UNIVERSITY
OF TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE

38123 Povo – Trento (Italy), Via Sommarive 14
http://www.disi.unitn.it

REVIEWING PEER REVIEW:
A QUANTITATIVE ANALYSIS OF PEER REVIEW

Fabio Casati, Maurizio Marchese, Katsiaryna Mirylenka and Azzurra Ragone

February 2010

Technical Report # DISI-10-014

Reviewing peer review: a quantitative analysis of peer review

Fabio Casati, Maurizio Marchese, Katsiaryna Mirylenka and Azzurra Ragone
Dipartimento di Ingegneria e Scienza dell'Informazione
University of Trento, Via Sommarive, 14 38100 Trento, Italy
{casati,marchese,kmirylenka,ragone}@disi.unitn.it

ABSTRACT

In this paper we focus on the analysis of peer reviews and reviewers behavior in a number of different review processes. More specifically, 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 from conference evaluation processes and we discuss the results implications and their eventual use toward improving the analyzed peer review processes. A number of unexpected results were found, in particular: (1) the low correlation between peer review outcome and impact in time of the accepted contributions and (2) the presence of an high level of randomness in the analyzed peer review processes.

1. INTRODUCTION

Peer review is simultaneously one of the most established and most controversial aspects of research assessment. Virtually every active researcher has - at one time or another - gained great benefit from a reviewer who helped to correct unnoticed (and sometimes serious) errors, suggested ways to clarify or improve results and their description, or brought to the attention other related work that was of great interest. At the same time most researchers have also experienced some of their papers or research proposals being blocked by reviews that seemed very superficial at the best and in some cases even mendacious. As scientists, we live in fear of the second kind of review and hoping to get the first. Yet the process through which this happens is little known and investigated with a scientific approach.

Since the role of the peer review process is fundamental for science (to select quality contributions), for innovation (to select relevant research proposals and patents) and for people (to provide proper credits assignment as well as career advancement), we think it is crucial to monitor it more closely and quantitatively. This in order to provide to all involved stockholders (program chairs, reviewers, authors, funding institutions as well as the final users - scientists) all relevant information in the most transparent way.

As a motivating example let's take a look to the typical review process that this paper is going through. The conference is a very relevant, authoritative and competitive conference in a specific and interdisciplinary domain. But little is known publicly in regard to:

- the specific peer review process model (i.e. whether it is a single phase or a multiple phase process with discussion or rebuttal; whether the number of reviews per paper is decided a priori or it is adjusted during the process).
- monitoring of eventual reviewer's biases (i.e. rating, affiliation, topic, country, gender, clique biases..); and if this is done, what are the applied counter-measures and whether or not they are transparently shared.
- the statistical accuracy of the current evaluation for a specific contribution from its ideal evaluation (to be defined by the the program chairs, but for instance where all the experts in the program committee evaluate the contribution)
- known a-priori criteria that the program chairs have decided to use as quality indicators to monitor the impact in the scientific community of the conference contributions in time (i.e. citations of accepted conference papers in time, number of extended version of the papers published in Journals, etc.).

We think that all these information - as well as others - should be considered in every scientific evaluation process in order to make it more effective and transparent.

The focus of our work is to analyze and understand the characteristics of current review processes in academia. In parallel, we aim to highlight strengths and weaknesses of the processes and to study quantitatively aspects potentially related to three core features that should, in principle, characterize every evaluation system: *quality, fairness and efficiency*. Once these aspects have been sufficiently investigated and understood, there might be the possibility to propose new models for the evaluation, that could provide further metrics and algorithms for supporting the set up, the management and the improvement of the process. Specially, 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, efficient papers distribution among reviewers and the statistical accuracy of the review results). In the past these quantitative analysis were somehow hinder by the manual and paper-based nature of the evaluation procedures. Nowadays, with the use of computer supported environment (e.g. Conference Management Systems) quantitative analysis are feasible and - in our view - should be included in every review process.

The main contributions of the paper are: (1) the definition of an initial number of metrics capable to analyze quantitatively the three core dimensions of the process: quality, fairness and efficiency; (2) once the model and the analysis tools have been defined, we have applied them to a number of data sets from conference papers review processes. Moreover a number of interesting (and some unexpected for us) aspects of the review system have been found. In brief for the specific analyzed data sets:

- there is a significant degree of randomness in the analyzed review processes, more marked than we initially expected; the disagreement among reviewers is high and there is very little correlation between the rankings of the review process and the impact of the papers as measured by citations.

- it has been 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 have shown 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

- we have also focused on efficiency-related metrics and we have shown that it is possible to device statistical approaches to tune review process parameters to improve quality while keeping the overall effort under control.

The results obtained from the application of the proposed analysis framework and related metrics are capable to characterize the different review processes. Through the computed results we are able make useful comparisons between the different processes and to drawn some general remarks and lessons learned.

The paper is structured as follows. In Section 2 we provide a brief description of the related work. Section 3 presents the dimension of our analysis, while the subsequent sections details the proposed metrics, analysis and related results and lessons learned. Conclusion and discussion of future work closes the paper.

2. RELATED WORK

Peer review has been widely studied in the last years, therefore this section, due to lack of space, is far from being complete.

Even if the origins of peer review date back to the Greek time, the first journal that introduced the peer review process as we know it today has been the *Medical Essays and Observations*, first published in 1731 [13].

Peer review is one of the most debatable topic, every scientist has an opinion on it, and sometime opinions are not very positive. Indeed, just to cite a few, peer review has been defined as a crude and understudied, but indispensable process[6] or as a process "whose effectiveness is a matter of faith rather than evidence" [12] These arguments are mostly based on the assumption that reviewers sometime are not completely objective, but are, instead, biased e.g. on gender, affiliation, country, status, or, worst, they have malicious intents against rival scientists they do not like [10].

Peer review has been analyzed and studied by several researchers, however, we notice that such analysis are not straight comparable, as they refer to review processes coming from different disciplines and different journals. Indeed, sometime even analysis done in the same field can lead to contradictory results [5].

Several issues related to peer review have been investigated by scientists: (i) if peer review is really able to improve the quality of a paper and corrects consistent errors, (ii) if the bias introduced by reviewers could have a significant impact in the review process, and (iii) if having open or double-blind review process could lead to better (or worst) results. For what concern the first item a study was conducted by Goodman et al. [4] who tried to measure the quality of the papers submitted to the Annals of Internal Medicine between March 1992 and March 1993 before and after the peer review process. They did not find any substantial difference in the manuscripts before and after publication. Indeed, they state that peer review was able to detect only small flaws in the papers, such as figures, statistics and description of the results. An interesting study was carried out by Godlee et al. [3]: they introduced deliberate errors in papers already accepted by the British Medical Journal(BMJ). Godlee et al. report that the mean number of major errors detected was 2 out of a total of 8, while there were 16% of reviewers that did not find any mistake, and 33% of reviewers went for acceptance despite the introduced mistakes.

For what concern studies on bias, there are works that have found *affiliation* bias (meaning that researchers from prominent institutions are favored in peer review) [1], bias in favor of US-based researchers [11], or *gender* bias against female researchers [15].

Finally, several studies have been done on *open* or *double- blind* review process. For the former, there is the problem that less researchers agree to review a paper if the review is not blind, but, then, they spend more time in doing the review and are less harsh and more courteous [14]. While, for double-blind review the main problem is that it is really difficult to enforce this policy, as authors always introduce (deliberately or by mistake) elements that help reviewers to identify them [7].

In our work, as described in the subsequent section, we have defined several metrics to study and understand the peer review process. Furthermore, as we did these analysis on a set of conferences from the computer science field, we do not claim that these results are general, but we think that they can help to understand better this process and maybe can give useful suggestions for future improvements.

3. PEER REVIEW METRICS DIMENSIONS

We define a number of metrics for evaluating quantitatively peer review processes. The purpose of such metrics is to help us understand and improve peer review process along three dimensions: *quality*, *fairness* and *efficiency*. Briefly, *quality* is related to the final result: a review process ensures quality if the best contributions are chosen. *Fairness*, in our approach, is related to the monitoring of the contributions assignment process to the reviewers: a process is unfair if the acceptance of the contribution depends on the particular set of reviewers that review it. *Efficiency* is related to the time spent in preparing and assessing the contributions and to the the statistically accurate review results: a process is efficient if the best proposals are accurately chosen with minimal time spent both by authors in preparing the contribution and by reviewers in performing the reviews. 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 only at quantitative review data (e.g. marks) and not natural language comments.

In the following, we present results mainly from just one review process, due to lack of space. However we have done analysis for several anonymous review process data (at the moment six), which differ in size (number of reviewers and papers), in the criteria used and/or in the review process (one-phase or two-phases with discussion among review). We present results from more then one process when there are significance discrepancies in the results obtained for different processes.

4. PROCESS QUALITY 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". In the next sub sections we explore

the rationale behind a number of proposed process quality-related metrics.

4.1 Mark distribution

A very simple analysis is to look at the distributions of the marks (following the experimental scientist motto: "always look at your data"). Analyzing the distribution of marks in review processes with different mark scales, we notice that the way reviewers give marks can be influenced by the scale itself. In Figure 1 we plot distribution of marks from processes where (1) marks range from one to ten (no half-marks allowed); (2) from one to seven (no half marks); (3) from zero to five with half-marks.

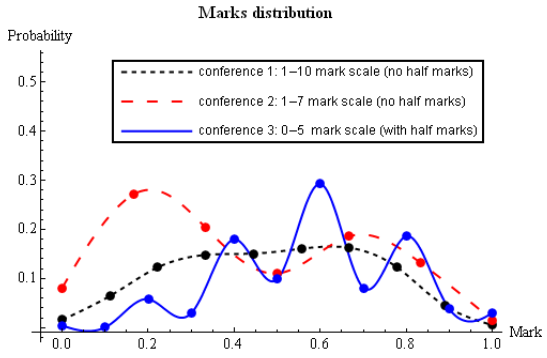

Figure 1: Examples of mark distributions in conferences with three different scales

In case (3) we notice that reviewers tend not to give half marks, indeed the curve has many oscillations; while if we consider almost the same scale - that is the doubled scale with integer marks instead of half marks, as in (1) - the mark distribution appears concentrated around the middle of the ratings scale [1]. Moreover, in case (2) reviewers tend not to give the central mark (4 in this case), but to give lower or higher marks (here the most frequent marks is 2). It seems that such a scale "forces" the reviewer to take a decision, avoiding the central marks which corresponds to a *neutral* mark.

4.2 Divergence

In this metric we do assume that, somehow, we have the "correct" or "ideal" ranking for the contributions. We assume 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. In the literature, the typical metric for measuring a difference between two rankings is the Kendall τ distance [8]. This metric, however, computes the difference in the exact position of the elements between two ranks, while in the review process the main issue is not to be in 3rd or 10th position, whether to be accepted versus to be rejected. To better capture this specific property, we use 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. [9], adapted to our scenario.

[1]Please note that in the figure the scale is normalized

DEFINITION 1 (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.

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. In Figure 2 are depicted three different divergence curves resulting from the fact that (i) the two rankings are correlated; (ii) they are uncorrelated i.e. independent (the analytical results for this case is the straight line in the figure); (iii) they are inversely correlated.

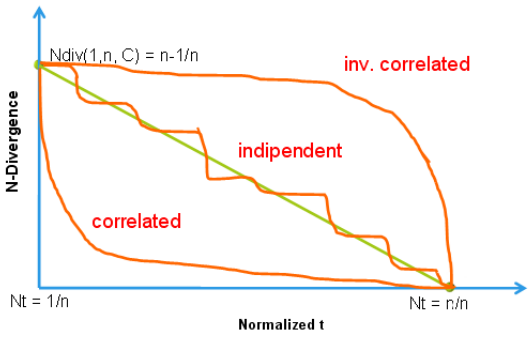

Figure 2: Examples of different divergence curves types

Divergence a posteriori. A program chair may decide that quality can be measured *a posteriori* (years after the completion of the review process), by measuring for instance the impact that the paper had e.g. by counting citations. Hence, we can compare this ranking a posteriori with the one produced by 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).

We can then apply the proposed divergence measure, using as rankings 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.

We can examine the difference in the ranking in the top k contributions, with $k < t$ where in this case $t = n$ with n is the number of accepted contributions. In Figure 3 is depicted the divergence between the ranking of the conference C1 and the ranking a posteriori given by the citation counts[2]. From Figure 3 we can notice that the two rankings are uncorrelated, that is the ranking of the conference is not reflected at all by the citation-count one.

We have found a similar result for all our actual data sets. One could argue that the peer review process is not good at predicting which papers will have more citations in the subsequent years, and then which papers will have more *impact*.

[2]The conference was held in 2003, so we were able to compute citations received in the subsequent years using Google Scholar

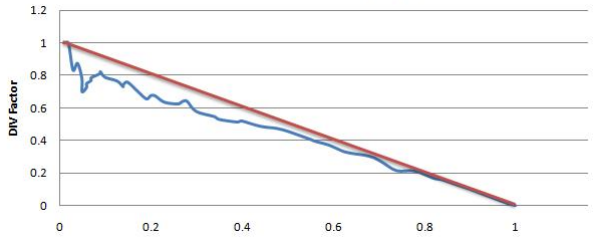

Figure 3: Normalized divergence computed for accepted papers at C1

On the other hand, one could say that the aim of peer review process *is not* the selection of high-impact papers, but simply to filter *junk* papers and accept only the ones above a certain quality threshold.

In Figure 4 we plot the divergence for conference C3 between the ranking *before* and *after* discussion among reviewers. We notice that the discussion consistently changes the fate of the contributions submitted, e.g. by looking at the 1/3 of the top papers in the two rankings, 32% of these papers differ.

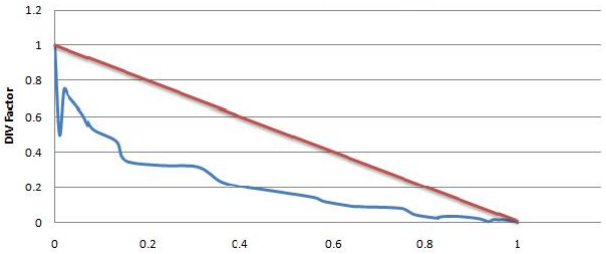

Figure 4: Normalized divergence computed for C3 after discussion

Divergence for robustness analysis. 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 would like the review process to be *robust* to minor variations in one of the marks. When reviewers need to select within, say, a 1-10 score range for a criteria, 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, see Section 5). With a robustness metric we try to assess how much precision is important in the mark assignment process. To this end, we apply a small positive/negative variation ϵ to each marks (typically multiple of the process's mark granularity 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.

Intuitively, what we do with the mark variation is a naive way to transform a mark into a random variable with a certain variance, reflecting the indecision of a reviewer on a mark. The higher the divergence, the lower the robustness.

This analysis is useful in order to assess the impact and the size of a perturbation on the mark value. This computation will be used both for the interpretation of the disagreement metric (how big is the disagreement and what effect does it have, see Section 4.3), as well as for the interpretation of the strength of a rating bias (how big is a detected bias) and, if any, for the "unbiasing" procedure (if eliminating the bias could lead to significant different results), see Section 5.

Figure 5 presents results for Conference C1 (1-10 scale, no half marks), the vertical line divides the accepted papers from the rejected ones. We firstly applied a perturbation of $\delta = 1$ (meaning that we randomly selected among three possible variations of the marks: -1/0/1), then $\delta = 2$ and finally $\delta = 3$. The analysis of the data suggests that is a quite robust process. Indeed, we can notice that with $\delta = 1$ the fate of only the 9% of the papers is affected (meaning that 9% of the papers that were initially accepted change their fate because of the perturbation) while with $\delta = 2$ the percentage is 15% and with $\delta = 3$ (a very large variationa) is 22%.

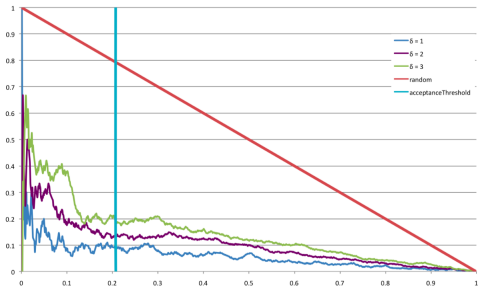

Figure 5: Example of a robust review process

In Table 1 we report results also for other conferences, ordered in increasing divergence values for $\delta = \epsilon$. Obviously, the δ changes depending on the mark granularity of the conference. We notice that no other process is more robust than conference 1. For conference 2, when $\delta = 1$ the fate of ca. 18% of the papers could change. Looking at these results one could argue that probably having a "shorter" scale (1-7 scale for conference 2) makes the process less robust, as possible indecisions of reviewers between marks have *more* impact on the final result of the review process. Also for conference 3 and 4 the process is not as robust as one should have expected. Indeed, even if their scale (0-5, with half marks) is, in principle, comparable to the one of conference 1 (1-10, no half-marks), because, as we noticed (see Figure 1), reviewers tend to not use half marks, an indecision between near full marks, can have an important impact on the fate of the contribution.

	granularity (ϵ)	$\delta = \epsilon$	$\delta = 2\epsilon$	$\delta = 3\epsilon$
C1	1,0	0,088	0,150	0,219
C3	0,5	0,143	0,230	0,306
C2	1,0	0,178	0,319	0,444
C4	0,5	0,180	0,290	0,360

Table 1: Computation of robustness for different conferences. Average standard error is ca. $0,004$

4.3 Disagreement

Here, we compute first how much the marks of a reviewer i differ from the marks of the other $r_z - 1$ reviewers for a specific criterion j and for a specific contribution z (Definition 2). 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 3), and, finally, over all the criteria (Definition 4).

DEFINITION 2. ***(Disagreement of a reviewer on a criterion and***

on a contribution)
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 $r_z - 1$ 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:

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

DEFINITION 3. ***(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 4. ***(Disagreement of a review phase)*** *Let q be the number of criteria in a review phase, then the disagreement over all the criteria is:*

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

The rationale behind these metrics is that in a review process we expect some kind of agreement between reviewers. While 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 the typical 3 reviewers per contribution) is to evaluate based on consensus or majority opinion. Intuitively, we assume that the ideal ranking is the one that we would obtain by having all experts review all papers. Then 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 come back to this issue also when discussing the quality vs effort trade-off in Section 6.

In Table 2 we collect in the first row the average normalized disagreement of review phase for four conferences. We normalize the disagreement value in order to make it comparable among different conferences. Both for comparison and to assist in the interpretation of the results, we also report in the same table, the average disagreement we have obtained in two simulations: (i) reshuffle experiment: where we have randomly exchanged the actual marks given by the reviewers; (ii) random experiment: where we have generated a new random (flat) distribution of marks in the available range of marks unrelated with the actual marks distribution in our data set. The reshuffle experiment mimics the case in which one reviewer is marking a certain number of contributions, but her marks are randomly given to other unrelated contributions, while her reviewed proposals get the marks of other randomly selected reviewers. So we are sampling from the correct marks distribution function, i.e. the actual one of the analyzed review phase, but we randomize the association between marks and contributions. We would have expected these reshuffling disagreements to be significantly higher than the actual one, since, again, we would have expected an higher correlation between the opinions of a team of experts, than a random one. However, this does not seem the case in our data sets, especially for conferences 3 and 4. Indeed, it seems that there is almost no difference between giving marks to specific proposals and randomly distributing such marks over all the proposals. On the other hand, the average disagreement is constantly lower than the random one (from 50 % to 60 % lower). This is expected since we would hope that a group of experts in a domain would tend to agree better than a completely random process.

	Conference 1	**Conference2**	**Conference 3**	**Conference 4**
Avg disagr.	0.276	0,306	0.260	0.219
Reshuffled	0.361	0,398	0.289	0.257
Random	0.434	0,488	0.436	0.449

Table 2: Average disagreement for all conferences.
Average standard error is ca. $0,005$.

Finally, we have tried to explore the behavior of the disagreement as a function of the number of reviews per reviewers.

We expected that more papers reviewers read, the less disagreement would be, as it is more easy for them to have an idea of the overall quality of the papers submitted and thus to judge them more appropriately. Figure 6 shows that indeed the average disagreement slightly decreases, but there is no a clear trend, as we expected.

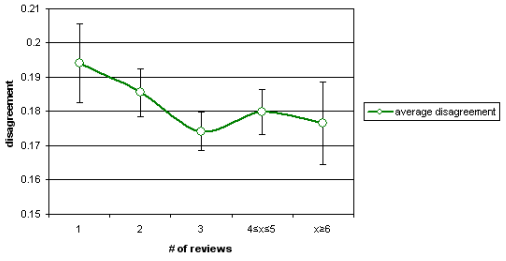

Figure 6: Average disagreement vs. number of reviews made

4.3.1 Band Agreement

Finally, we compute the *band agreement*. Our goal here is to study the correlation in the decisions of reviewers on very good and very bad papers. The approach is based on clustering conference marks in "bands" and measuring the probabilities of giving the mark from a particular band in condition that a mark from another "band" has already been given.

To this end, all marks were divided into non overlapping bands: (i) strong reject;(ii) weak reject;(iii) borderline;(iv) weak accept; (v) strong accept. We also computed the overall probability of a paper to belong to each group.

We have analyze the behavior of reviewers in two different conferences: conference 1 without threshold of marks for acceptance and conference 3 where such a threshold is present. Results are presented in Figure 7 and 8.

We note that for conference 1 (without threshold) if somebody gave a mark from the strong reject band then other reviewers will give a mark from the weak or strong reject "band" with higher probability (these probabilities are significantly bigger than the overall probability also shown in the figure). The same can be said about the "strong accept" band. So, in this case, we can say that reviewers "strongly" agree on very good and very bad papers.

For conference 3 (with threshold) the situation is different: overall probability is biased in the direction of "weak accept" band. Here, we could suggest that when there is a mark threshold reviewers tend not to give very low marks since they know that even a

mark from a "borderline" band and under threshold will eventually "kill" a contribution (they tend to be polite !). A more detailed analysis shows that if somebody gives a mark from the "strong reject" band, this increases the probability of giving marks not only from strong and weak reject bands (by 14% and 63% correspondingly) but also from borderline band (by 11%). In the "strong accept" set the probability of others giving a "weak accept" mark is 20% higher than the overall probability. So we can say that we have marks biased towards the "weak accept" band but reviewers still agree on very bad and very good contributions.

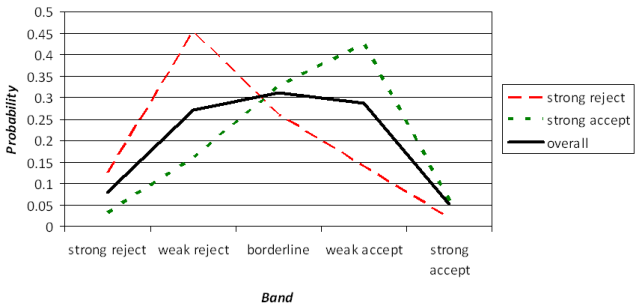

Figure 7: Band Agreement for Conference 1

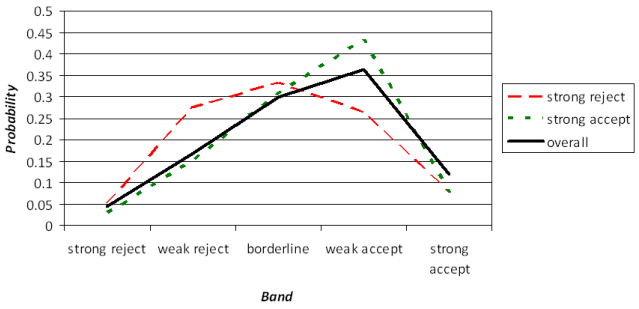

Figure 8: Band Agreement for Conference 3

4.4 Quality: lessons learned

From the exploration of the quality dimension we can derive a number of interesting findings:

- Monitoring mark distribution is useful since it is specific for the particular review process and marks scale. It is thus convenient to design and adapt the scale for specific purposes.
- The divergence metric is a practical metric to compare the actual ranking of the conference against various target ranking. It is useful to test the robustness of the process and the relevance of variation of the marks. The application of the divergence metric has uncovered (for the available data sets) that there is little correlation between the ranking of contributions obtained in the review process and the actual impact (citation counts) of the same contributions in the community.
- the measurement of the disagreement among reviewers is a useful metric to check and monitor the degree of process's randomness. There are some indications that a proper sample of reviews (in number) lower the disagreement. Moreover, having a high disagreement value means, in some way, that the judgment of the involved 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.

5. 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 a given set of experts. 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.

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. 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 here is not to try to identify what is the appropriate quality standard for the specific review process 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 diminish. This has nothing to do with the relative quality of the paper with respect to the other submissions. It is merely a bias 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 aim to identify, measure, and expose the most significant biases so that the review chairs can decide if they indeed correspond to unfair review results that need to be compensated before taking the final decision.

Different kinds of bias can be identified: rating, affiliation, topic, country, gender, clique bias. Using the available datasets, we focused on *rating* bias, namely when reviewers are positively (negatively) biased i.e. they *consistently* give higher (lower) marks than their colleagues who are reviewing the same proposal. The way to compute the bias value is very similar to that described for the disagreement metric:

$$\phi_i^j = \mathrm{M}_i^j - \mu_i^j \,. \tag{5}$$

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 given criteria (e.g. 3 in an evaluation scale form 1 to 5). This is computed by simply calculating the variance of the given mark for the specific criteria. 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.

Once biases are identified, a number of actions can be taken by

the review chairs. One could be to compensate for the specific paper under review with additional reviews. Another action could be to apply automatic or semi-automatic unbiasing algorithms: a simple one could be to 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, if we take all reviewers r that have a bias greater than b and that have done a number of reviews higher than n_r, and subtract b from all marks of r (or from the top-k biased reviewers), we can obtained a new debiased ranking. By comparing the obtained debiased ranking with the original ranking (for instance using the divergence metrics) we can assess the overall impact of the unbiasing procedure on the particular review process.

5.1 Biases identification and effect of unbiasing procedure

Applying the proposed rating bias metric, we were able to identify on actual review data for four conferences groups of potentially behavioral biased reviewers. These are all reviewers with an accepting or rejecting behavior greater then the minimum mark granularity. Table 3 reports for each analyzed conference (values here are not normalized in order to compare directly with the mark granularity): (i) the top accepting bias value (ii) the top rejecting bias value (iii) the percentage of accepting biased reviewers (iv) the percentage of rejecting biased reviewers (v) the mark granularity (vi) the number of reviewers

The table shows that even with simple metrics it is relatively easy to detect rating biases. Moreover, following the simple unbiasing algorithm outlined in the previous subsection, it is also possible to measure the effect of the bias on the review process. The bottom lines of Table 3 reports the percentage of affected papers.

	C1	C2	C3	C4
top accepting	3,4	1,52	2,8	1,3
top rejecting	-2,8	-2,06	-2	-2,6
> + |min bias|	5%	9%	4%	1%
< - |min bias|	4%	8%	2%	3%
mark granularity	1	1	0,5	0,5
no. Reviewers	943	103	151	382
Divergence at acceptance threshold	9%	11%	10%	6%

Table 3: Examples of computed accepting and rejecting biases in four review processes

5.2 Fairness-related metrics: lessons learned

From the above analysis we derive a number of interesting lessons or hints:

- the percentage of biased (accepting or rejecting behavior) is an important parameter to monitor by the review chairs and it is relatively easy to detect it through the application of simple metrics;
- it is also possible to device simple and automatic unbiasing procedures; they do not need to be applied as black boxes, but together with the analysis of the divergence between the actual ranking and the unbiased one. Divergence can provide quantitative data about the effect of biases on the final review process. Data that again can be used by the review chairs to better characterized and monitor their evaluation processes.

Our future work in the dimension of fairness-related metrics includes identifying - with similar models - other types of biases related to affiliation, topic, country, gender, clique bias and other aspects rather than limiting the analysis to accepting or rejecting biases. The challenge here is to collect and have access to the appropriate specific metadata.

6. EFFICIENCY-RELATED METRICS

Here 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.

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), 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 only raw review data (essentially marks).

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. Correspondingly, the contribution preparation effort is the number of submissions $N_C$multiplied by the average time spent in preparing each submission $\overline{t_w}$. Reviews and submissions can span across N_P phases. [3]

In the ideal case from a quality perspective, all reviewers are equally experts and 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.

Our hypothesis here is that beyond a certain review time threshold t_{rx} and contribution 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.

Essentially in all real cases (both conference, journal or project's proposals evaluation) the actual review process is far away from the above ideal case. It is therefore of interest to have some analysis and quantitative data and metrics to measure how far we are from the ideal case.

Informally, making the review process efficient requires reducing the effort minimizing the quality degradation. In our current work we have analyzed in some details the following parameters: (i) the number of reviews per paper; and (ii) the number of papers per reviewers;

6.1 Reducing the number of reviews

A line of investigation is around reducing the number of reviews for submissions whose fate is clear. Assume that the review process is structured in as many phases as the maximum number of review-

[3]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.

ers 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 (i.e. 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 and 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. 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 addition to the deterministic analysis mentioned above, which is conservative, we can also us perform a statistical analysis relying on the fact that reviewers' marks exhibit some correlation (see our analysis in Section 4). 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 band disagreement measures to help estimate the confidence associated to the estimate. Notice that implementing the above process requires either a multi-phase review, or requires to give to 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 formal analysis of such process is part of our current research work.

6.2 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. Here we investigate the efficiency of the review process from the view of an efficient ("optimal" number of papers per reviewer) review distribution among reviewers and the statistical accuracy of the review results.

6.2.1 "Optimal" number of papers per reviewer

Our working hypothesis is that in all evaluation processes there are different groups of contributions to evaluate, typically: immature, average, good and eventually excellent papers. We presume that if a reviewer estimates contributions only from one group, her evaluation scale will tend to expand, i.e. contributions from the same group could end up with very diverse marks. If a reviewer would have access to contributions belonging to different groups, the scale could be more realistic and probably more correct. Consequently, we would like to estimate how to distribute the papers among reviewers in a way such that each reviewer will have at least one paper from each group. The idea is to use statistical information about the distribution of the average marks for individual contributions (either an expected one or an historical one where available) in order to identify typical clusters of contributions for a given review process. Then use statistical approaches to compute the needed number of papers per reviewer in order to maximize the probability - with a specified confidence level - to have in the set of reviews at least one paper from each cluster.

In order to show a possible implementation of this idea, we first study a posteriori the distribution of the average marks for individual papers and for one criterion (for example for the most significant one among the marks of the conference). (Figure 9 shows such average marks distribution for one of the analyzed conference). This information is used to evaluate the general behavior of the sample as we use it as an estimation of the mean values density function.

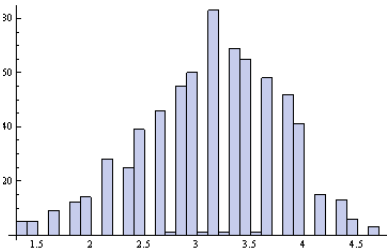

Figure 9: Histogram of the average marks for individual papers for conference 3

On the basis of such distribution, we then determine appropriate boundaries for papers clusters. In the following analysis, we chose three clusters - *immature*, *average* and *good/excellent* papers - with the following range $[0, 2.7]$; $[2.7, 3.7]$; $[3.7, 5]$ correspondingly.

As initial parameters in our statistical approach we have: (1) estimate of average mark distribution; (2) user's selection for cluster boundaries; (3) user's selection of desired confidence level $(1 - \alpha)$. The confidence level represents the probability that at least one paper from the group with minimal probability (p_{min}) will be assigned to a reviewer (i.e. α is the probability that in the set of papers for each reviewer there will not be the paper from the minimal probability group). Then if we provide that the reviewer reviews at least one paper from this group with the probability $1 - \alpha$, the papers from the others groups will appear with higher probabilities. If n is the desired value for the number of papers per reviewer, than - assuming that we have a large number of observations - we can estimate n as: $\alpha = (1 - p_{min})^n$ hence

$$n = log_{1-p}(\alpha). \tag{6}$$

If the number of observations (N) is not very large (i.e. the group probability changes significantly if we pull one paper out) then we can approximate the solution with the expansion:

$$\alpha = (1 - p_{min}) \left(1 - p_{min} \frac{N}{N-1}\right) \cdot \ldots \cdot \left(1 - p_{min} \frac{N}{N-n+1}\right) \tag{7}$$

In this case, we can't obtain an analytical expression for n, but we can estimate it using the following approximated computation:

1. Set initial parameters: average mark distribution, cluster boundaries, $1 - \alpha$.
2. Calculate the cluster distribution $\{p_1, p_2, ..., p_k\}$, where k is the number of paper cluster, $p_i = \frac{N_i}{N}$, $i = 1, ..., k$, N - total number of papers, N_i - number of papers in the ith group.
3. Find minimal p_i , $i = 1, ..., k$. Define it as p_{min}.
4. Obtain n from equation 7.

This approach can be used to estimate the quality of the peer review process dynamically (collecting and analyzing marks distribution from reviews as they are coming in during the evaluation process) or a posteriori (to check within which confidence level the initial assumption - each reviewer have had at least one paper from each cluster - has been met).

Results from an a posteriori analysis are reported in Figure 10 with real data from two conferences. Review chairs could have seen from the graphs that reviewers with a small number of papers have had a small probability of reviewing the papers from all the groups: for these conferences the probability of reviewing a paper from the "immature" cluster ranges from 0.45 to 0.51 for conference 3 and from 0.47 to 0.38 for conference 1, if the reviewer has received only 4 papers to review. In these conferences, in order to have a confidence level around 0.8-0.9 that each reviewer has seen a contribution from every cluster, each reviewer should have been assigned around 9-12 contributions for conference 3 and 10-14 for conference 1.

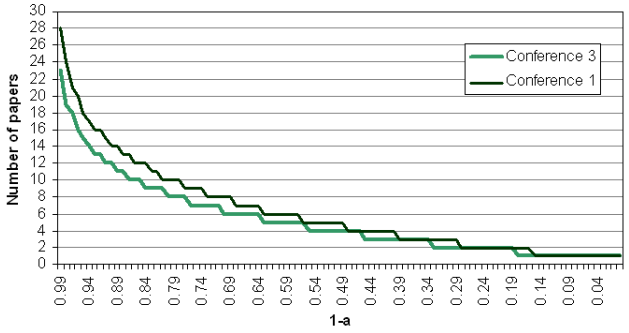

Figure 10: Number of papers per reviewer for different values of $1-\alpha$ ($0.99 \leq 1-\alpha \leq 0.01$)

6.2.2 *Evaluation of the accuracy of a review*

Another direction of investigation is the estimation of the accuracy of the marks obtained from the reviewers as a function of the number of papers reviewed. The approach is based on the classical Central Limit Theorem that states that for n (number of reviews) that tends to infinity - the distribution of average mark per paper approaches the normal distribution. So we can assume that the average review marks for a contribution $(x_i, i = 1, 2, ..., n)$ belong to the normal distribution $N(\mu, \sigma)$. The sample mean $\hat{\mu}(n) = \frac{1}{n}\sum_{i=1}^{n} x_i$ for each contribution is the estimation of the mathematical expectation value μ of the mark (i.e. its "ideal value") for a particular contribution. Our goal here is to choose n so that the error of estimation would be less then ε with probability $(1-\alpha)$:

$$P\{|\mu - \hat{\mu}(n)| < \varepsilon\} = 1 - \alpha.$$

so that μ falls into confidence interval $(\hat{\mu}(n) - \varepsilon, \hat{\mu}(n) + \varepsilon)$ with confidence level $(1-\alpha)$.

If σ is known, then confidence interval for unknown mathematical expectation μ with confidence level $(1-\alpha)$ can be computed analytically as:

$$\hat{\mu}(n) - u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu}(n) + u_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \qquad (8)$$

where $u_{\frac{\alpha}{2}}$ - quantile of the standardized normal distribution defined by the confidence probability $(1-\alpha)$, $\varepsilon = u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ - accuracy (limiting error) point estimate of mathematical expectation. The most common values of the confidence probability $(1-\alpha)$ are 0.9, 0.95, 0.99, 0.999. An analysis of the formula 8 shows that:

1. larger n correlates with smaller confidence intervals, hence the estimation is more accurate;
2. increasing the probability confidence $(1-\alpha)$ leads to increase of the confidence interval length.
3. if we fix accuracy ε and confidence probability $(1-\alpha)$ then from the formula $\varepsilon = u_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ we can obtain "optimal" (minimal) amount of sampling (i.e. n_{min}), that will provide the desired accuracy.

Unfortunately, in real cases σ is not known and cannot be estimated a priori. And even a posteriori, we only have estimates for σ, since in all realistic cases the number of reviews is limited and far from infinity! However, we can make use of such point estimates of σ (standard deviation) obtained either a posteriori or dynamically using current marks for a single contribution. This approximation will not lead to analytically correct results (σ is supposed to be known in the above method), but it allows to get an approximated estimate of the accuracy behavior depending on n.

We carried out a number of analysis with actual data from conference 1. Figure 11 shows the results we obtained using the computed (a posteriori) average value for the sample standard deviation σ=1.51.

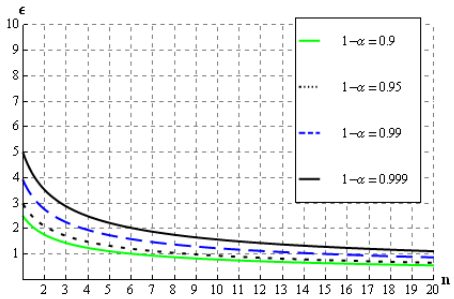

Figure 11: Accuracy versus amount of sampling for σ = 1.51

In this realistic case (average value for σ) in order to have a confidence level of 0.9 that the we have an accuracy around ± 1 absolute marks around the 'ideal' mark we would need around 6 reviews per paper. However, the figures clearly shows that improving this accuracy is going to be hard since the accuracy curves level off (decreases very slowly).

Another useful approach is to acknowledge that σ is unknown, and use statistical approaches for obtaining the confidence interval of an unknown mathematical expectation μ from a random variable X with unknown normal distribution $N(\mu, \sigma)$. Specifically we can write [2]:

$$\hat{\mu}(n) - t_{\frac{\alpha}{2};n-1} \cdot \frac{S}{\sqrt{n}} < \mu < \hat{\mu}(n) + t_{\frac{\alpha}{2};n-1} \cdot \frac{S}{\sqrt{n}}, \qquad (9)$$

where $t_{\frac{\alpha}{2};n-1}$ - quantile of the Student's distribution defined by the confidence probability $(1-\alpha)$ and by the number of degrees of freedom $n-1$; $\hat{\mu}(n)$ and S - point unbiased estimates of the normal distribution parameters; $\varepsilon = t_{\frac{\alpha}{2};n-1} \frac{S}{\sqrt{n}}$ - accuracy (limiting error) point estimate of mathematical expectation.

Given a specific sample of actual marks, equation (9) can be used to compute the confidence interval for μ while it cannot be used to find directly the required amount of sampling. However, we suggest, that it can be used to estimate (either in real-time or a posteriori) whether the number of reviewers for given paper is/was enough to determine μ with a defined accuracy, or if more reviews are/were needed. We note here, that similar but informal procedures are currently used in many review processes: for instance in the case were there is a relevant disagreement in the opinions among experts for a specific contribution, the review chairs can decide to include other reviewers in the evaluation. Our statistical approach provides

a mathematical base for such procedures and adds a more quantitative dimension with a detailed estimate of accuracy of the process for a given confidence level.

As an example, in Figure 12 we show how the suggested statistical approach could be used to estimate the accuracy "on-the-fly" during a review process for a particular contribution and adding more reviewers as a function of the desired target confidence level. The data for specific example are based from a contribution from conference 1 with 6 reviews and corresponding marks for the most relevant criterion, equals to $(5, 8, 7, 5, 4, 4)$. In the analysis we sorted marks by review date and computed the accuracy of the estimation (depending on the confidence probability) for first k reviews for k in the range $(3, 4, 5, 6)$, as if dynamically adding new reviewers (Figure 12).

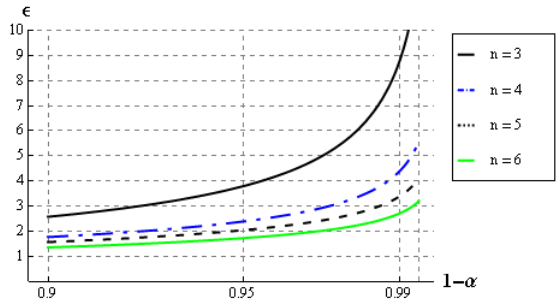

Figure 12: Accuracy of estimation versus confidence probability depending on the considered number of marks n. Marks sample = $(5, 8, 7, 5, 4, 4)$

The accuracy curves show the increase of accuracy in the process as a function of the confidence level (x-axis) and of the number of reviews added (individual curves). For instance, for a confidence level of 0.90 the accuracy in the estimate of the mark values improves from ca. ± 2.5 with three review to ± 1.2 when all 6 reviews are considered.

6.3 Efficiency-related metrics: lessons learned

Our preliminary investigations along the efficiency dimension has identify a number of statistical approaches able to provide information about a possible trade-off between effort and quality of the review process. Briefly:

- it is possible to provide real-time feed-backs on the status of submissions whose fate is clear
- it is possible to provide statistical indications on: (1) the "optimal" number of reviews per reviewer optimizing the overall coverage of the quality of the sampled set of reviews; (2) the accuracy of the evaluation process of a specific contribution as a function of the number of reviews done and of the desired confidence level.

7. CONCLUSIONS

In this work we have presented and discussed the results of the analysis on three different dimensions (quality, fairness and efficiency) of conference review process data. As already stated, we do not claim that our results are general, but we think that they can give useful hints in order to improve current peer review process. In the near future we want to extend the analysis to more conferences, also from fields different from computer science.

8. REFERENCES

[1] Ceci S.J., Peters D.P. Peer review: A study of reliability. *Change*, 14(6):44–48, 1982.

[2] David Brink. Statistics. Ventus Publishing ApS, 2008.

[3] Godlee F., Gale C.R., Martyn C.N. Effect on the quality of peer review of blinding reviewers and asking them to sign their reports a randomized controlled trial. *JAMA*, 280(3):237–240, 1998.

[4] Goodman S.N., Berlin J., Fletcher S.W., Fletcher R.H. Manuscript quality before and after peer review and editing at annals of internal medicine. *Annals of Internal Medicine*, 121(1):11–21, 1994.

[5] T. Jefferson, P. Alderson, E. Wager, and F. Davidoff. Effects of editorial peer review: a systematic review. *JAMA*, 287(21):2784–2786, 2002.

[6] Kassirer J.P.,Campion E.W. Peer review: Crude and understudied, but indispensable. *Journal of American Medical Association*, 272(2):96–97, 1994.

[7] Katz, D. S., Proto, A. V., and Olmsted, W. W. Incidence and nature of unblinding by authors: our experience at two radiology journals with double-blinded peer review policies. *Amer. J. Roentgenol.*, 179:1415–1417, 2002.

[8] Kendall M.G. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

[9] Krapivin M., Marchese M., Casati F. Exploring and understanding citation-based scientific metrics. In J. Zhou, editor, *First International Conference Complex 2009, Revised Papers, Part 2*, pages 1550–1563, Shanghai, China, February 23-25, 2009.

[10] Lawrence P.A. The politics of publication. *Nature*, 422(6929):259–261, 2003.

[11] Link A.M. Us and non-us submissions an analysis of reviewer bias, 1998.

[12] Smith R. Peer review: a flawed process at the heart of science and journals. *JRSM*, 99(4):178, 2006.

[13] Spier, R. The history of the peer-review process. *Trends Biotechnol.*, 20:357–358, 2002.

[14] Walsh, E., Rooney, M., Appleby, L., and Wilkinson, G. Open peer review: a randomised controlled trial. *Brit. J. Psychiat.*, 176:47–51, 2000.

[15] Wenneras C.,Wold A. Nepotism and sexism in peer-review. *Nature*, 387:341–343, 1997.