 3 reviewers, the typical number) is to evaluate based on consensus or majority opinion.

Intuitively, assume that we consider the ideal ranking to be the one that we would obtain by having all experts review all papers, and that we assigned each contribution to only 3 reviewers to make the review load manageable. If the disagreement is high on most or all contributions, we cannot hope that the opinion of 3 reviewers will be a reasonable approximation or estimate for the ideal ranking. We will get back to this issue when discussing the quality vs effort trade-off.

In order to measure the disagreement, we compute the euclidean distance between the marks given by different reviewers on the same contribution.

Definition 6 (Disagreement of a reviewer on a criterion). *Let j be a criterion and $M^j_{i_z}$ be the mark set by the reviewer i for the criterion j assigned to a contribution*

z. Then, a disagreement $\phi^j_{i_z}$ among r_z reviewers on a contribution z is the euclidean distance between the mark given by the reviewer i, and the average $\mu^j_{i_z}$ of those given by the others $r-1$ reviewers:

$$\phi^{\mathrm{j}}_{\mathrm{i_z}} = \mid \mathrm{M}^{\mathrm{j}}_{\mathrm{i_z}} \text{-} \mu^{\mathrm{j}}_{\mathrm{i_z}} \mid \tag{1}$$

with:

$$\mu^{\mathrm{j}}_{\mathrm{i_z}} = \frac{1}{\mathrm{r_z}} \cdot \sum_{\mathrm{k} \neq \mathrm{i_z}}^{\mathrm{r_z}} \mathrm{M}^{\mathrm{j}}_{\mathrm{k_z}} \tag{2}$$

Definition 7 (Disagreement of a review phase on a criterion). *Let n be the number of the contributions and r_z be the number of reviewers assigned to a contribution z, then the disagreement over all contributions on a criterion j is the average disagreement:*

$$\Phi^{\mathrm{j}} = \frac{1}{\mathrm{n}} \cdot \sum_{\mathrm{z}=1}^{\mathrm{n}} \cdot \frac{1}{\mathrm{r_z}} \sum_{\mathrm{k}=1}^{\mathrm{r_z}} \phi^{\mathrm{j}}_{\mathrm{k_z}} \tag{3}$$

Definition 8 (Disagreement of a review phase). *Let q be the number of criteria in a review phase, then the disagreement over all the criteria is:*

$$\Psi = \frac{1}{\mathrm{q}} \cdot \sum_{\mathrm{j}=1}^{\mathrm{q}} \Phi^{\mathrm{j}} \tag{4}$$

Analogously to the above definitions and via obvious extensions we can also define:

Definition 9 (Disagreement of a reviewer on a contribution). *Let q be the number of criteria in a review phase, then the disagreement of a reviewer i on a contribution z is:*

$$\gamma_{\mathrm{i_z}} = \frac{1}{\mathrm{q}} \cdot \sum_{\mathrm{j}=1}^{\mathrm{q}} \phi^{\mathrm{j}}_{\mathrm{i_z}} \tag{5}$$

Definition 10 (Average disagreement of a reviewer). *Let Z be the number of contributions a reviewed by a reviewer i, then the average disagreement of a reviewer i is:*

$$\overline{\gamma}_{\mathrm{i}} = \frac{1}{\mathrm{Z}} \cdot \sum_{\mathrm{z}=1}^{\mathrm{Z}} \gamma_{\mathrm{i_z}} \tag{6}$$

Definition 11 (Disagreement of a review phase on a contribution). *Let r_z be the number of reviewers in a review phase on a contribution z, then the disagreement of a review phase on a contribution is:*

$$\Gamma_{\mathrm{z}} = \frac{1}{\mathrm{r_z}} \cdot \sum_{\mathrm{i}=1}^{\mathrm{r_z}} \gamma_{\mathrm{i_z}} \tag{7}$$

To illustrate operatively the disagreement metric, we use again the data from Example 1 in Figure 2 and compute the disagreement for each paper and each reviewer, w.r.t. the the others two reviewers. For each paper (Paper ID) we have (i) the marks given by

each reviewer for a given criterion (M_1, M_2, M_3), (ii) the average score obtained by each contribution after the review phase (*Avg review mark*)), (iii) the reviewers that have had that contribution in charge, (iv) the disagreement value computed for each reviewer and over all the criteria, (v) the average disagreement per paper and (vi) the normalized average disagreement for paper.

The minimum average disagreement among reviewers is zero, while - in this specific set - the maximum potential average disagreement is 6, corresponding to the following mark assignment: M_1(1,1,1), M_2(10,10,10), M_3(10,10,10). Notice that for paper ID=2 the disagreement is zero, while the highest disagreement pertains to paper ID=9. Moreover, we can compute the average disagreement of the overall (specific) review process, as well as its variance (absolute and normalized values are included in Figure 2).

Paper ID	M_1	M_2	M_3	Avg review mark	Reviewers ID	Disagreement(r_x,r_y,r_z)	Avg disagreement per paper	Norm. avg disagreement per paper
1	{10,10,10}	{10,10,10}	{9,10,10}	9.89	{r_1,r_2,r_3}	{0.16,0.16,0.33}	0.21	0.04
2	{10,10,8}	{10,10,8}	{10,10,8}	9.33	{r_1,r_2,r_5}	{0,0,0}	0	0
3	{9,9,9}	{8,8,8}	{10,10,10}	9.00	{r_1,r_2,r_3}	{0,1.5,1.5}	1	0.17
4	{9,10,10}	{8,8,8}	{9,9,10}	9.00	{r_1,r_2,r_4}	{1,1.5,0.5}	1	0.17
5	{9,8,9}	{10,6,8}	{8,10,9}	8.56	{r_2,r_3,r_4}	{0.16,1.83,0.66}	0.88	0.15
6	{7,8,7}	{9,8,9}	{8,8,8}	8.00	{r_2,r_3,r_4}	{0.5,1,0}	0.50	0.08
7	{6,6,6}	{7,7,7}	{8,9,8}	7.11	{r_3,r_4,r_5}	{1.66,0.16,1.83}	1.22	0.20
8	{6,6,6}	{8,7,7}	{7,9,7}	7.00	{r_1,r_4,r_5}	{1.5,0.83,1}	1.11	0.19
9	{5,1,3}	{6,10,8}	{10,10,9}	6.89	{r_2,r_4,r_5}	{5.83,2.66,4.16}	4.22	0.70
10	{1,2,1}	{5,2,3}	{1,1,1}	1.89	{r_1,r_4,r_5}	{1.16,2.16,1.33}	1.55	0.26
							AVG = 1.17	**AVG = 0.19**
							STD. ERR. = 0.37	**STD. ERR. = 0.06**

(a) Average disagreement and normalized disagreement per paper

Reviewer ID	Papers reviewed	Avg disagreement per reviewer	Norm. avg disagreement per reviewer
r_1	{1,2,3,4,8,10}	0.64	0.07
r_2	{1,2,3,4,5,6,9}	1.38	0.15
r_3	{1,3,5,6,7}	1.26	0.14
r_4	{4,5,6,7,8,9,10}	1	0.11
r_5	{2,7,8,9,10}	1.66	0.18
			AVG = 0.13
			DEV. STD. = 0.04

(b) Average disagreement and normalized disagreement per reviewer

Fig. 2: (Normalized) disagreement calculated on the data set

If the disagreement value, given three reviewers for each contribution, is high this means that, probably, three reviewers are not enough to ensure a high quality review process. Indeed, having a high disagreement value means, in some way, that the judg-

ment of the three peers is not sufficient to state the value of the contribution itself. So this metric could be useful to improve the quality of the review process as could help to decide, based on the disagreement value, if three reviewers are enough to judge a contribution or if more reviewers are needed in order to ensure the quality of the process.

Robustness A mark variation sensitivity analysis is useful in order to assess if a slight modification on the value of marks could bring a change in the final decision about the acceptance or rejection of a contribution.

The rationale behind this metric is that we expect the review process to be *robust* to minor variations in one of the marks. In a way we can see it as yet another measure of randomness in the review process. When reviewers need to select between, say, a 1-10 score for a set of metrics, often they are in doubt and perhaps somewhat carelessly decide between a 7 or an 8 (not to mention the problem of different reviewers having different scoring standards, discussed next). With this metric we try to assess how much precision is important in the mark assignment process, i.e., how much a slight difference in the mark value can affect the final (positive or negative) assessment of a contribution. To this end, we apply a small positive/negative variation ϵ to each marks (e.g., ± 0.5), and then rank the contributions with respect to these new marks. Assuming that we accept the top t contributions, we then compute the divergence among the two rankings in order to discover how much the set of accepted contributions differs after applying such a variation.

Definition 12 (naive ϵ-divergence). *Let ρ_a be the actual ranking and ρ_s be the ranking after the mark variation phase, $\mathcal{C}$ a set of contributions, $n = |\mathcal{C}|$ the number of contributions submitted and ranked according to ρ_a and ρ_s, and $t = |\mathcal{C}_{\mathcal{A}}|$ the number of actually accepted contributions, we call divergence of the two rankings $Div_{\rho_a,\rho_s}(t, n, \mathcal{C})$ the number of elements that differ between the two sets (or, t minus the number of elements that are equal)*

The higher the divergence, the lower the robustness. Intuitively, what we do with the mark variation is a a naive way to transform a mark into a random variable with a certain variance, reflecting the indecision of a reviewer on a mark. Reviewers do have indecision among similar marks especially if the number of possible marks for a criterion is high. Informally, if we assume that each reviewer has an indecision area for the mark (e.g., in case of a 1-10 scoring range, we assume that a reviewer has a hard time discriminating between, say, a 6 and a 7), when the process is not robust to some marks differing of 1 from the actual ones, then it means that a more or less random decision between a 6 and a 7 given by a reviewer can determine the fate of the contribution. Notice that a process that is not robust is not necessarily low quality. For example, if for a criterion we have only two marks, 0 and 1, it is natural that the process is not robust but the granularity of the mark set is so coarse that it is unlikely for the reviewer to be undecided. The problems arise when the granularity is fine and the robustness is low.

To illustrate operatively the mark sensitivity analysis, we use again the data from Example 1. In Fig. 3a we modify the marks given to one of the contributions taken from the data set of Example 1, by subtracting $\epsilon = 1$ in turn to each of the three marks given

Paper ID	Original Avg review mark	Outcome	Δ	Modified Avg review mark	Modified outcome		
8	7	accepted	-1	**6.89**	rejected		
9	6.89	rejected	+1	**7**	accepted		
					Acc. Threshold = Avg mark >= 7		

(a) Example of perturbation $\epsilon = \pm 1$ applied to two papers from the data set

Paper ID	Avg review mark	Avg modified review mark	Rank(Actual review mark)	Rank(Modified review mark)
1	9.89	9.78	1	1
2	9.33	9.22	2	2
3	9.00	8.89	3	3
4	9.00	8.89	3	3
5	8.56	8.44	5	5
6	8.00	7.89	6	6
7	7.11	7.00	7	7
8	7.00	6.89	8	9
9	6.89	7.00	9	8
10	1.89	2.00	10	10
$NDiv_{\rho m,\rho a}(8, 10, 10) = 0.12$				

(b) Normalized divergence calculated between the rankings obtained before and after the perturbation ϵ, where $\epsilon = -1$ for accepted papers and $\epsilon = +1$ for rejected papers

Fig. 3: Sensitivity computed for the data set

by the three reviewers. Then, we fix an acceptance threshold and see how the acceptance changes according to the small modification we made to the marks assigned. In our simple case, changing the average review mark given to paper with ID=8 and paper with ID=9 also changes the acceptance of them. In fact, in this case we applied $\epsilon = -1$ to paper with ID=8 and $\epsilon = +1$ to paper with ID=9 and we see that the recomputed average essentially swaps them in the ranking. So, in the perturbated case we have that paper with ID=8, which was accepted, is now rejected and, conversely, paper with ID=9, which was rejected, is now accepted.

This computation can be extended to all papers and the results for our data from Example 1 (see Fig. 3b). We applied such modification to the entire data set of papers, subtracting $\epsilon = -1$ to the accepted papers and adding $\epsilon = +1$ to the two papers that were rejected. After we computed the new ranking obtained from the perturbation ϵ, we calculate the normalized divergence $NDiv_{\rho_m,\rho_a}(8, 10, 10)$ between the rankings obtained with and without the ϵ modification. The normalized divergence we found actually tells us that the ranking has been slightly modified and in particular that the set of accepted papers is changed.

The statistical version of the above metric is the ϵ-robustness. With this metric, we replace each mark m with a random variable M uniformly distributed between $m - \epsilon$ and $m + \epsilon$ (this distribution represent the reviewer's uncertainty). When we do this, rankings become stochastic, so each ranking ρ has a certain probability of occurring. We call this stochastic ranking deriving from the perturbation of the marking as ρ_ϵ.

Definition 13 (ϵ-divergence). *Let ρ_a be the actual ranking and ρ_ϵ the random variable defining rankings after perturbations. Let $\mathcal{C}$ be a set of contributions, $n = |\mathcal{C}|$ the number of contributions submitted, and $t = |\mathcal{C}_\mathcal{A}|$ the number of actually accepted contributions. We call* stochastic divergence *of the two rankings $Div_{\rho_a,\rho_\epsilon}(t, n, \mathcal{C})$ the random variable denoting the probability that the divergence between ρ_a and the (random) ranking ρ_ϵ has a given value ϵ.*

Definition 14 ($\delta\pi$-robustness). *We say that a ranking is ϵ-robust if the probability of the ϵ-divergence being less than δ is equal or greater than π.*

In this way we can compute the robustness of the process and see how much the process is sensitive to the mark variation.

Fairness-related metrics A property that, we argue, characterizes a "good" review process, and specifically the assignment of contributions to reviewers, is *fairness*.

A review process is *fair* if and only if the acceptance of a contribution does not depend on the particular set of reviewers that assesses it among the set of experts E. Formally, assume that we are able to compute, at the start of the review process (after the contributions have been submitted but before they are assigned) the probability of a contribution c being accepted. We denote this probability as $P^c_{accept}(t_{start})$. Assume now that we can compute the same probability after the contributions have been assigned to reviewers. We denote this probability as $P^c_{accept}(t_{assign})$.

We define an assignment process as *fair* with respect to contribution c iff:

$$P^c_{accept}(t_{start}) = P^c_{accept}(t_{assign})$$

In other words, an assignment is unfair if the reviewers selected for contribution c give marks which are different than what a randomly selected set of reviewers (among the committee members) would give.

Correspondingly, we define a peer review process as fair iff:

$$P^c_{accept}(t_{start}) = P^c_{accept}(t_{end}) \quad \forall C \in \mathcal{C}$$

where $P^c_{accept}(t_{end})$ is the probability of the final acceptance of the contribution [2].

The problem with unfair assignments is that *the assignment* is affecting or determining the fate of the paper: a different assignment would have yielded a different result. Again, to a certain extent, this is normal, natural, and accepted: different reviewers do have different opinions. The problem we are trying to uncover is when reviewers are *biased* in various ways with respect to their peers. For example, a common situation is the one in which a reviewer is consistently giving lower marks with respect to the other reviewers for the same contributions, perhaps because he or she demands higher quality standards from the submission.

Our aim is not to try to identify what is the appropriate quality standard or to state that reviewers should or should not be tough. However, if different reviewers have different quality standards, when a contribution has the "bad luck" of being assigned to one such tough reviewer, the chances of acceptance are lower. This has nothing to do with the relative quality of the paper with respect to the other submissions, it is merely a biasing introduced by the assignment process and by the nature of the reviewer, that is rating contributions using, de facto, a different scale than the other reviewers. Fairness metrics try to identify, measure, and expose the most significant biases so that the chair can decide if they indeed correspond to unfair review results that need to be compensated before taking the final decision. As such they can be indicators of quality but also can provide hints to the chairs to compensate quality problems. In the following we illustrate different kind of biases and define a metric to discover biased reviewers.

- **Rating bias**: Reviewers are positively biased if they *consistently* give higher marks than their colleagues who are reviewing the same proposal. The same definition applies for the opposite case, when we talk about negatively biased reviewers. We also refer to these two cases as *accepting behavior* and *rejecting behavior*.
- **Affiliation bias**: it is a rating bias but computed over contributions written by authors with certain affiliations.
- **Topic bias**: a rating bias but computed over contributions concerning certain topics or research areas.
- **Country bias**: a rating bias but computed over contributions whose authors are coming from a certain country or continent (for instance, American reviewers with respect to papers written by European authors).
- **Gender bias**: a rating bias but computed over contributions written by authors of a certain gender.

[2] Between t_{assign} and t_{end} the probability could change because the contribution has been revised, a senior reviewer entered the process, an author feedback is provided, or a de-bias phase is performed.

– **Clique bias**: a rating bias but computed over contributions written by authors which belong to the same clique inside a certain research community.

The way to compute the bias value is very similar to that described for the disagreement metric (see Definition 6), the difference is that the domain may be restricted to papers with certain topics or affiliations (depending on the kind of bias we are looking at), and that the sign of the disagreement coefficient ϕ_i^j has been preserved, basically replacing equation (1) with the following one:

$$\phi_{\mathrm{i}}^{\mathrm{j}} = \mathrm{M}_{\mathrm{i}}^{\mathrm{j}} \text{-} \mu_{\mathrm{i}}^{\mathrm{j}} \tag{8}$$

This time the sign of the equation is important in order to discover positive or negative biases. Indeed, if the value of ϕ_i^j is constantly positive, this means the reviewer tends to give always higher marks with respect to other reviewers; while if the value of ϕ_i^j is constantly negative then the reviewer tends to give always more negative marks than other reviewers.

A variation on the rating bias is the **variance bias**, which occurs when a reviewer always gives marks that are very close to (or far from) the threshold for a mark. This is computed by simply calculating the variance of the mark.

As for the disagreement metrics, there are several scopes to which we can apply the bias metric. For example, we can measure the bias for a single reviewer and for a particular criterion, the bias over a review phase, and the bias over all the criteria. In the experiments we will assess the biases on actual review data and discuss how they can be compensated.

In Fig. 4 we first report the computed bias per reviewer per paper for the data of Example 1 and then the computed average bias per reviewer in which we highlighted the *most accepting* behavior and the *most rejecting* behavior. This information could be useful to a Conference Chair to improve the fairness of the review process.

We now address the difference between the actual ranking and the ranking obtained by compensating the biases. To compensate, we modify the marks by adding or removing the bias values so that on average the overall bias of the most biased reviewers is reduced. In particular, we take all reviewers r that have a bias greater than b and that have done a number of reviews higher than nr, and subtract b from all marks of r (or from the top-k biased reviewers). We call this ranking *debiased ranking* $\rho_{b,nr}$.

Definition 15 (bias compensation divergence). *It is the value* $Div_{\rho_a,\rho_{b,nr}}(t,n,\mathcal{C})$

2.3 Efficiency-related metrics

Efficiency refers to the effort spent in determining which contributions are accepted, and in particular the trade-off between effort and quality of the review process. It considers both the effort in writing contributions and in reviewing them.

Assumptions The basic working assumption of this section is that the quality-effort trade-off exists and that, in general, if a paper or proposal is *long*, and is reviewed very *carefully* by a *large* number of reviewers (all the ones the chairs consider to be experts),

Paper ID	M_1	M_2	M_3	Avg review mark	Reviewers ID	Bias(r_x,r_y,r_z)
1	{10,10,10}	{10,10,10}	{9,10,10}	9.89	{r_1,r_2,r_3}	{0.16,0.16,-0.33}
2	{10,10,8}	{10,10,8}	{10,10,8}	9.33	{r_1,r_2,r_5}	{0,0,0}
3	{9,9,9}	{8,8,8}	{10,10,10}	9.00	{r_1,r_2,r_3}	{0.5,-1.5,1.5}
4	{9,10,10}	{8,8,8}	{9,9,10}	9.00	{r_1,r_2,r_4}	{1,-1.5,0.5}
5	{9,8,9}	{10,6,8}	{8,10,9}	8.56	{r_2,r_3,r_4}	{0.16,-0.33,0.66}
6	{7,8,7}	{9,8,9}	{8,8,8}	8.00	{r_2,r_3,r_4}	{-1,1,0}
7	{6,6,6}	{7,7,7}	{8,9,8}	7.11	{r_3,r_4,r_5}	{-1.66,-0.16,1.5}
8	{6,6,6}	{8,7,7}	{7,9,7}	7.00	{r_1,r_4,r_5}	{-1.5,0.16,1}
9	{5,1,3}	{6,10,8}	{10,10,9}	6.89	{r_2,r_4,r_5}	{-5.83,1.66,4.16}
10	{1,2,1}	{5,2,3}	{1,1,1}	1.89	{r_1,r_4,r_5}	{-1.16,2.16,-1.33}

(a) Bias per paper

Reviewer ID	Papers reviewed	Avg bias per reviewer	Norm. avg bias per reviewer
r_1	{1,2,3,4,8,10}	-0.16	-0.02
r_2	{1,2,3,4,5,6,9}	-1.36	-0.15
r_3	{1,3,5,6,7}	0.04	0.00
r_4	{4,5,6,7,8,9,10}	0.71	0.08
r_5	{2,7,8,9,10}	1.07	0.12
			AVG = 0.13
			STD. ERR. = 0.05

(b) Bias per reviewer

Fig. 4: Bias computed for the data set

the selection is more informed than the case in which, say, one page proposal is briefly looked at by a couple of reviewers. Time is a precious resource, so the challenge is how to reduce the time spent while maintaining a "good" selection process that indeed selects the "best" proposals. A separate issue that we do not address (also as it is hard to measure) is the fact that a process is affected by the quality of the reviewers and the amount of discussion or the presence of a face to face discussion. For now we limit to metrics that we can derive from review data or from simple surveys. We also do not discuss here what could be a somewhat opposing, but intriguing argument that peer review tends to favor incremental paper rather than breakthroughs and that therefore peer review sometimes kills innovation. Assessing this is among our current research efforts, but is not discussed in this document.

In the following we identify metrics that can help us understand if the review process is efficient. The reviewing effort of a review phase is the total number of reviews N_R multiplied by the average time $\overline{t_r}$ (e.g., measured in person-hours) spent per review in that phase:

$$E_R = N_R \cdot \overline{t_r}$$

Correspondingly, the contribution preparation effort is the number of submissions multiplied by the average time spent in preparing each submission:

$$E_W = N_C \cdot \overline{t_w}$$

Reviews and submissions can span across N_P phases. For simplicity, in the above definitions and in this section we use the average reviewing or writing time instead of considering the time spent by each reviewer or author and the fact that different phases may require different reviewing or writing effort per contribution. We also assume that the set of experts is the same for all phases. The extension of the reasoning done here to remove these assumptions is straightforward.

In the ideal case from a quality perspective, all reviewers read all contributions for as long as they need to take a decision, and contributions are as long as they need to be for the reviewers to fully grasp their value. With respect to the review time and contribution length, we assume in particular that as the review time and contribution length grow, the reviewer is able to narrow down the *uncertainty/error* on the review marks he or she wants to give. In other words, it will increase the confidence that the correct mark for the contribution is within a given interval. This is graphically depicted in Figure 5 and 6 where we schematically plot the mark error function σ_r as a function of time and length respectively.

These schematic figures also show that beyond a certain time threshold t_{rx} and length threshold l_x the mark uncertainty remains constant. Reading a 10 pages paper for 4 hours or 4 days is not likely to make a difference (if we are in doubt between giving a 6 and a 7 we will probably still be in doubt), but one minute versus four hours does.

In summary, the ideal process from a quality perspective is as follows:

- The number of reviews $N_R = N_{RC} = |\mathcal{R}| * |\mathcal{C}|$, that is, the number of reviewers multiplied by the number of papers (all the reviewers rate all the contributions);

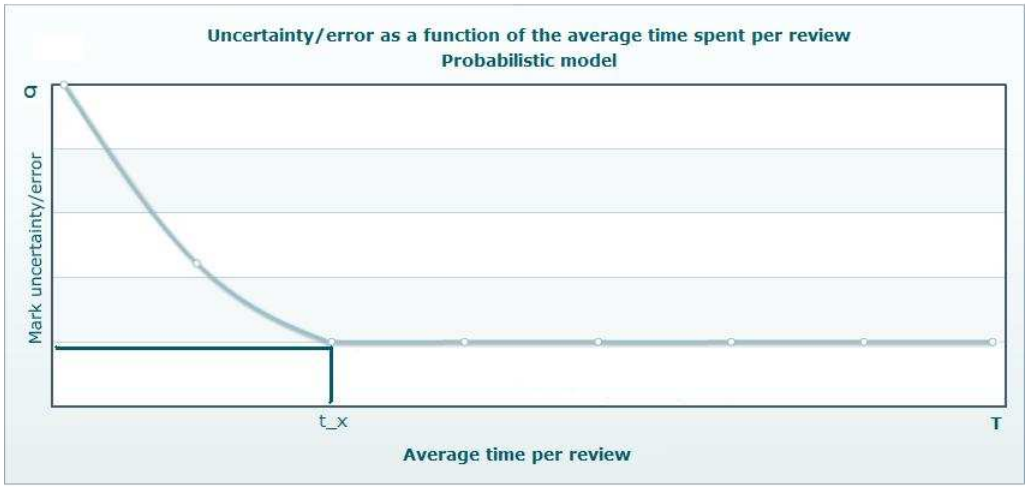

Fig. 5: Given a constant effort, this is a possible model of the uncertainty of a review with respect to the average time spent by the reviewers

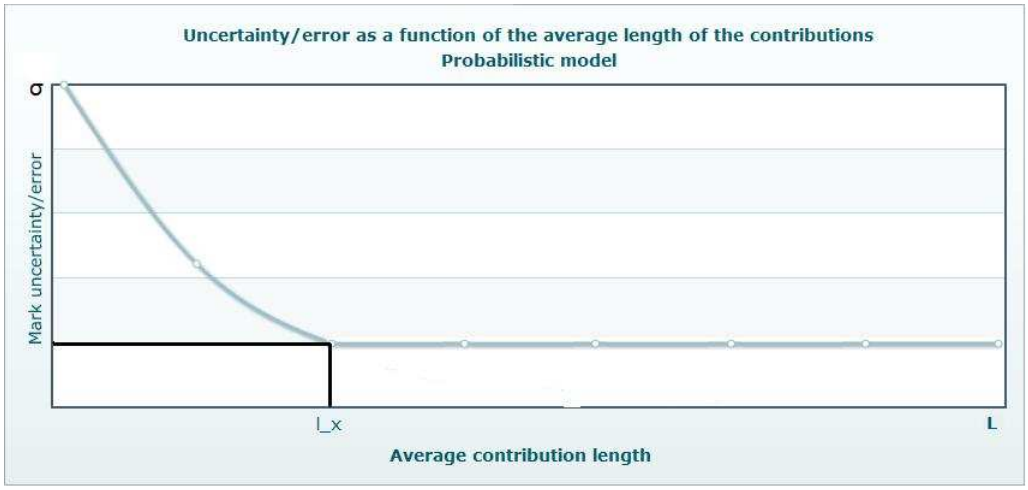

Fig. 6: Given a constant effort, this is a possible model of the uncertainty of a review with respect to the average length of the contributions

– The average time spent in reviewing $\overline{t_r} = t_{rx}$, where t_{rx} is the time that minimizes the error in the review process;
– The contribution length $l_c = l_x$, where l_x is the length that minimizes the error in the review process (and the corresponding effort is t_{lx}).

The reviewing effort is $E_R = t_{rx} * N_{RC}$. The writing effort is $E_W = t_{lx} \cdot N_C$

Informally, making the review process efficient requires reducing the effort minimizing the quality degradation. To this end we can act on the following parameters:

– the time devoted to each review or the length of the contribution (we assume the latter to be somewhat correlated to the writing effort);
– the number of reviewers per paper and the number of papers per reviewers;
– the number of phases, with the aim of i) putting less effort on submissions that are clearly good or clearly bad and more effort on borderline submissions, and ii) avoiding multi-phase processes that require update of the submission at each phase, if we recognize that having the extra phases is not beneficial to the process.

We now see some metrics that we think will help us understand proper values for these parameters.

Reducing the number of reviews A line of investigation is around reducing the number of reviews, and specifically reducing the number of reviews for submissions whose fate is clear. This is the main reasons for having phases, where at each phase focus and effort is placed on the submissions for which a decision has not been taken yet. We consider now in particular the case in which the submission remains unchanged from a phase to the next (as in the Sigmod example discussed earlier).

Assume that the review process is structured in as many phases as the maximum number of reviewers per papers (say, we plan to have at most four reviews for a paper, so at most four phases). The analysis we want to make is to understand which is the earliest phase at which we can stop reviewing a given paper, because we have a sufficiently good approximation of the fate of the paper, which is the one we would get with the four reviews. In particular, given the number t of submissions we can accept (as long as they get marks above a minimal acceptance threshold), we want to estimate the earliest point (the minimum number of reviews) so that we can state whether a paper will or will not be in the top t. As an example, if a paper has two strong reject reviews, it is impossible for it to end up in the acceptance range, so we can stop the review process for this paper after two reviews. Similarly, if the paper has three rejects, we can confidently skip the fourth review. Stopping reviews for guaranteed acceptance is more complex as it depends also on the marks of other papers (being above a threshold is not enough as it is a competitive process) but essentially it always amounts to verify if there is a possible combination of marks for the missing reviews that can change the ranking to the point that the paper can end up in the reject bin.

In Fig. 7 we show the results of such deterministic approach for a specific case where $|\mathcal{R}| = 5$, for each criteria M_j the reviewer can assign a mark between $\{1, \ldots, 10\}$ with no half-marks and with a acceptance threshold $T = 7.0$. The darker areas in the bottom of the diagram indicate the cases where the fate (rejection) of the contribution

is already finalized and no further review will change it. The shadow areas in the upper part indicate the symmetric cases where the acceptance of the contribution is sure, in this case that is based on a simple threshold.

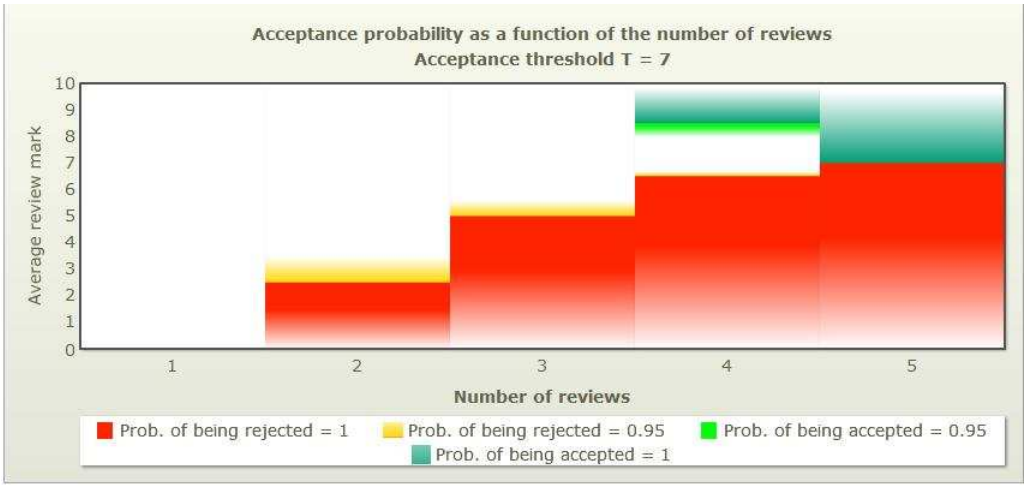

Fig. 7: Deterministic acceptance and rejection of a contribution. The figure shows, given an average mark for r reviews, what is the probability that the contribution will be accepted or rejected according to all the possible marks that could be given in the remaining $|\mathcal{R}| - r$ reviews

In addition to the deterministic analysis done above, which is conservative, we could perform a statistical analysis relying on the fact that reviewers' marks should exhibit some correlation (if three reviewers give a weak reject mark, it is unlikely that the fourth will give a strong accept to bring the paper above the threshold). In general, after each phase, we can estimate the probability of each paper ending up in the accept or reject bin, and to do so we can also leverage our previous disagreement measures to help estimate the confidence associated to the estimate. This is also indicated in Fig. 7 only for explanatory purposes with the shadowed areas on top or below the deterministic rejection/acceptance areas.

Notice that implementing the above process requires either a multi-phase review, or requires giving reviewers a priority on what they should review so to increase the chances that the reviews they would have to do later may not be needed because the fate of the contributions has already been determined. The result of the analysis (the extension of the area that denotes when we can stop reviewing a paper) is informally described in the figure but is not formally described here. The formal analysis is part of our current research work.

Reducing the number of phases The goal here is to reduce the writing time by reducing phases. It applies to processes where different phases require different submissions

(a la FET-Open). The metric we leverage is the *Distance between rankings obtained after different review phases*. This metric is computed when a review process has multiple phases and contributions have to be ranked at the end of each phase according to the marks they got. The level of correlation can help to assess if, in a given review process, multiple phases with modified submissions are really needed. In order to measure this we use again the Kendall τ distance. In fact, the problem can be mapped to that of evaluating quality a posteriori, where we take the (supposedly more accurate) ranking after the second phase as a way to measure the quality of the results in the first phase. As in the quality case, we can only measure this for the contributions accepted after the first phase and that therefore went through the next phase.

The analysis of the distance between rankings may have a significant impact on reducing the effort. If the distance is high, then the two rankings are not correlated and one can question whether the first phase is significant at all. In this case, we would be rejecting many proposals that would get an excellent rank in the second phase. If the distance is low, then one phase is enough and we can spare the extra effort in writing and reviewing. Again, the formulation of a concrete suggestion for which values of Kendall should suggest to combine the phases is part of our current research.

Reducing the review criteria Another interesting dimension is the reduction of the criteria used for the review. The point here is not so much for papers - reviewers do not spend much time if they need to rate a paper with one, two or five criteria - but for project proposals. Specifically, in some cases proposals are evaluated along a set of criteria (e.g., *impact on society*) and authors have to write pages (spend effort) to demonstrate that their proposal satisfies the criteria. If we realize that certain criteria and associated marks are on average irrelevant for the final decision or can be predicted by looking at other marks, then we *may* decide to spare authors the related writing effort. To this end, we measure the *Correlation between criteria*, as correlation indicates the strength and direction of a linear relationship between two random variables. To assess the correlation between criteria we use the Pearson's correlation factor:

Definition 16 (Pearson's correlation factor). *Let x and y be two random variables, σ_{xy} be the covariance, σ_x and σ_y be the standard deviations, the Pearson's correlation factor is defined as follow:*

$$\rho_{\mathrm{xy}} = \frac{\sigma_{\mathrm{xy}}}{\sigma_{\mathrm{x}}\sigma_{\mathrm{y}}} \tag{9}$$

with $-1 \leq \rho_{xy} \geq 1$

Note that if:

- $\rho_{xy} > 0$, x and y are directly correlated, or positively correlated;
- $\rho_{xy} = 0$, x and y are independent;
- $\rho_{xy} < 0$, x and y are inversely correlated, or negatively correlated.

Therefore, if the Pearson's correlation factor between two criteria x and y is $\rho = 0$, this means that the two criteria are independent, so it is useful to have both of them in the

review process, while if $\rho = 1$ then the two criteria are correlated, so it is useless to differentiate and it is better to group the criteria[3].

As a variation, we can compute the kendall distance between the ranking of a specific criterion and the final ranking, again to study which criterion is more significant for the final result. Again, the actionable part of this metric (the indication of which values should suggest combining marks) is under investigation.

Effort-invariant choices An additional line of investigation is around effort-invariant choices, that is, varying review process parameters to improve quality while keeping the effort constant.

Review time versus number of reviews. The first issue in this category that we analyze is trading time spent on a review in exchange of an increase in the number of reviews. As we have seen before, the review effort is given by the time t_r spent reviewing the contribution times the total number of reviews assigned N_R:

$t_r = \frac{E_R}{N_R}$

The trade off is between ranking a paper with few high quality (low uncertainty, as per the meaning defined above) reviews or many lower quality (more uncertain) reviews. Assuming that we can actually estimate, via surveys, the accuracy of the review based on time (on average, or for each reviewer), then we can identify which is the best trade-off.

Committee size versus number of contributions per reviewer. A final effort-invariant trade-off we consider is the size of the expert group versus the number of contributions each reviewer gets to review. We assume that effort E requested to each reviewer, average time per review t_r, and average preparation effort t_w are kept constant, and also constant and decided a priori is the number $N_{RP}$of reviews we wish to have for each paper. We also assume that the number of submissions $|\mathcal{C}|$ is known or can be estimated.

The number of papers or contributions per reviewer N_{PR} is equal to $\frac{N_{RP} \times |\mathcal{C}|}{|\mathcal{R}|}$. Figure 8 shows the plotting of this function, where N_{PR} varies from a minimum value of 1 (each reviewer gets one paper, which means that the size of the expert group is equal to the total number of reviews) to a maximum value of $|\mathcal{C}|$. The question we try to address is what is the optimal point to be selected on the curve in the figure. In general, the point cannot be freely chosen because chairs typically decide the effort they can ask from each reviewer. This effort is defined by N_{PR} x t_r and depending on how demanding chairs are, the corresponding line in the chart can be above or below the maximum value for N_{PR}. Depending on the effort, therefore, N_{PR} can vary between 1 and the smallest between $|\mathcal{C}|$ and $\frac{E}{t_r}$, rounded to the lower integer. We refer to this value as M_{PR} (maximum number of papers per reviewer).

Given that all values of N_{PR} between 1 and M_{PR} are acceptable, the point is how to fix this value - and consequently how to fix the size of the review committee. Intuitively, we believe that giving many papers per reviewer is preferable as the reviewer can form an opinion on the quality distribution of the papers and can mark them accordingly. If

[3] Incidentally, as a particular case, this metric could also be useful to assess quality if we find a high correlation between the reviewer confidence and the overall quality mark, e.g., we may find that reviewers with a low level of confidence always give low rate or high rate.

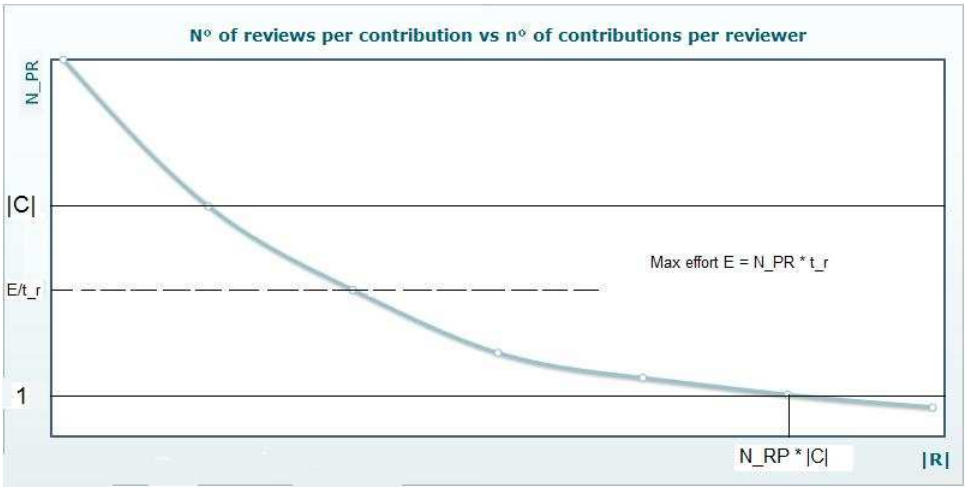

Fig. 8: Given a constant effort, a constant number of reviews and time per review, this is a possible model of the number of reviews per contribution with respect to the number of contributions assigned to each reviewer

a reviewer only sees one paper it may be hard to state whether this is good enough, and impossible to estimate if this is in the top k. The problem is particularly significant with junior reviewers who do not have an idea of the quality level to expect.

More quantitatively, there are two observations we make in this respect: the first is that high values of N_{PR} are good because they allow us to detect biases. We have more samples to make the case and determine if a reviewer has an accepting or rejecting behavior and possibly compensate. The second is that this raises the need for additional metrics to explore if there is a correlation between quality and N_{PR}. Specifically, what we propose is to compute the spread, agreement, robustness, and quality metrics for the various values of N_{PR} in our datasets and examine if there is a correlation between N_{PR} and these metrics. We will perform this analysis in the following sections and discuss conclusions and implications from the perspective of this metric.

3 Experiments on conference papers

3.1 Review process and dataset description

Metrics described in section 2 have been applied also to conference papers, adapting them to the different assessment procedure. This section presents the three datasets we use for the analysis, one related to a large conference (more than 500 submissions), another related to a medium-size conference (less than 500 submissions), and the third one related to a small workshop (less then 50 submissions) with predominantly young reviewers. All the datasets available refer to conferences in the ICT area. In the following, for each dataset, we describe both the review process, basic information and statistics on the dataset and the computed metrics.

3.2 Large conference

Parameter	Details	Alias
$\mathcal{C}$	$\mid\mid \mathcal{C} \mid > 500$	"Submissions" or "papers"
$\mathcal{SR}$	$\mid\mid \mathcal{SR} \mid > 40$	"Senior reviewers"
$\mathcal{R}$	$\mid\mid \mathcal{R} \mid > 500$	"Reviewers"
$\mathcal{M}$	$\mathcal{M}$ = {"overall evaluation", "significance", "novelty", "relevance to the conference", "presentation", "technical quality", "reviewer's expertise"}	"Criteria"
T	$T =$ unknown	-
π	Each paper has been assigned to 3 reviewers and 2 senior reviewers	-

Table 1: $p = \{\mathcal{C}, \mathcal{E}, \mathcal{M}, T, \pi\}$

The conference covered topics related to computer science. Following our evaluation process model, we define the characteristics of the process in Table 1. The number of papers per reviewer varies from 1 to 13 and the marks range from 0 (lowest) to 10 (highest), with no possibility of half-marks. The review procedure was the following: two program committee members (one primary and one secondary) and three reviewers have been assigned to each paper. Papers have been subject to blind peer review, so reviewers were not aware of the identities or affiliations of the authors. On average, we have 3 reviews per paper, and an overall set of more than 500 reviewers. A reviewer could also be an author of a submitted paper. The overall number of analyzed reviews is approximately 3000. After the peer review process, 20,7 % of papers have been accepted.

Fig. 9 shows the distribution of mark values, the majority of the values are concentrated between 3 and 8. A lot of marks given by reviewers are very close to the acceptance threshold which ideally might have been between 5 and 6 (the exact value is not specified in the data set). The analysis is done considering only the first criterion (overall evaluation), which is the most important one to look at when deciding the fate (acceptance/rejection) of a paper.

Looking at Fig. 9, the first thing we noticed is that the distribution is basically flat if we consider the values between 4 and 7. The expected value for the marks related to the first criterion is 5.4 and the standard deviation is 2.02.

3.3 Computation of quality-related metrics

Divergence a posteriori Here, we compute the divergence between the ranking of the conference and the ranking a posteriori given by the citation counts. We recall that

Fig. 9: Mark distribution for the *overall evaluation* criterion in the large conference

the conference was held in 2003, so we were able to compute how many citations the papers got in the subsequent years. We computed the divergence value only for the accepted papers, as only for these we have citations; indeed the rejected ones have not been published (at least not in the same version) so they did not get any citation. In Fig. 10 we highlight the divergence value for t=1/3A $NDiv_{\rho_c,\rho_a}(1/3A, A, A) = 0.63$ and t=2/3A $NDiv_{\rho_c,\rho_a}(2/3A, A, A) = 0.32$ [4].

The former value means that by looking at the top 1/3 of the papers in the two rankings, 63% of these papers differ. If - and this is again an assumption that is not proven - we extend the reasoning to all papers instead of considering only the top 1/3, this means that in a process that accepts 1/3 of the submissions, the review process and the random selection process have the same quality in approximating the citation-based ranking. For completeness, we also compute the value of the Kendall τ distance among the two rankings, the normalized value is $NK_{tau} = 0.49$. We recall that a value $NK_{tau} = 0$ means that the papers are in the same order, while $NK_{tau} = 1$ means that no one of the papers is in the same position w.r.t the two rankings. The value $NK_{tau} = 0.49$ confirms again our previous result (based on the divergence metric) that the review process of such a conference has performed poorly w.r.t. the citation-based count, therefore papers that were ranked very good in the conference got less citations of papers that were lower in the ranking, and that happened for almost 50% of papers.

Disagreement between reviewers Following definitions in Section 2.2 we have computed the average disagreements between reviewers and we report the results of the computation in Fig. 11. In order to compare the different conference dataset among

[4] We indicate with A the number of accepted papers, which we recall was 20.7% of total submissions

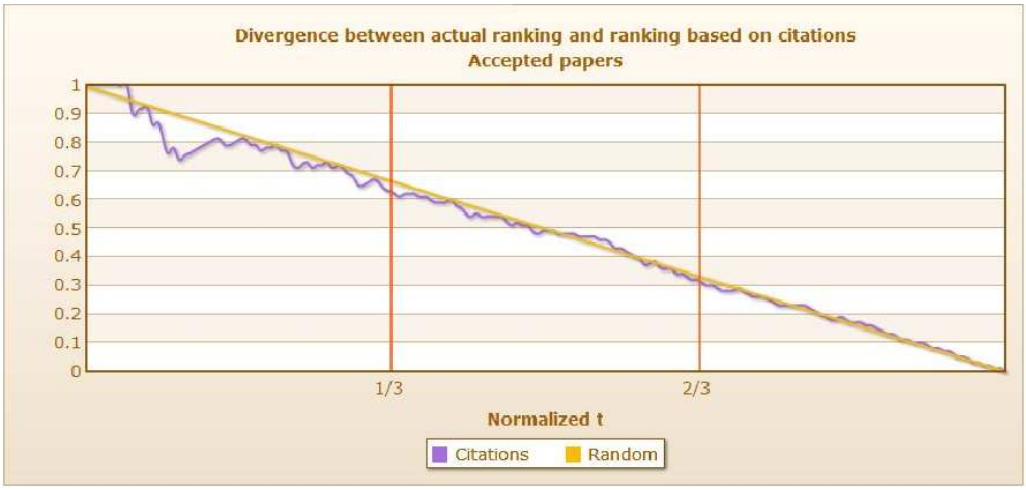

Fig. 10: Normalized divergence calculated for accepted papers

themselves, all values in this section are normalized (i.e. divided by the highest possible value: maximum disagreement, maximum bias, etc..).

Here, the values computed for the disagreement are not surprising: the disagreement computed with the actual data is consistently lower than the random one, and lower than the one computed with the reshuffling of the marks. Such a straight difference could be due to the fact that here the reviewers use consistently the entire scale, so marks assigned to the papers are, at the end, more distributed.

Then we also computed the average disagreement w.r.t. the number of reviews done and we have not found any particular correlation.

Robustness In order to assess the impact of a perturbation on the mark values and find how this can affects the number of accepted/rejected papers we compute the robustness of the process. We applied a perturbation of $\epsilon = 1$, meaning that we randomly selected among three possible variations of the marks: -1/0/1. Note that we apply a stochastic variation to each one of the six marks. Then we compute the divergence between the actual ranking and the "perturbated" one, computed as already explained in Section 2.2. In Fig. 12 are shown results for t=A, that is the number of accepted papers. The divergence value is only $NDiv_{\rho_a,\rho_\epsilon}(A, C, C) = 0.06$ [5]. This is a very low value, as a perturbation in the mark value of $\epsilon = 1$ impacts only the 6% of the papers. Summing up, the analysis of these data suggests that the process is quite robust.

[5] C is the total number of submitted contributions

	Φ_1	Φ_2	Φ_3	Φ_4	Φ_5	Φ_6
Original	0.27	0.28	0.28	0.26 (+-0.01)	0.29 (+-0.01)	0.28
Reshuffled	0.34 (+-0.01)	0.34 (+-0.01)	0.33 (+-0.01)	0.30 (+-0.01)	0.35 (+-0.01)	0.35 (+-0.01)
Random	0.49 (+-0.01)	0.49 (+-0.01)	0.49 (+-0.01)	0.49 (+-0.01)	0.49 (+-0.01)	0.49 (+-0.01)
$\Psi = 0.28$						

(a) Average disagreement Ψ and standard error σ

$\overline{\gamma}$	N° of reviews made	N° of reviewers
0.18 (+-0.02)	x>=10	6
0.16 (+-0.01)	x=9	7
0.17 (+-0.01)	x=8	18
0.20 (+-0.01)	x=7	20
0.18 (+-0.01)	x=6	43
0.18 (+-0.01)	x=5	70
0.19	x=4	173
0.18	x=3	236
0.19	x=2	214
0.19 (+-0.01)	x=1	158

(b) Average disagreement $\overline{\gamma}$ and standard error σ computed separately for reviewers grouped by number of reviews made

Fig. 11: Computed average disagreement

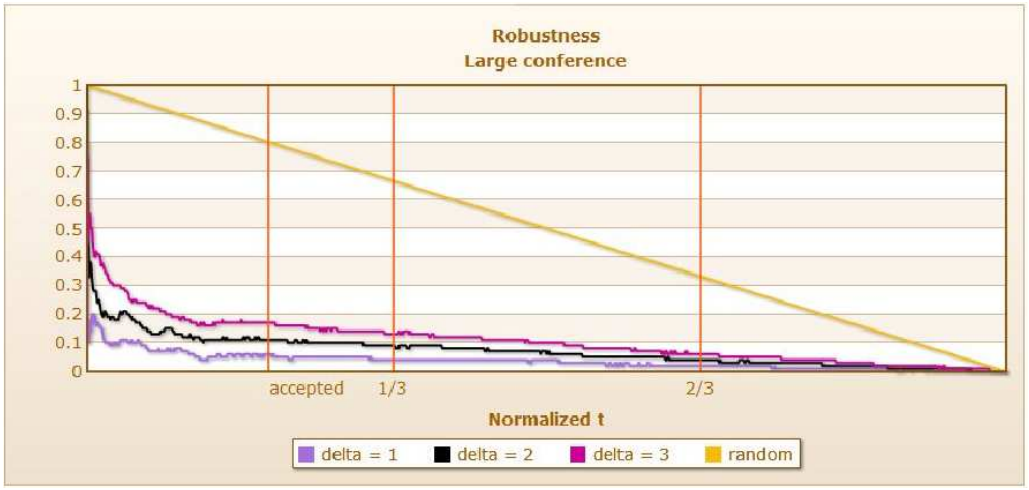

Fig. 12: Robustness for the large-size conference

Fairness-related metrics The same metrics used in the project analysis has been exploited for papers. Here, we only changed the threshold fixed to determine if a reviewer is biased or not. In this case we only consider reviewers who assessed at least 4 papers, with an average bias greater than 1 for the positive case and lower than -1 for the negative case (please remember that in this data set, ± 1 is the minimum marks difference). The obtained results are collected in Fig. 13 and Fig. 14. As previously said, the disagreement and the bias values have been computed using the metrics defined in Section 2.2.

Reviewers with ID=796 and ID=347 (Fig. 13) have given higher marks than the others on more than 10 papers. In particular, their marks have been more than one point higher, on all the criteria, with respect to the average of the marks given by the other reviewers to the same paper. Remarkable is the case of the reviewer with ID=1461 (see Fig. 13), who, despite the low number of reviews he made, has given marks that are almost 3.5 points higher on every criterion than those given by the other reviewers. Quite interesting is also the case of the reviewer with ID=2497 (see Fig. 14), who gave marks that are more than 2 points lower on every criterion than those given by the others.

We perform a simple unbiasing procedure: considering only reviewers having a bias greater than 1, we add or subtract the individual computed bias in order to obtain a new "unbiased" rank. Finally, we compute the divergence between the two rankings, the biased and the "unbiased" one, obtaining $NDiv_{\rho_1,\rho_2}(A, C, C) = 0.09$. The divergence value is computed for the accepted papers, the result means that the 9% of the papers could be affected by the unbiasing procedure. As the 0.09 value is greater then the one computed for the minimal stochastic variation (0.06), the "unbiased" procedure could have a real effect in this case, if the conference chair would decide to use it.

ID reviewer	number of reviews made	avg bias				
796	13	1.10		1189	4	2.53
347	11	1.42		781	4	1.14
1109	8	1.13		2619	4	1.17
1380	8	2.24		2676	4	1.21
1284	8	1.10		2639	4	1.71
2716	7	1.18		2684	4	1.14
890	7	1.02		2831	4	1.13
2664	6	1.56		1038	4	1.31
769	6	1.30		548	4	1.67
778	6	1.23		2810	4	1.67
2079	6	1.07		2834	4	1.84
2843	6	1.43				
1294	6	1.92				
2883	5	1.58				
765	5	1.14				
820	5	1.16				
822	5	1.37				
364	5	1.01				
2838	5	1.12				
2771	5	1.08				
2570	5	1.98				
2615	5	1.34				
1500	4	1.44				
1503	4	1.74				
1387	4	1.20				
2535	4	1.47				
1493	4	1.10				
1432	4	2.35				
1921	4	2.02				
2495	4	1.13				
1870	4	1.15				
2508	4	2.30				
1468	4	1.24				
1461	4	3.44				

Fig. 13: Positively biased reviewers - people who reviewed more than 3 papers

ID reviewer	number of reviews made	avg bias				
2696	8	-1.42		387	4	-2.71
1257	8	-1.47		1104	4	-1.44
2734	7	-1.39		478	4	-1.02
2818	7	-1.40		2908	4	-1.38
379	7	-1.91		637	4	-1.31
2497	7	-2.52		755	4	-2.38
1406	6	-1.83				
1288	6	-1.02				
767	6	-1.03				
976	6	-1.11				
804	5	-1.12				
972	5	-1.18				
1972	5	-1.15				
1912	5	-2.78				
766	5	-1.10				
2709	5	-1.57				
554	5	-1.34				
345	4	-1.47				
1507	4	-1.31				
1528	4	-1.73				
2534	4	-2.29				
369	4	-1.76				
2046	4	-1.58				
1670	4	-1.10				
1717	4	-1.83				
1755	4	-1.64				
1471	4	-1.39				
2545	4	-1.52				
2678	4	-1.28				
2590	4	-2.06				
807	4	-1.07				
2668	4	-2.13				
2762	4	-1.09				
2862	4	-2.10				

Fig. 14: Negatively biased reviewers - people who reviewed more than 3 papers

3.4 Computation of efficiency-related metrics

Correlation between criteria In Fig. 15 we have reported the correlation computed between each pair of criteria. The first criterion ("Overall evaluation") is the overall judgment for a given paper. Through the calculation of Pearson's correlation coefficient it has been found that the overall acceptance of a paper strongly depends on three core features: the *significance* of the paper, its *novelty* and its *technical quality*. The paper relevance to the conference and also the presentation do not have the same importance with respect to the paper acceptance.

	Overall evaluation	Significance	Novelty	Relevance to the conf	Presentation	Technical Quality	Reviewer expertise
Overall evaluation	1	0,84	0,7	0,42	0,53	0,89	-0,04
Significance		1	0,7	0,44	0,47	0,86	-0,04
Novelty			1	0,42	0,39	0,74	-0,04
Relevance to the conf				1	0,37	0,49	0,13
Presentation					1	0,58	0,12
Technical Quality						1	-0,02
Reviewer expertise							1

Fig. 15: Correlation between criteria

Another aspect came out from the analysis of the correlation: the confidence score, in general, is not correlated with the other criteria. This is coherent with the fact that conceptually the quality of the proposal is independent from the expertise of the reviewer in the research field.

3.5 Medium-size conference

The second conference dataset we analyze refers to a medium-size conference held in 2008, with more than 200 submissions, more than 100 reviewers, no senior reviewers, with only one criterion to decide the overall evaluation of the paper. The mark scale ranges from 1 to 7, with no half marks. No threshold was defined, each paper was assigned to 3 or 4 reviewers and the peer review process was blind. The number of reviews analyzed is more than 800. The acceptance rate of the conference was 18%. The peer review process is summarized in Table 2.

Fig. 16 shows the mark distribution. It is interesting to note that the most frequent marks is 2, and that there is a low probability to get a mark in the middle of the scale (4 in this case).

Parameter	Details	Alias
$\mathcal{C}$	$\mid \mathcal{C} \mid > 200$	"Papers"
$\mathcal{SR}$	$\mid \mathcal{SR} \mid = 0$	"Senior reviewers"
$\mathcal{R}$	$\mid \mathcal{R} \mid > 100$	"Reviewers"
$\mathcal{M}$	$\mathcal{M}$ ={"overall evaluation"}	"Criteria"
T	T = undefined	-
π	Each paper has been assigned to 3 or 4 reviewers	-

Table 2: $p = \{\mathcal{C}, \mathcal{E}, \mathcal{M}, T, \pi\}$

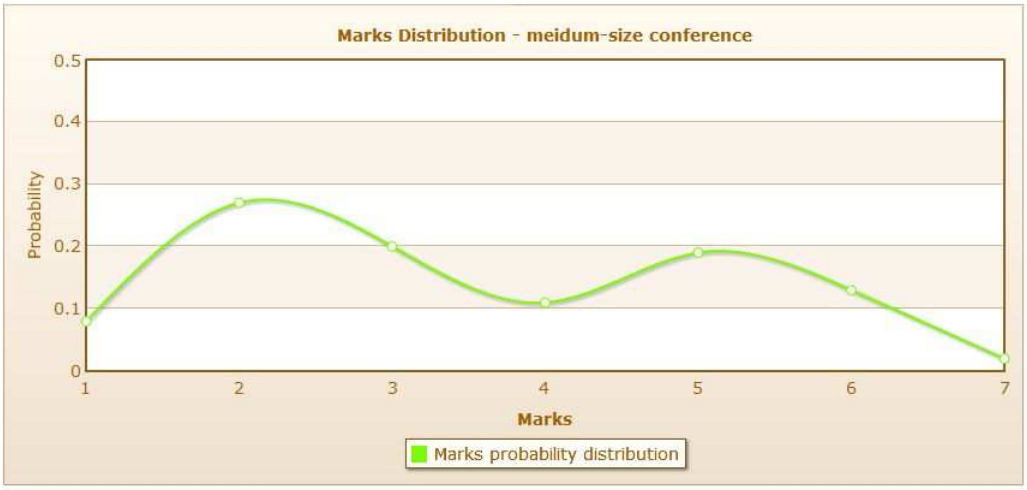

Fig. 16: Marks distribution

We computed the **disagreement** between reviewers, the computed values are depicted in Fig. 17 we recall that the disagreement values are normalized,i.e. zero represents a perfect agreement among reviewers, while 1 represents the maximum disagreement. The disagreement here is slightly higher than the one computed for the large-size conference (0.32 in the medium-size one w.r.t. 0.28 in the large-size one). While the disagreement computed reshuffling the marks here is higher than the original one (0.40 in the medium-size one w.r.t. 0.32 in the large-size one) as one would expected.

	Φ_1
Original	0.32 (+-0.05)
Reshuffled	0.40 (+-0.01)
Random	0.50 (+-0.01)
Ψ = 0.32 (+-0.05)	

Fig. 17: Disagreement

Then, we analyzed the **robustness** of the process in order to assess the effect of a perturbation in the mark values and how such a perturbation is reflected in the number of accepted/rejected papers. We applied a perturbation of $\epsilon = 1$, that is we randomly selected among three different variations in the mark values: -1/0/1. Then, we computed the divergence between the two rankings, the actual one and the one obtained with the "perturbated" marks, as already explained in Section 2.2. The results are depicted in Fig. 18. The divergence value for t=A (number of accepted papers), and C total number of contributions, is $NDiv_{\rho_a,\rho_\epsilon}(A, C, C) = 0.17$, this means that a perturbation of $\epsilon = 1$ impacts the 17% of papers. We notice that the divergence value computed here cannot be compared with the one computed for the large-size conference, the same is for robustness. Indeed, in both cases we applied a perturbation $\epsilon = 1$, but while in the large-size conference the scale ranges from 0 to 10, here the scale ranges from 1 to 7, therefore a variation $\epsilon = 1$ has a higher impact in this case than in the former one, and this explains the higher divergence value here w.r.t. the large-size conference.

We computed the **bias** values for all reviewers. The results are depicted, respectively, in Fig. 19 (positive bias) and Fig. 20 (negative bias). We show only reviewers with a bias value greater than $|1|$, coupled with the number of reviews made. Also in this case we have performed the unbiasing procedure, adding or subtracting the computed bias from the original marks and computing the new rank with the "unbiased" marks. We then computed the divergence between the two rankings, the biased and the "unbiased" one: $NDiv_{\rho_1,\rho_2}(A, C, C) = 0.11$, for t=A (accepted papers) the divergence is 0.11, this mean that the unbiasing procedure could impact the fate of 11% of the accepted papers.

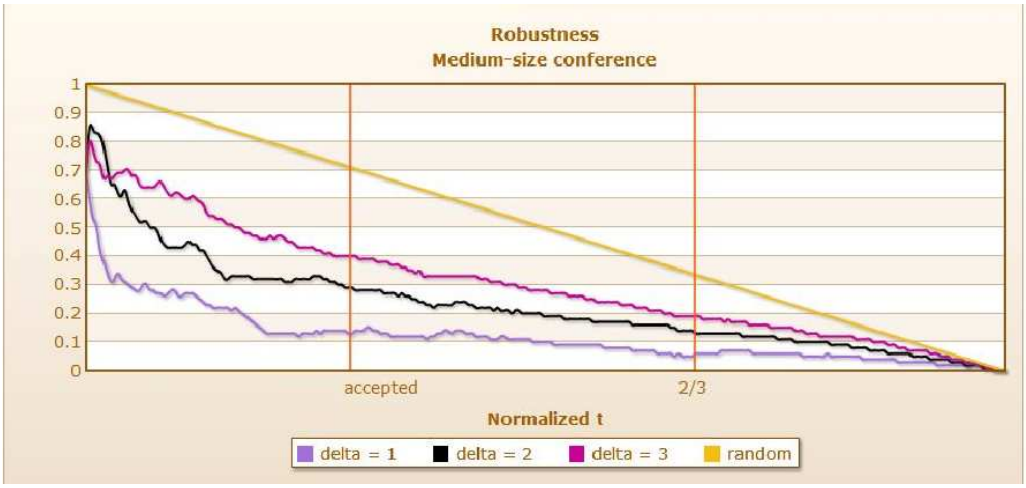

Fig. 18: Robustness

Reviewer ID	N° of reviews	Avg Bias
29	10	1.52
75	10	1.52
41	10	1.28
82	5	1.23
5	10	1.17
2	10	1.12
34	10	1.07
80	9	1.06
98	9	1

Fig. 19: Positive bias

Reviewer ID	N° of reviews	Avg Bias
103	3	-2.06
12	10	-1.78
44	8	-1.42
20	10	-1.22
28	10	-1.2
56	5	-1.17
30	10	-1.03
91	10	-1.02

Fig. 20: Negative bias

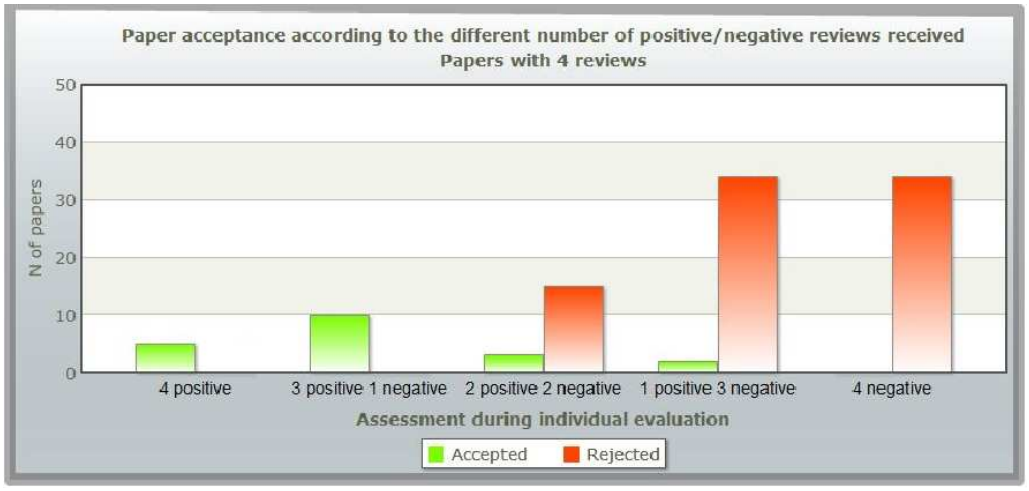

Fig. 21: Acceptance of the paper according to the number of positive/negative reviews received - papers with 4 reviews

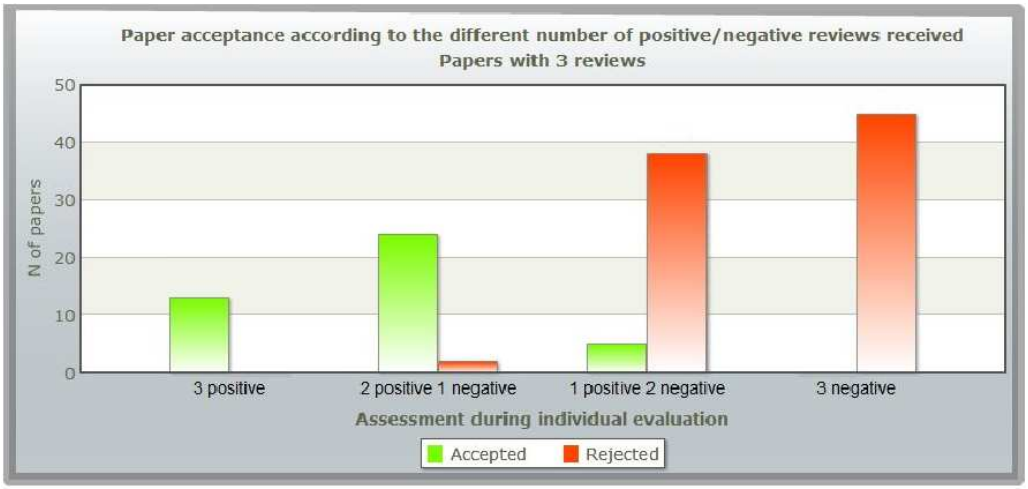

Fig. 22: Acceptance of the paper according to the number of positive/negative reviews received - papers with 3 reviews

3.6 Small, informal conference

The data set refers to a small conference held in 2009, with only a small number of submissions (45 papers), where contributions where reviewed (and also submitted) by 45 young reviewers (essentially Ph.D students). The marks range between 1 and 5 with no half marks allowed. The process was a blind process, reviewers were aware of the identities and the affiliations of the authors. A threshold was not defined, but at the end 28 papers on 45 have been accepted for presentation (62% acceptance rate). Each paper has been assigned to 3 reviewers and no senior reviewers were involved in the process. The characteristics of the peer review process are depict in Table 3.

Parameter	**Details**	**Alias**
$\mathcal{C}$	$\mid \mathcal{C} \mid = 45$	"Papers"
$\mathcal{SR}$	$\mid \mathcal{SR} \mid = 0$	"Senior reviewers"
$\mathcal{R}$	$\mid \mathcal{R} \mid = 45$	"Reviewers"
$\mathcal{M}$	$\mathcal{M}$ ={"overall evaluation"'}	"Criteria"
T	T = undefined	-
π	Each paper was assigned to 3 reviewers and 0 senior reviewers	-

Table 3: $p = \{\mathcal{C}, \mathcal{E}, \mathcal{M}, T, \pi\}$

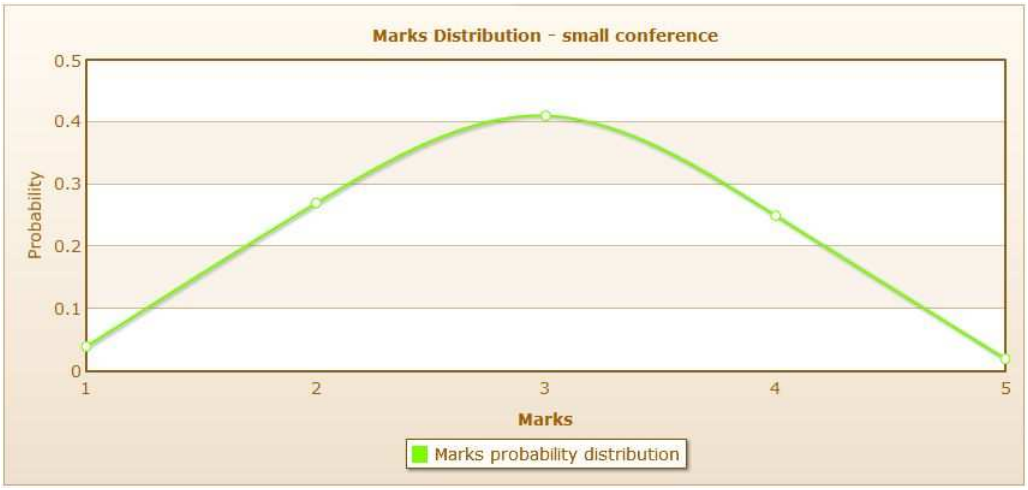

Fig. 23: Marks distribution

Fig. 23 shows the marks distribution, the most frequent mark is 3, right in the middle of the scale, while there is a very low probability to give very extremely marks. In order to assess the **robustness** of the process, following the definition and procedure of the ϵ-divergence metric defined in Section 2.2, and taking into account the fact that our marks are discrete values, we generated 10 stochastic rankings for $\epsilon = 1$, namely we have three possible variations of our mark: $-1/0/1$. We compute the divergence value $NDiv_{\rho_a,\rho_\epsilon}(28, 45, 45) = 0.12$, as shown in Fig. 24 for $t = 28$ (number of accepted papers) the divergence is equal to 0.12. This means that on average a stochastic modification of $\epsilon = 1$ could affect the 12% of papers, that is the final decision on ca. 3 papers could change.

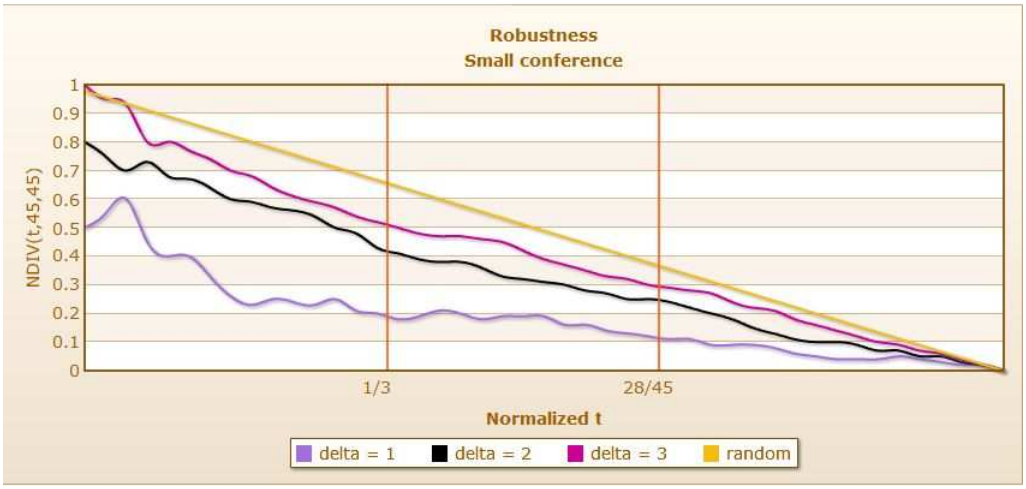

Fig. 24: Robustness for the small-size conference

For what concerns the **disagreement**, we can observe that this is constanly lower than the random one, and lower than the reshuffled one. Comparing these results with the one for large-size conference, we notice that this is slightly lower: 0.26 w.r.t. 0.28 (agreement over all criteria), but essentially,there is not a big difference between the two conferences. While if we compare the disagreement of the small-size conference with the one of the medium-size, we notice that this is considerably lower (0.26 w.r.t.0.32).

The results of the **bias** analysis are shown in Fig. 26, where we can notice that there are, respectively, three very positive biased reviewers (ID 70, 22, 28) and three very negative biased reviewers (ID 25, 20, 42), who constantly give marks which are one point higher (or lower) than the others. In Fig. 26 we highlighted all the reviewers having a bias greater than 0.5, as, we recall, the scale in this case was smaller than in the other data sets (1 to 5, no half marks), so even a variation of 0.5 could be important. We then perform the **"unbiasing" procedure** subtracting or adding the computed bias from the original marks. Once computed the new ranking obtained with the new "unbiased"

	Φ_1	N° of contributions
Original	0.26 (+- 0.03)	45
Reshuffled	0.32 (+-0.04)	45
Random	0.52 (+-0.04)	45
Ψ = 0.26 (+- 0.03)		

Fig. 25: Average disagreement

marks, we compute the divergence between the original ranking and the "unbiased" one as $NDiv_{\rho_1,\rho_2}(28, 45, 45)$ for $t = 28$ is equal to 0.14.

As before, 28 is the number of accepted papers, a divergence of 0.14 means that the "unbiasing" procedure could have been affected ca. 4 papers (14% of 28) that could have been accepted instead of rejected (and vice versa).

ID reviewer	Avg bias	N° of reviews
70	1.17	3
22	1.00	3
28	1.00	3
21	0.83	3
47	0.83	3
12	0.67	3
64	0.67	3
36	0.50	3
18	0.50	3
13	0.50	3
41	0.50	3
46	0.50	3
17	0.33	3
33	0.33	3
37	0.33	3
49	0.33	3
29	0.33	3
54	0.33	3
44	0.17	3
50	0.17	3
35	0.00	3
38	0.00	3
43	0.00	3
45	0.00	3
32	0.00	3
31	0.00	3
24	-0.17	3
56	-0.33	3
34	-0.33	3
16	-0.33	3
69	-0.33	3
26	-0.33	3
27	-0.33	3
15	-0.50	3
40	-0.50	3
48	-0.50	3
23	-0.50	3
19	-0.67	3
14	-0.67	3
11	-0.67	3
53	-0.67	3
51	-0.83	3
25	-1.00	3
20	-1.17	3
42	-1.17	3

Fig. 26: Biased reviewers

3.7 Findings and lessons learned

We now provide some considerations and derive some lessons comparing the results from the three conferences themselves. We remark that for the conference dataset we did not present any results about the optimal number of reviewers per papers, as this is part of our current work. From the present analysis we derive some interesting hints, that will be investigated further on the additional datasets we are obtaining:

- Comparing the actual ranking of the first conference with the ranking based on the citation-count, we found a very low correlation between the two. Indeed, the divergence values obtained is not that far from what a random selection process would give. This means that the review process has apparently performed very poorly. We did not do the same analysis for the other two datasets as the medium-size conference was held in 2008 and the small one in 2009, so we could not really assess a citation-based ranking as the two conferences were too recent. The divergence analysis we do is essentially for the accepted papers, so the interesting hints for our work is now to identify a way to perform the divergence analysis also for papers that were rejected and in particular for papers ranking very low. This is hard to do as we do not have citations for those papers, since they were not published. So we have to identify another way to do this. The goal of the analysis is to see if an assumption that seems reasonable - that is, the assumption that papers that rank very low in the process and are obviously junk never get high in the citation count, which is the current measure of impact of a paper- actually holds, or whether instead from this perspective the review process is poor also in filtering out the very bad papers.
- Comparing the value of the disagreement in the three conferences, we found that the disagreement value was never high for any of the three conferences, and, above all, was always consistently lower than the random one and lower than the reshuffled one. We also notice that we did not find a clear difference between the conference with young reviewers and the other two conferences. This is again contrary to our initial assumption that we would find more disagreement among young reviewers. We need more datasets however to confirm or disprove the assumption. One more interesting analysis to be done is to identify if the reviews for conferences are in fact done by the PC members or by the students and to see if disagreement or quality of the rankings change in the two cases or it is the same.
- We did not find any correlation between the disagreement among reviewers and the number of reviews made. We expected that the disagreement could decrease with the increasing of the number of reviews made, but the dataset did not confirm our hypothesis. Part of our current work - to convince us that the assumption we made was indeed wrong, is the computation of this on more datasets but also the analysis of the stochastic distribution of marks of reviewers with only one paper to review vs that of reviewers with many papers to review, to see if they differ.
- The values for the divergence among the actual ranking and the unbiased one show that an unbiasing procedure would have a significant effect on the final result. This is an important message for PC chairs who may not be necessarily aware of the issue. Our future work - but again this require significant research - includes identifying biases related to topics and other aspects rather than limiting the analysis to accepting or rejecting biases.

4 Conclusions and future directions

This document has provided a brief overview to existing review processes, the modeling of peer review, an analysis of some of these process, and some critical analysis of current community review processes. It seems that there are no clear rules for a successful review process. For example, in our analysis we have found that there is a significant degree of randomness in the review process, more marked than we initially expected; there is very little correlation between the rankings of the review process and the impact of the papers as measured by citations. On the other hand, open peer review carries the risk of potential bias, increasing conflict between author and reviewers, and a decrease in reviewer's willingness to be properly critical. Moreover, the value of peer review lies also in specific comments and advice rather than in general or abstract measures of 'quality' on which there is usually little agreement.

The question is, do current practices in peer review process need to be improved ? And if the answer is positive, then the related question is: how can we improve them ?

The work reported in this paper aimed to answer - at least preliminarily and partly - these questions. We think that there is a common consensus on the fact that traditional peer review processes need to be improved. This is even more evident is we take into account all the new ICT technologies and related tools and media, that at present are not fully used in current practices.

By applying our analysis and framework to available review data sets, we have been able to reveal a number of interesting features about peer review processes.

What we want to address now are the future plans and direction of the present line of research and its implication and relationships to the next phases of the LiquidPub project.

Our future plans include:

1. Extend our framework analysis to community review processes and combine it with the metrics and tools developed specifically for these data sets. Furthermore, knowing more details about the semantics associated to the available peer review data could allow the introduction of more sophisticated metrics, borrowing algorithms and techniques used in the social networks analysis and in the collaborative filtering processes.
2. Apply the proposed metrics and theoretical framework to a larger number of data sets. In order to overcome the relevant privacy issues that we have encountered in the first year, we are working towards agreements with interested stakeholders to develop a specific plug-in to their information systems in order to collect anonymously the needed aggregated data directly from their database.
3. Explore the design and development of an ICT support tool for peer reviews, in terms of supporting the selection of reviewers, the assignment of contributions, the analysis of review results, and the efficiency of the overall process as well as rendering the whole review process open and transparent.

In addition to the above contributions to the LiquidPub platform, we present below some further applications describing how the analysis framework and metrics proposed in this document can aid users, for instance, in a future liquid conference (or journal, workshop) publication process.

Selecting PC committee members
Computing the reputation of researchers in reviewing (both published and unpublished) paper may assist in deciding who to invite as PC members. This, we believe, does not only provide a relatively strong incentive for researchers in the community to review (or rate) other papers, but also to review them "properly" (in other words, to take the review process more "seriously"). This is because the closer the researcher's review result is to the group's result, then the higher his reputation as a reviewer is; hence, the higher his probability for playing the role of a PC member is.

Aggregating reviewers' results
When the review results need to be aggregated, a reliability measure may be attached to each review result. How reliable is the result provided by a given reviewer depends, amongst other things, on the possible bias of the result, whether the reviewer is in a competitive or a collaborative relationship with the author(s) of the paper being reviewed, whether there exists strong dependencies amongst reviews, and so on.

Deciding who & how many reviewers should review a given paper
The reputation of a single reviewer may assist in deciding whether a proposed reviewer is reliable to review a given paper. This reliability measure depends on measures that have already been discussed above, such as possible bias, possible cooperative/competitive relations with the authors of the paper to be reviewed, the reviewer's confidence, etc.

To conclude, we believe that the proposed analysis framework (and its future extensions) can support the development of new ad more efficient review processes that will have the additional properties of being more open (i.e. community driven vs. clique behavior), more transparent (in terms of shared, accessible and quantitative monitoring frameworks) and more efficient in terms of minimizing the time spent by *both* authors and reviewers.

References

1. KENDALL M.G. A new measure of rank correlation. *Biometrika 30*, 1-2 (1938), 81–93.
2. KRAPIVIN M., MARCHESE M., CASATI F. Exploring and understanding citation-based scientific metrics, 2008.