How healthy are software engineering conferences?

Bogdan Vasilescu\textsuperscript{a}, Alexander Serebrenik\textsuperscript{a}, Mark van den Brand\textsuperscript{a}, Ekaterina Pek\textsuperscript{b}, Tom Mens\textsuperscript{c}

\textsuperscript{a}Eindhoven University of Technology, PO Box 513, 5600 MB Eindhoven, The Netherlands
\textsuperscript{b}University of Koblenz-Landau, Universitätsstraße 1, 56070, Koblenz, Germany
\textsuperscript{c}Université de Mons, Place du Parc 20, 7000 Mons, Belgium

Abstract

In this article we study the health of software engineering conferences with respect to community stability, openness to new authors, inbreeding, representativeness of the PC with respect to the authors’ community, availability of PC candidates, and scientific prestige. We analyze this for 9 conferences over a period of more than 10 years. We observe differences between conferences with a wide scope and those with a more narrow scope. We also analyze which conferences perform better than others according to the aforementioned criteria. This knowledge can be used by prospective authors to decide in which conferences to publish and by conference steering committees or PC chairs to assess their selection process.

Keywords: scientometrics, software engineering, conferences

1. Introduction

In computing science, and especially in software engineering, scientific publications in international conferences (as opposed to journals) are often considered the most important way of disseminating research results [1]. The preference for conference publications is motivated by such arguments as: the young age and high dynamism of the field requiring shorter turnaround time between submission and publication than journals typically offer (to avoid results becoming obsolete before their publication) [2]; the increased visibility and publicity associated with presenting a paper and discussing it with peers [2]; the prestige associated with publishing at highly-selective (low acceptance rate) venues [3]; the increasing importance given to conference publications by decision makers assessing scientists, both in the USA [4] as well as Europe [3].
However, the fundamental role of conferences in computing science is not undisputed [2, 5–12]. The reported criticism is focused around the limited number of pages, too little time to revise a paper after receiving comments from reviewers, the ultimately higher impact of referenced peer-reviewed journal publications, and the increased volume of submissions. To keep the review quality high and the reviewer workload low, the latter requires the programme committee (PC) to grow larger. However, “the number of experienced reviewers does not appear to be growing at the same rate” [7], resulting in a “shrinking pool of qualified and willing PC candidates” [8].

Moreover, most software engineering conferences are based on a single-blind peer reviewing scheme, i.e., the reviewers know the names of the authors, but not vice versa. This may increase the risk of conferences becoming closed communities, and they may suffer to some extent from inbreeding. We understand openness as the readiness to accept newcomers, either as authors or PC members. Indicative of low openness—closed communities—are, e.g., inviting roughly the same group of people to the PC each year, or preferring to accept papers by known authors that have previously published in the same conference. In addition, we understand inbreeding as the favouritism of the PC towards papers (co)authored by PC members. While theoretically everybody can contribute to any conference, in practice some conferences tend to attract more “new faces” than others. Both problems are well-recognised [7, 8, 13], as can be witnessed by the following quotes. Crowcroft et al. [7] argue that “there is a distinct perception that papers authored by researchers with close ties to the PC are preferentially accepted with an implicit or overt tit-for-tat relationship”. Similarly, Birman and Schneider [8] question the quality of reviews, but suggest that “work by famous authors is less likely to experience this phenomenon, amplifying a perception of PC unfairness”. Therefore, it is useful to study to which extent and for which conferences such symptoms as inbreeding, closed nature, or shortage of PC candidates occur and, if so, what are the causes and consequences of this occurrence.

In this article we assess the health of software engineering conferences with respect to several factors: community stability (author and PC turnover), openness to new authors, inbreeding, representativeness of the PC with respect to the authors’ community, availability of PC candidates, and scientific prestige. In general, our findings suggest that software engineering conferences are healthy: balanced PC turnover (high enough to avoid inbreeding, yet low enough to ensure continuity and coherence), high openness to new authors (“new” in terms of both turnover with respect to previous years as well as not having published at that conference ever before), and moderate inbreeding (in terms of fraction of papers co-authored by PC members). Nonetheless, some conferences perform better than others according to the aforementioned criteria. This knowledge can be used by prospective authors when deciding to which conferences to submit their work, and by conference steering committees and PC chairs to assess their selection process.

The remainder of this article is organised as follows. Section 2 describes our selection of conferences, the metrics proposed to characterise the health factors,
and the data extraction process. Section 3 details the statistical analysis carried out and its results. Section 4 discusses related work, and Section 5 concludes.

2. Methodology

2.1. Data extraction

Numerous software engineering conferences are organised every year. Moreover, papers addressing software engineering topics are also being solicited by wider-scoped computer science conferences. In our study, we decided to focus on the conferences studied in [13]: European Conference on Software Maintenance and Reengineering (CSMR), Generative Programming and Component Engineering (GPCE), IEEE International Conference on Program Comprehension (ICPC), International Conference on Software Engineering (ICSE), IEEE International Conference on Software Maintenance (ICSM), and Working Conference on Reverse Engineering (WCRE). Of these 6 conferences, only ICSE has a wide coverage of the software engineering domain, while the others focus on a specific subdomain (maintenance, reverse engineering, program comprehension, and generative programming). Because of this, we decided to add three more conferences that also have a wide coverage of software engineering, namely IEEE/ACM International Conference on Automated Software Engineering (ASE), ACM SIGSOFT Symposium on the Foundations of Software Engineering (FSE), and International Conference on Fundamental Approaches to Software Engineering (FASE).

The data we have analysed was restricted to the main research track of each conference: number of papers submitted and accepted, the authors of accepted papers, and the composition of the programme committee. In order to facilitate replication of our study, we have created a webpage and a replication package containing all the data and tooling developed during our work. The dataset can be found at www.win.tue.nl/mdse/conferences/.

For all considered conferences, most of the data of all accepted papers and their authors was extracted from the DBLP records [14]. The extracted data covers a period of at least ten years, as can be seen in Table 1. Data about the composition of the programme committee and number of submitted papers to each conference was retrieved from the websites of each conference and online proceedings volumes. For earlier editions we used the Wayback machine\(^1\) to analyse no-longer-available websites as well as announcements posted by conference organisers in Usenet newsgroups.

Since we are integrating data from different sources, the names of authors and PC members are not necessarily consistent, while it is critical to know the identities of persons if we wish to check for signs of inbreeding. For example, Mark van den Brand is also known as Mark G. J. van den Brand or M. G. J. van den Brand. To match multiple aliases for the same person we performed

\[^1\text{archive.org/web/web.php}\]
identity merging [15, 16], and manually checked the results in a post-processing step.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ASE</td>
<td>1994</td>
<td>2012</td>
<td>No</td>
<td>55</td>
<td>Own charter(^2), SIGSOFT PC policy(^3)</td>
</tr>
<tr>
<td>CSMR</td>
<td>1997</td>
<td>2012</td>
<td>Yes</td>
<td>40</td>
<td>Yes; no guidelines on PC renewal</td>
</tr>
<tr>
<td>FASE</td>
<td>1998</td>
<td>2012</td>
<td>Yes</td>
<td>42</td>
<td>Yes(^4)</td>
</tr>
<tr>
<td>FSE(^5)</td>
<td>1993</td>
<td>2012</td>
<td>No</td>
<td>49</td>
<td>SIGSOFT PC policy</td>
</tr>
<tr>
<td>GPCE(^6)</td>
<td>2000</td>
<td>2012</td>
<td>Yes</td>
<td>37</td>
<td>Yes; no guidelines on PC renewal</td>
</tr>
<tr>
<td>ICPC(^7)</td>
<td>1997</td>
<td>2012</td>
<td>Yes</td>
<td>41</td>
<td>Yes(^8)</td>
</tr>
<tr>
<td>ICSE</td>
<td>1994</td>
<td>2012</td>
<td>Yes</td>
<td>117</td>
<td>SIGSOFT PC policy</td>
</tr>
<tr>
<td>ICSM</td>
<td>1994</td>
<td>2012</td>
<td>Yes</td>
<td>53</td>
<td>Yes(^9)</td>
</tr>
<tr>
<td>WCRE</td>
<td>1995</td>
<td>2012</td>
<td>Yes</td>
<td>43</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 1: Software engineering conferences considered in the study. Those conference with a wide coverage of the domain are indicated in boldface.

2.2. Metrics

Table 2 shows all metrics we have used. Some of these coincide with those used in [13]. The basic metrics count how many persons belong to the community (authors \(A\) and PC members \(C\)) of a conference for a particular year. To determine whether a conference community for a particular year is **stable**, we use two **sliding window** metrics \(NA\) and \(NC\), that count the number of Newcomers (Authors as well as programme Committee members) during several previous years. We also use their relative counterparts \(RNA\) and \(RNC\). To study **openness** of a community surrounding a conference, the sliding window metric \(PNA\) counts the number of Papers by New Authors, i.e., those papers for which none of the authors published at previous editions of this conference. Similarly, we use its relative counterpart \(RPNA\). To quantify **inbreeding** (influence of PC membership on paper co-authorship), we calculate the ratio of accepted papers co-authored by programme committee members who served at least once in recent years. To assess the **representativeness of**

\(^2\)www.ase-conferences.org/Charter.html
\(^3\)www.sigsoft.org/about/policies/pc-policy.htm
\(^4\)www.easst.org/fase/fasech
\(^5\)We do not distinguish between the years when FSE is colocated with the European Software Engineering Conference (and is known as ESEC/FSE) and the regular editions of FSE.
\(^6\)Formerly Semantics, Applications and Implementation of Program Generation (SAIG).
\(^7\)Formerly Workshop on Program Comprehension (WPC) and International Workshop on Program Comprehension (IWPC).
\(^8\)www.program-comprehension.org/ICPC-ProgramCommittee-v1.1.pdf
\(^9\)conferences.computer.org/icsm/PC-Guidelines.pdf
the PC for the community, we apply the ratio of PC members that never have co-authored a paper in several preceding editions of the same conference.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>[acronym in [13]]</td>
<td>Definition</td>
</tr>
<tr>
<td>Basic metrics</td>
<td></td>
</tr>
<tr>
<td>#A(c,y)</td>
<td>number of distinct Authors for conference c in year y</td>
</tr>
<tr>
<td>#C(c,y) [#PCmem]</td>
<td>number of PC members for conference c in year y</td>
</tr>
<tr>
<td>Stability</td>
<td></td>
</tr>
<tr>
<td>#NA(c,y,n)</td>
<td>Number of New Authors for conference c in year y that were not author in years y−n to y−1</td>
</tr>
<tr>
<td>#NC(c,y,n) [#(real)newPCmem]</td>
<td>Number of New PC members for conference c in year y that were not PC member in years y−n to y−1</td>
</tr>
<tr>
<td>RNA(c,y,n)</td>
<td>Author turnover = Ratio of New Authors for conference c in year y w.r.t. years y−n to y−1</td>
</tr>
<tr>
<td>RNC(c,y,n) [(real)newPCprop]</td>
<td>PC turnover = Ratio of New programme Committee members for conference c in year y w.r.t. years y−n to y−1</td>
</tr>
<tr>
<td>Openness</td>
<td></td>
</tr>
<tr>
<td>#PNA(c,y,n)</td>
<td>Number of Papers of conference c in year y by New Authors for which none of the co-authors has published at this conference in years y−n to y−1</td>
</tr>
<tr>
<td>RPNA(c,y,n)</td>
<td>Ratio of Papers (by New Authors) for conference c in year y for which none of the co-authors has published at this conference in years y−n to y−1</td>
</tr>
<tr>
<td>Inbreeding</td>
<td></td>
</tr>
<tr>
<td>RAC(c,y,n) [PCaccProp]</td>
<td>Ratio of accepted papers for conference c in year y co-authored by programme committee members who served at least once during years y−n to y</td>
</tr>
<tr>
<td>Representativeness</td>
<td></td>
</tr>
<tr>
<td>RCnA(c,y,n)</td>
<td>Ratio of PC members for conference c in year y that never have co-authored a paper at preceding instances of c between y−n and y−1</td>
</tr>
<tr>
<td>Sustainability</td>
<td></td>
</tr>
<tr>
<td>SR(c,y,n)</td>
<td>Ratio between the number of core authors that have not served on the PC in years y−n to y and #C(c,y).</td>
</tr>
<tr>
<td>Prestige</td>
<td></td>
</tr>
<tr>
<td>#SP(c,y) [#subm]</td>
<td>Number of Submitted Papers for conference c in year y</td>
</tr>
<tr>
<td>RA(c,y) [accRate]</td>
<td>Ratio of accepted papers for conference c in year y</td>
</tr>
<tr>
<td>CI(c)</td>
<td>Conference Impact of conference c = SHINE h-index for c between 2000 and 2012.</td>
</tr>
<tr>
<td>Workload</td>
<td></td>
</tr>
<tr>
<td>RL(c,y) [revCoeff]</td>
<td>Review Load for conference c in year y. ( RL(c,y) = #SP(c,y)/#C(c,y) )</td>
</tr>
</tbody>
</table>

Table 2: Metrics used to assess conference health.

To study the sustainability of the PC-candidates pool, we first introduce the notion of core author for a given conference, being a person who
frequently (co)authored papers published at that conference during the current or previous four editions. Specifically, we consider an author to be core author if either: (i) she has (co)authored papers in at least 3 out of the 5 most recent editions; or (ii) she has (co)authored papers in at least 2 out of the 3 most recent editions. A core author is therefore a very active member of the author community, who probably deserves to serve on the PC. To obtain a measure of the sustainability of the PC-candidates pool (its rejuvenation capacity), the SR (Sustainability Ratio) metric represents the ratio between the number of core authors that have not served on the PC at previous editions of this conference (PC candidates) and the size of the PC at that time. To quantify the review workflow experienced by PC members, we compute the ratio between the number of submitted papers and the size of the PC.

Finally, we relate an accepted prestige measure, conference impact, to the number of submitted papers and the acceptance ratio. We computed the conference impact CI based on the Simple H-INdex Estimator (SHINE) [17], a conference-specific variant of Hirsch’s h-index for quantification of an individual’s scientific research output [18, 19]. For a given conference c, CI(c) = SHINE(c, 2000, 2012) = 40 means that conference c has 40 papers, each with at least 40 citations between 2000 and 2012. The earliest SHINE data available is from 2000. Since computation of the h-index can be inaccurate for recent years (due to late propagation of citation information) we use the entire available history of conference citations since 2000 until 2012.

2.3. Data analysis

Using the R project for statistical computing [20] we visualise and statistically analyze the data to detect patterns and trends, with the aim to detect (counter-)evidence of conference health, e.g., signs of inbreeding or openness.

To detect monotone trends we compute Spearman rank correlation ρ between the values of the metrics and the time axis: since the latter is monotonically increasing, strong correlation (either positive or negative) indicates presence of a trend in the metric (either increasing or decreasing, respectively) [21]. Similarly, to verify presence of linear trends (between different metrics rather than a metric and the time axis) we compute the Pearson correlation r.

To verify claims such as “conference A tends to have higher values for metric m than conference B”, we need to compare multiple distributions of m (one for each of the 9 conferences). Traditionally, comparison of multiple groups follows a two-step approach: first, a global null hypothesis is tested, and then multiple comparisons are used to test sub-hypotheses pertaining to each pair of groups. The first step is commonly carried out by means of ANOVA or its non-parametric counterpart, the Kruskal-Wallis one-way analysis of variance.

---

10Some of the core authors may be PhD students who, despite being very active, are typically not officially invited to serve on the PC. Distinguishing between PhD students and PhDs goes beyond the scope of this article and is considered as future work.
by ranks [22]. The second step uses the $t$-test or the rank-based Wilcoxon-Mann-Whitney test [23], with Bonferroni correction [24, 25]. Unfortunately, the global test null hypothesis may be rejected while none of the sub-hypotheses are rejected, or vice versa [26]. Moreover, simulation studies suggest that the Wilcoxon-Mann-Whitney test is not robust to unequal population variances, especially in the unequal sample size case [27, 28]. Therefore, one-step approaches are preferred: these should produce confidence intervals which always lead to the same test decisions as the multiple comparisons. To this end, we employ the recently-proposed multiple contrast test procedure $\tilde{T}$ [29] using the traditional 5% family-wise error rate. $\tilde{T}$ is robust against unequal population variances. A more comprehensive discussion of $\tilde{T}$ goes beyond the scope of this article.

For ease of presentation (given 9 conferences, one has to report the results of $9 \times 8 \div 2 = 36$ comparisons per metric), we use the $\tilde{T}$-graphs proposed in [30] to summarise the results as a directed acyclic graph. For a particular metric, nodes of the graph correspond to conferences, and edges to results of pairwise comparisons (there is an edge from $A$ to $B$ if $A$ tends to have higher values for that metric than $B$). Because transitivity is respected by $\tilde{T}$ (as opposed to, e.g., the traditional pairwise Wilcoxon-Mann-Whitney tests [31]), we omit direct edges between $A$ and $B$ if there is a path from $A$ to $B$ passing through at least one other node $C$.

A special case of comparison of multiple distributions is the comparison of two distributions (e.g., wide-scoped conferences versus narrow-scoped ones). We need to test whether one of two samples of independent observations tends to have larger values than the other. The Wilcoxon-Mann-Whitney two-sample rank-sum test [32, 33] is not robust against differences in variance [27, 28], and the $\tilde{T}$ procedure as described above cannot be used to compare two distributions [29]. We therefore prefer the two-distributions equivalent of the $\tilde{T}$ procedure, i.e., we perform two sample tests for the nonparametric Behrens-Fisher problem [28], and compute confidence intervals for the relative effect of the two samples (using the R package nparcomp [34]). If the relative effect $p(A, B) > 0.5$ then $B$ tends to be larger than $A$. Therefore, we accompany visualisations of the evolution of different metrics by $\tilde{T}$-graphs (for a more rigorous view of the relations between different conferences), and we report the relative effect $p(\text{wide},\text{narrow})$ (to support claims about the relation between conferences when grouped into wide-scoped and narrow-scoped).

To avoid clutter when reporting $p$-values, regardless of the statistical procedure applied, we superscript the test result using the following convention: no superscript corresponds to $p \geq 0.05$, * corresponds to $0.01 \leq p < 0.05$, and ** corresponds to $p < 0.01$.

3. Results

3.1. Workload

We start our discussion of the health of software engineering conferences with an overview of their general state: Are software engineering conferences
attracting more submissions? Do the programme committees grow? Is there evidence of increased reviewing load?

Figure 1a displays the variation of the number of submitted papers per year \(\#SP(c, y)\), for each of the conferences. Increasing trends are confirmed for ICSE (\(\rho = 0.87^{**}\)), FASE (\(\rho = 0.76^{**}\)) and CSMR (\(\rho = 0.71^{**}\)): these conferences tend to receive more submissions each year. The other conferences exhibit less clear (increasing or decreasing) trends.

However, not all conferences receiving increasingly more submissions resort to increasing the size of their programme committees. Inspection of Figure 2a reveals increasing trends \(\#C\) for FASE (\(\rho = 0.85^{**}\)), ICSM (\(\rho = 0.78^{**}\)), GPCE (\(\rho = 0.77^{**}\)), CSMR (\(\rho = 0.76^{**}\)) and ICPC (\(\rho = 0.74^{**}\)): these tend to increase their programme committee size over the years. The other conferences exhibit less clear trends. This paradoxically suggests that narrow-scoped conferences that tend to receive less submissions than the wide-scoped ones (\(p(\text{narrow,wide}) = 0.888^{**}\)) tend to increase the programme committee size.

WCRE seems to behave more like a wide-scoped conference in this respect, while FASE—more like a narrow-scoped one.

Overall, when comparing the wide-scoped conferences to the narrow-scoped ones, we observe that the former receive more submissions but have smaller PCs, hence higher review load than the latter. Some program committee members engage external reviewers, i.e., the true review load might be lower than the ratio \(RL\) of the number of submissions and the number of PC members.

Figure 3a displays the variation of the review load \(RL\). ICSE displays an increasing trend, since it receives increasingly more submissions each year, but its PC tends not to increase accordingly. Other trends are visible for ASE (increasing, \(\rho = 0.75^{**}\)), and for GPCE (decreasing, \(\rho = -0.65^{*}\)). Recognition
Figure 2: Variation of the number of programme committee members $\#C(c, y)$ per year. Wide-scoped conferences (squares) have smaller PCs than narrow-scoped ones (circles): $p(\text{narrow,wide}) = 0.317^{**}$.

Figure 3: Variation of the review load $RL$ per year. Wide-scoped conferences (square) have higher review load than narrow-scoped ones (circle): $p(\text{narrow,wide}) = 0.964^{**}$.
of an extremely high review load for ICSE has lead the program co-chairs and steering committee of ICSE to adopt the Program Board model in 2014. This model follows a two-phase review process: first, a broad PC reviews the submissions, then a much smaller Program Board initiates, monitors, and guides the discussions with the PC members. In this way a balance can be found between a reduced review load and high quality of reviews. The Program Board model has recently been used the International Requirements Engineering Conference (RE) and the International Conference on Model-Driven Engineering Languages and System (MoDELS).

3.2. Stability

To assess the community stability we investigate two turnover aspects: the PC turnover and the author turnover.

**PC turnover**

PC turnover reflects one of the aspects of conference stability. Inviting PC members from previous editions helps to ensure continuity and coherence. However, conferences are frequently subject to charters or guidelines that require PC renewal. Therefore, conferences must strike a balance between PC turnover and continuity.

When studying the PC turnover w.r.t. the previous year $RNC(c, y, 1)$, we observed a wide variation (Figure 4a): the lowest observed PC renewal ratio was 12% for WCRE 1998, and the highest was 93% for GPCE 2007. Using Spearman rank correlation no clear monotonic trends could be inferred.
for $RNC(c, y, 1)$. To check whether the chosen duration (one year) affects our results, we repeated the analysis by looking at a longer period of 4 previous years. The PC turnover rate $RNC(c, y, 4)$ showed similar behavior: no clear trend, and a big range of values for the metric (from 9% for ASE 2001 and WCRE 2010, to 85% for GPCE 2007).

Next, we use $\tilde{T}$-graphs to compare distributions of RNC-metrics for different conferences. Two groups of conferences become apparent in Figure 4b: the wide-scoped ICSE, FSE and FASE and the narrow-scoped GPCE, have consistently higher values of $RNC(c, y, 1)$ than the five other considered conferences (WCRE, ICPC, ASE, ICSM and CSMR). We conjecture that presence of GPCE in the high-turnover group of conferences is either due to relevance of the GPCE topics to a broader scientific community, or due to the small PC of GPCE: it is easier to renew a large fraction of a small rather than a large PC. However, while the PC of GPCE is smaller e.g., than that of ICSE, it is not significantly different in size than the PC of FSE or FASE (as resulted from applying $\tilde{T}$ to the values of $#C$).

Wide-scoped conferences ICSE, FSE and FASE, and narrow-scoped GPCE have consistently higher PC turnover than the other conferences. GPCE’s high PC turnover is not due to a smaller PC. ASE (wide-scoped) appears to be an outlier with respect to the other wide-scoped conferences.

**PC charter availability**

The PC turnover rate often depends on external factors, such as the presence of some implicit or explicit policy or charter requiring part of the PC to be renewed every year. Conference charters commonly recommend that no PC member should serve four consecutive terms. The ACM SIGSOFT policy, applicable to ICSE, FSE and ASE, requires at least one-third of the PC members to change each year. Our results confirm that ICSE and FSE (as well as FASE and GPCE) always conform to this requirement of the ACM SIGSOFT policy. While ASE should also adhere to this policy, this is true for only 10 out of 18 editions considered, with the most recent noncompliance being in 2010.

ICSE and FSE always conform to the “at least one third” PC renewal policy, while ASE does not.

The official charter of FASE (established at FASE 2004) requires that about 50% of the PC members should be chosen from among PC members of the previous two editions. Let us loosely interpret “about 50%” as the interval between 40% and 60%. Although FASE did not always satisfy this requirement, it has been adhering to this charter regulation since 2009, and the threshold of 50% has always been exceeded since the establishment of the charter.

**Author turnover**

The author turnover is another indicator of conference stability. One can expect that conferences attract local researchers, that might not be ready to
participate in the subsequent edition organised at a different location. However, one can also expect a relatively stable group of “core” researchers that are likely to contribute to a number of conference editions. Similar tradeoffs as with PC turnover are in place. On the one hand, a very unstable community might fail to achieve a critical research mass. On the other hand, a very stable community, in which the same authors publish over and over again, can be a sign of inbreeding.

We observed that all considered conferences are very dynamic and have high author turnover $RNA(c, y, 1)$ with respect to the previous edition: from 2000 onwards values exceed 70% and can reach as high as 97% for FASE 2003, suggesting high openness to new authors. Overall, the lowest author turnover rate of 58% is observed for ICPC in 1999, and the highest one of 100% for FASE 1999. Given these high values, we hypothesise that the “new” authors are not necessarily new but rather returning after a short period of absence.

Figure 5a visualizes $RNA(c, y, 4)$ to take the four preceding years into account. We still observe high turnover (ranging from 49% for WCRE in 2006 to 86% for ASE in 2006), but the results become less extreme. The $T$-graph of Figure 5b further reveals differences between the wide-scoped ASE and FASE, and the narrow-scoped ICPC, WCRE, ICSM and CSMR: the wide-scoped conferences tend to have higher author turnover. This is not surprising since wide-scoped conferences have a larger pool of tentative authors that can contribute to them. This finding is concurrent with the observation that such software engineering conferences as ASE, CAV, FASE, FM, FSE, ICSE, ISSTA are quite interdisciplinary [35].
All conferences have high author turnover: since 2000 there are more than 70% new authors with respect to the previous edition, and more than 50% with respect to the previous four editions. Wide-scoped conferences tend to have higher author turnover than the narrow-scoped ones.

We emphasize that attracting new authors and renewing the PC (cf. Section 3.2) do not necessarily prevent inbreeding in wide-scoped conferences (e.g., the new authors could also be PC members). We therefore further investigate openness and inbreeding in Sections 3.3 and 3.4, respectively.

3.3. Openness

To evaluate openness, i.e., the ability of a conference to attract new authors, we study the evolution of $RPNA(c, y, 4)$—the fraction of papers published at conference $c$ in year $y$ for which none of the co-authors has previously published at conference $c$ in the 4 preceding years (Figure 6a). The lower this value, the less “open” the conference is to new authors. By focusing on the percentage of papers rather than the percentage of authors, we can avoid the phenomenon of “new faces”, e.g., junior co-authors of researchers that have already published at the conference. Moreover, by looking at 4 previous years only, we remove the impact of the amount of historic data on the evaluation of openness.

At the high end of the scale (more open communities) the $\tilde{T}$-graph of Figure 6b reveals ASE, FASE and GPCE: ASE is consistently more open than ICPC, WCRE and ICSM. Similarly, FASE is consistently more open than ICPC, WCRE, ICSM, CSMR and ICSE. At the low end of the scale (more closed communities) no statistically significant ranking can be inferred between any of the
conferences. Overall, the clearest trend is exhibited by ICSE: over the years, the percentage of papers for which none of the co-authors has ever published at ICSE in the preceding four editions is decreasing ($\rho = -0.675^{**}$). Hence, ICSE is becoming increasingly less open.

ASE, FASE and GPCE are very open communities (have a low entrance barrier). In contrast, ICSE is becoming increasingly less open.

3.4. Inbreeding

To evaluate inbreeding we study the evolution of $RAC(c,y,0)$, the fraction of papers co-authored by PC members in the same year (Figure 7a). Again, the $\bar{T}$-graph reveals differences between the wide-scoped ICSE, FSE, FASE, and ASE, and the narrow-scoped ICSM, CSMR, WCRE, ICPC, and GPCE when considered as two groups: the wide-scoped conferences tend to be less inbred. The values range between 0% (no inbreeding at all) for FASE 2002 or GPCE 2007 and 71% for WCRE 2006 (high inbreeding). Overall, WCRE, ICPC and ICSM tend to be the most inbred (i.e., tend to have higher values of $RAC(c,y,0)$ than other conferences), while GPCE, FSE and FASE tend to be the least inbred. However, both FASE ($\rho = 0.83^{**}$) and GPCE ($\rho = 0.66^{*}$) are becoming increasingly inbred. Repeating the analysis with a longer time window using $RAC(c,y,4)$ confirms the previous ranking.

Figure 7: Wide-scoped conferences (squares) usually have lower $RAC(c,y,0)$ values (are less inbred) than the narrow-scoped ones (circles): $p(\text{narrow,wide}) = 0.174^{**}$.

WCRE (on average 47% of the papers accepted each year are co-authored by PC members), ICPC (45%) and ICSM (38%) tend to be the most inbred conferences. In contrast, GPCE (14%), FSE (17%) and FASE (9%) tend to be the least inbred. Overall, wide-scoped conferences tend to be less inbred.
Systä et al. [13] have observed negative linear correlation between $RAC(c, y, 0)$ and $RNC(c, y, 1)$: “the less there is PC turnover, the greater is the proportion of PC papers among the accepted papers”. However, they also noticed that the negative correlation does not necessarily hold for individual conference series, e.g., for CSMR the correlation was found to be reversed. We have replicated their study for our larger set of conferences and longer period, and we have observed a similar phenomenon (Figure 8): $RAC(c, y, 0)$ and $RNC(c, y, 1)$ show a moderately-strong negative linear correlation ($r = -0.60^{**}$). The moderate correlation occurs in particular for the narrow-scoped ($r = -0.55^{**}$) rather than the wide-scoped ($r = -0.33^{**}$) conferences. At the level of individual conference series, only GPCE ($r = -0.56$) confirms this trend, but the correlation is statistically significant only at 90% confidence level.

![Figure 8](image)

Figure 8: For narrow-scoped (circles) rather than wide-scoped (squares) conferences, higher PC turnover is associated with smaller fractions of PC papers among the accepted papers.

When observing the whole set of conferences or focusing on narrow-scoped conferences only, higher PC turnover is associated with lower inbreeding. However, the claim does not typically hold at the level of individual conference series.

3.5. Is the PC representative of the conference community?

Ensuring a right balance between continuity and renewal is not the only sign of a healthy PC. We believe that PC members should be representative of their respective communities, i.e., they should largely be established authors within those communities. However, not all PC members should be expected to have
published at a conference before. For example, PC chairs often invite some PC members with industrial affiliation or background, who typically do not publish. In the case of FASE, the charter explicitly mentions that “the PC should include at least 10% of members with industrial affiliation or background”. Similarly, senior researchers may be invited to serve on the PC, even if they prefer to publish at more prestigious venues or in journals instead of conferences. Nevertheless, we expect high representativeness of the PC, given that all considered conferences are well-established and we analyse at least ten years of history.

To investigate this, we studied $RCnA(c, y, 4)$, the fraction of PC members not having (co)authored papers at conference $c$ during any of the preceding four editions. The higher the values, the less representative we claim the PC to be. Overall, we observe a wide range of values (Figure 9a): the lowest is 12% for ICPC 2001 (the earliest ICPC edition for which the metric can be computed), and the highest is 85% for FSE 1997. The $\tilde{T}$-graph of Figure 9b reveals that values of $RCnA(c, y, 4)$ for WCRE are lower than for all other conferences except ICPC; ICSM and ICPC are both lower than any of CSMR, FSE, ASE and FASE. Narrow-scoped conferences have more representative PCs than wide-scoped ones. For example, the WCRE PC is consistently more representative of its community than the PCs of all other conferences except ICPC. ICSM and ICPC also have representative PCs.

ICPC exhibits the clearest increasing trend ($\rho = 0.874^{**}$), suggesting that its PC is becoming increasingly less representative of the ICPC community. However, one should also take into account the fact that the time series for ICPC started from a value that was much lower than for the other conferences,
and that the highest value (for 2012) still falls within the same range as for the other conferences.

ICPC starts off in 2001 by having the most representative PC out of all conferences, but its PC is becoming increasingly less representative over the years.

3.6. Is the pool of PC candidates sustainable?

We next investigate whether the conference communities comprise core authors that have not served on the PC at previous editions of the conference. Presence of such unsung heroes among the core authors can be seen as a measure of conference health: such core authors, either senior researchers or PhD students, can and should serve as a pool of candidates for PC membership in the future; moreover, they can contribute to increasing the representativeness of the PCs for their respective communities, should this be desired. Alternatively, the low number of unsung heroes can be seen as a sign of degeneration of community/field. However, there is high variation in the number of PC members and number of authors between the different conferences (e.g., in 2007 ICSM had 89 PC members, while ICSE had only 35). Therefore, we study variation across conferences of $SR(c,y)$, the average number of unsung heroes per PC member.

Using the Spearman correlation we confirm (Figure 10a) increasing trends for ASE ($\rho = 0.908^{**}$), FSE ($\rho = 0.829^{**}$) and ICSE ($\rho = 0.675^{**}$)—they have increasingly more sustainable pools of PC candidates; in contrast, GPCE ($\rho = -0.929^{**}$), WCRE ($\rho = -0.824^{**}$), CSMR ($\rho = -0.645^{*}$) and to a lesser extent ICSM ($\rho = -0.59^{*}$) exhibit decreasing trends—finding qualified PC candidates is becoming more challenging.
There is high potential to renew the PC members for ICSE, FSE and ASE from within the core authors publishing at these conferences. In contrast, CSMR, WCRE, ICSM and GPCE have increasingly lower potential to renew the PC from within their author communities.

3.7. Prestige

It is generally believed that the number of submissions to a conference is directly proportional to its scientific impact [36]. To verify this intuition, we study the relation between the conference impact factor $CI(c)$ and the number of submissions $#SP(c,y)$. Since $CI(c)$ has a single value per conference series, we contrast it against the mean number of submissions. Computing the mean is meaningful since for each conference series the distribution of $#SP(c,y)$ over the years is close to normal (the Shapiro-Wilk test fails to reject the normality hypothesis at 95% confidence level). We observe very strong positive and statistically significant linear correlation ($r = 0.93^{**}$), confirming that more prestigious conferences attract more submissions.

The higher the scientific impact of a conference, the more submissions it attracts.

It is also generally believed that the acceptance rate of a conference is inversely proportional to its scientific impact [36]. To verify this intuition, we study the relation between the conference impact factor $CI(c)$ and the mean acceptance rate $RA(c,y)$, following the same reasoning as above. We observe moderate negative linear correlation ($r = -0.59$), suggesting that conferences with higher acceptance rates indeed have lower scientific impact. This conclusion is concurrent with the findings of Chen and Konstan [1] based on a study of 600 ACM conferences.

Higher impact conferences tend to have lower acceptance rates.

3.8. Discussion

In this section we combine and interpret the findings for each considered conference. Furthermore, when some of our indicators reveal that conference health is threatened we propose strategies to improve it.

CSMR has a low author turnover. Together with average openness, this suggests a relatively stable community. Moreover, the PC turnover is low, the inbreeding and representativeness are neither high nor low. This suggests that the CSMR authors and PC members are relatively disconnected. As a remediation strategy one can consider inviting some of the “unsung heroes” to join the PC. However, sustainability of the PC pool candidates decreases, i.e., finding qualified PC members becomes more difficult.

ICSM, ICPC and WCRE exhibit similar trends: all these conferences have low author and PC turnover and average openness. The PC members actively contribute to the paper body implying that the PC is representative for its author community but also high inbreeding. Sustainability of the PC for these
conferences is low and exhibits a decreasing trend. Improving on the PC sustainability as well as increasing the PC turnover could be achieved by reducing the size of the PC. This would, however, increase the PC members’ workload and, therefore, might endanger the quality of the reviews and negatively affect the scientific impact of the conference. Still, we have observed that not only ICSE, but also ASE and FSE with CI-values comparable to those of ICSM have much higher ratios of the number of submissions to the number of PC members. Hence, higher reviewer load is not necessarily detrimental for the scientific impact and might be beneficial for ICSM, ICPC and WCRE. As a countermeasure against inbreeding and closeness these conferences may consider changing the tradition single-blind review by a double-blind review scheme as this scheme is fairer to authors from less prestigious institutions, that are unlikely to be among the PC members [37]. Another alternative might consist in employing a double-open review scheme, although it may increase the reluctance of prospective PC members to join the PC.

GPCE, as opposed to CSMR, ICSM, ICPC and WCRE, has a high PC turnover and low inbreeding, suggesting that the PC members do not tend to publish at GPCE. However, the author turnover is low. This might be indicative of, on the one hand, relevance of the GPCE topics to the broader scientific community, or alternatively, of the failure to establish a well-defined core group. Sustainability of the PC for GPCE is low and exhibits a decreasing trend.

ICSE is a highly prestigious conference with high author and PC turnover, suggesting that both authors and PC members can be selected from a large pool of candidates. ICSE, however, appears to become increasingly more difficult to enter, which in the long run might put the sustainability of the candidate-author pool in jeopardy. The ICSE steering committee is aware of this, since they decided to adopt a Program Board model as of 2014.

Similarly to ICSE, FSE has high author and PC turnover as well as low inbreeding. Moreover, for FSE the sustainability of the PC is high and shows an increasing trend. Openness of FSE fluctuated greatly in early 2000s and seems to stabilize on a relatively low level, suggesting that the same warning as for ICSE seems to apply.

FASE, in general, exhibits typical features of wide-scoped conferences, e.g., high author and PC turnover, little inbreeding and a pool of PC candidates that seems to be highly sustainable in recent years. Unlike ICSE and FSE, FASE is very open, suggesting the future sustainability of the conference.

Unlike FASE, ASE has a low PC turnover but similarly to FASE, it is very open and has a high author turnover. Similarly to CSMR, the PC community of ASE seems to be disconnected from the author community (the representativeness is low). However, unlike CSMR, the sustainability of the PC candidates pool is higher, and therefore, we suggest inviting some of the “unsung heroes” to join the PC.
4. Related work

In this article, we built upon the work by Systä, Harsu and Koskimies [13], replicating and extending their inbreeding study of 6 software engineering conferences (ICSE, ICSM, ICPC, CSMR, WCRE and GPCE) observed during the 2004-2009 period\(^{11}\). Our study takes into account more conferences, uses a wider range of metrics, considers longer time periods, and uses T-graphs to perform pairwise comparison of metrics across conferences. We assess the health of software engineering conferences with respect to several factors (community stability, openness to new authors, inbreeding, representativeness of the PC with respect to the authors’ community, availability of PC candidates, and scientific prestige), and we track how each health factor evolves over time for each of the considered conferences. Additional health criteria could have been used, such as the quality of the published conference papers, or the amount of self-citations, but these fall outside the scope of the current work.

Similar in spirit, although not focusing on software engineering conferences\(^{12}\), is the work by Biryukov and Dong [38]. The authors investigate how the communities represented by different research subfields within computer science as well as the corresponding conferences are evolving and communicating to each other. They use DBLP data to survey the development of authors’ careers, and extract features that can help distinguish between conferences of different rank. For example, *population stability* (akin to our discussion of author turnover from Section 3.2) is recognised as “a candidate feature that helps to distinguish between the top and non-top venues”. They find that lower-rank conferences are characterised by higher turnover (typically the newcomers constitute about 75-85% and the leavers up to 88% of all authors) and high percentage of pure newcomers among the newcomers (about 75%), the latter suggesting high openness (cf. our discussion in Section 3.3). Our results computed suggest that wide-scoped conferences tend to have higher author turnover than narrow-scoped ones.

Also related, although again not focusing on software engineering conferences, are the works of Elmacioglu and Lee [39], Zhuang et al. [40] and Sakr and Alomari [41], who recognise the impact of PC quality on the conference quality. Elmacioglu and Lee [39] extract information about PC composition for a number of conferences from Calls for Papers published on DBWorld\(^{13}\), and construct a collaboration graph for authors thereat from the ACM Guide\(^{14}\). By dividing the set of analysed conferences into two groups (reputable and questionable\(^{15}\)), they show that (i) reputable conferences tend to have smaller PC sizes than less reputable ones (28.8 members on average as opposed to 69.6); (ii) PC members of reputable conferences typically have more publications than those of less reputable ones (complementary to our notion of representative-

\(^{11}\)The GPCE data is collected over the 2002-2009 period.

\(^{12}\)The only conference in common between us is FSE.

\(^{13}\)http://research.cs.wisc.edu/dbworld/

\(^{14}\)Currently the ACM Digital Library.

\(^{15}\)Distinction based on personal experience of Elmacioglu and Lee.
ness, cf. Section 3.5); and (iii) most reputable conferences have PC members with high closeness values in the collaboration graph (the more central a node is, the lower its total distance to all other nodes) on average. In our case all considered conferences are well-established and would likely be labeled as “reputable”. However, our results using the $T$-procedure reveal that ICSM has consistently the largest PC among the considered conferences (63.5 members on average), while FASE, GPCE and FSE have consistently the smallest ones (21.3, 23.1 and 26.7 members on average, respectively).

Similarly to us, Sakr and Alomari [41] also argue in favour of PC renewal: “it is quite unhealthy to have a fixed or slightly different list of members in the program committees for the different venues”, since this “may have intended or unintended negative effects in the fairness of evaluating the research contributions or in the quality and variability of the conference programs”. They analyse the composition of the PCs for four top-tier and prestigious database conferences (SIGMOD, VLDB, ICDE, EDBT) over a period of 10 years (2001–2010), and report the percentage of overlap in the PC between the different editions of each conference. Although their metric is similar in spirit to our $RNC(c, y, n)$ for different values of $n$, we cannot directly compare our results since they use a slightly different definition. Nonetheless, both their analysis and ours suggest that from the PC composition viewpoint, the considered conferences are relatively healthy.

Inbreeding has also been studied by Inanc and Tuncer [42], albeit with a different meaning. While we consider inbred the conferences for which PCs favour acceptance of papers submitted by PC members, they refer to a situation wherein PhDs are employed by the very same institution that trained them during their doctoral studies (denoted academic inbreeding). Using a dataset of scholars from Turkish technical universities, the authors show that inbreeding has negative consequences, affecting apparent scientific effectiveness as measured by one’s $h$-index.

Our “health assessment” of software engineering conferences can be further put in the context of quality evaluation (i.e., ranking) of scientific venues. We have used SHINE (the Simple H-INdex Estimator [17]) to rank the conferences and show e.g., that higher-impact conferences attract more submissions but tend to have lower acceptance rates. Numerous alternative approaches to ranking scientific venues have been proposed (e.g., [36, 43–50]), but they fall beyond the scope of this paper. For example, da Silva et al. [48] propose a ranking scheme for scientific conferences based on ranking the PC members. Their rank measure is based on the $h$-index of the PC (i.e., the maximum number $x$ of PC members such that each PC member has $h$-index [18] at least $x$) as well as the inequality (spread) of the $h$-indices of the PC members, computed using the Gini index [51]. The Gini index is one of a number of equivalent econometric inequality indices [52–54] commonly used in software engineering research.

\footnote{Their metric is defined as the ratio of the number of PC members in common between two editions and the number of distinct PC members of the same two editions.}
5. Conclusions

The goal of this article was to assess how the health of 9 software engineering conferences evolves over time, with respect to a variety of criteria: stability and representativeness of the PC, openness to new authors, inbreeding, availability of PC candidates and scientific prestige. We grouped conferences according to whether they had a wide or a narrow scope, and analyzed if these groups performed differently w.r.t. the aforementioned criteria.

Healthy conferences should strive for a stable community. PC turnover should be high enough to avoid inbreeding, while keeping PC members from previous editions to ensure continuity and coherence. Similarly, healthy conferences should strive for a good balance in author turnover: open enough to allow new authors to join, yet selective enough to keep the good reputation of the conference. Author turnover was consistently high for all considered conferences. In addition, author turnover was higher for wide-scoped conferences than for narrow-scoped ones.

We studied the representativeness of the conference PC by measuring the fraction of PC members that published papers at that particular conference before. A higher representativeness coincided with a higher inbreeding, measured by a high fraction of papers co-authored by PC members in the same year. A negative correlation could be observed between inbreeding and PC turnover, but not at the level of individual conference series.

We verified some general beliefs about the effect of scientific prestige of conferences and were able to confirm that the higher the scientific impact of a conference, the more submissions it attracts. We also observed that more prestigious conferences tend to have lower acceptance rates.

We could not identify a “holy grail” for the health of a conference, i.e., a small set of metrics that are universally responsible for conference health. Instead, we presented a range of analyses and countermeasures that can be useful for steering committees, prospective authors and researchers in different ways. Different strategies could be used to increase conference health: inviting “unsung heroes” to join the PC, reducing the size of the PC, replacing the traditional single-blind review process by either a double-blind or double-open review process, and reducing reviewer load by resorting to a Program Board review model.

Acknowledgements

We are grateful to Dr. Frank Konietschke for providing us with the (yet to be published) implementation of the $\tilde{T}$ procedure. In addition, this research has been partially supported by research project NWO 600.065.120.10N235 financed by the Dutch Science Foundation (Nederlandse Organisatie voor Wetenschappelijk Onderzoek, NWO). Moreover, part of this research has been carried out during the second author’s stay at the Université de Mons, supported by grant BSS-2012/V 6/5/015 of the Fonds de la Recherche Scientifique (F.R.S-FNRS). None of the funding agencies was involved in study design, in the collection,
analysis and interpretation of data, in the writing of the report, and in the
decision to submit the article for publication.

References


