Science of Computer Programming 89 (2014) 251–272

Contents lists available at ScienceDirect

Science of Computer Programming

www.elsevier.com/locate/scico

How healthy are software engineering conferences?

Bogdan Vasilescu [a], Alexander Serebrenik [a,*], Tom Mens [c], Mark G.J. van den Brand [a], Ekaterina Pek [b]

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

ARTICLE INFO

Article history:
Received 30 January 2013
Received in revised form 8 January 2014
Accepted 28 January 2014
Available online 31 January 2014

Keywords:
Scientometrics
Software engineering
Conferences
Empirical research

ABSTRACT

In this article we study the health of software engineering conferences by means of a suite of metrics created for this purpose. The metrics measure stability of the community, openness to new authors, introversion, representativeness of the PC with respect to the authors' community, availability of PC candidates, and scientific prestige. Using this metrics suite, we assess the health of 11 software engineering conferences over a period of more than 10 years. In general, our findings suggest that software engineering conferences are healthy, but we observe important differences between conferences with a wide scope and those with a more narrow scope. We also find that depending on the chosen health metric, some conferences perform better than others. This knowledge may 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.

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 venues with low acceptance rate [3]; the increasing importance given to conference publications by decision makers assessing scientists, both in the USA [4] and in 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]. This realisation brought a number of conferences to adopt 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 is sought between a reduced review load and high review quality.

* Corresponding author. Tel.: +31 40 2473595; fax: +31 40 2475404.
E-mail addresses: b.n.vasilescu@tue.nl (B. Vasilescu), a.serebrenik@tue.nl (A. Serebrenik), tom.mens@umons.ac.be (T. Mens), m.g.j.v.d.brand@tue.nl (M.G.J. van den Brand), pek@uni-koblenz.de (E. Pek).

0167-6423/$ – see front matter
http://dx.doi.org/10.1016/j.scico.2014.01.016

Most software engineering conferences follow a single-blind peer reviewing scheme, in which 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 introversion. 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 preferential acceptance of papers by known authors that have previously published in the same conference. We understand *introversion* as the prevalence of papers (co)authored by PC members among the accepted papers. 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]. 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 introversion, 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 criteria: community stability (author and PC turnover), openness to new authors, introversion, 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 introversion, 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 introversion (in terms of fraction of papers co-authored by PC members). Nonetheless, some conferences perform better than others according to the aforementioned criteria. In addition, we observe important differences between conferences with a wide scope and those with a more narrow scope. This knowledge can be used by conference steering committees and PC chairs, e.g., to assess composition of the PC, paper selection process and adherence to conference charters. Furthermore, prospective authors might consider conference openness as well as prestige when deciding to which conferences to submit their work.

The remainder of this article is organised as follows. Section 2 describes our research methodology, including the 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 findings. Section 4 discusses the results on a per conference basis. Section 5 surveys related work. The threats to validity, part of any empirical study, are presented in Section 6. Section 7 sketches directions for future work and Section 8 concludes.

2. Methodology

2.1. Data extraction

Numerous software engineering conferences are organised every year. Moreover, papers addressing software engineering topics are also solicited by wider-scoped computer science conferences. In our study, we focused on the conferences studied in [13]: International Conference on Software Engineering (**ICSE**), European Conference on Software Maintenance and Reengineering (**CSMR**), International Conference on Program Comprehension (**ICPC**), International Conference on Generative Programming and Component Engineering (**GPCE**), 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). To balance our sample in terms of scope, we added three more conferences with wide coverage of software engineering, namely International Conference on Automated Software Engineering (**ASE**), Symposium on the Foundations of Software Engineering (**FSE**), and International Conference on Fundamental Approaches to Software Engineering (**FASE**). Furthermore, to balance our sample in terms of age, we also included two younger conferences, namely Working Conference on Mining Software Repositories (**MSR**), and International Working Conference on Source Code Analysis and Manipulation (**SCAM**).

The data we analysed was restricted to the main research track of each conference: number of papers submitted and accepted (without distinguishing between long and short papers, if both were part of the main track), authors of the accepted papers, and composition of the programme committee. In order to facilitate replication of our study, we have published all the data and tooling developed during our work on GitHub at http://github.com/tue-mdse/conferenceMetrics. The dataset is described in more detail in [14].

For all considered conferences, most of the data of all accepted papers and their authors was extracted from the DBLP records [15]. 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 websites which were no longer available as well as announcements posted by conference organisers in Usenet newsgroups.

[1] http://archive.org/web/web.php.

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

Conference series	First ed. issued	First ed. included	Last ed. included	Included in [13]	CI	Charter availability
ASE[a]	1986	1994	2013	No	55	Own charter[b], SIGSOFT PC policy[c]
CSMR	1997	1997	2013	Yes	40	Yes, but not public; no guidelines on PC renewal[d]
FASE	1998	1998	2013	No	42	Yes[e]
FSE[f]	1993	1993	2013	No	49	SIGSOFT PC policy
GPCE[g]	2000	2000	2013	Yes	37	Yes; no guidelines on PC renewal
ICPC[h]	1992	1997	2013	Yes	41	Yes[i]
ICSE[j]	1975	1994	2013	Yes	117	SIGSOFT PC policy
ICSM[k]	1983	1994	2013	Yes	53	Yes[l]
MSR	2004	2004	2013	No	32	No[m]
SCAM	2001	2001	2013	No	15	Own charter + part of SIGSOFT PC policy
WCRE	1993	1995	2013	Yes	43	Yes, but not public; no guidelines on PC renewal

[a] Formerly Knowledge-Based Software Engineering Conference (KBSE) and Knowledge-Based Software Assistant (KBSA).
[b] http://www.ase-conferences.org/Charter.html.
[c] http://www.sigsoft.org/about/policies/pc-policy.htm.
[d] A public charter is in preparation and is expected for 2014.
[e] http://www.easst.org/fase/fasech.
[f] We do not distinguish between the years when FSE is co-located with the European Software Engineering Conference (and is known as ESEC/FSE) and the regular editions of FSE.
[g] Formerly Semantics, Applications and Implementation of Program Generation (SAIG).
[h] Formerly Workshop on Program Comprehension (WPC) and International Workshop on Program Comprehension (IWPC).
[i] http://www.program-comprehension.org/ICPC-ProgramCommittee-v1.1.pdf
[j] Formerly National Conference on Software Engineering.
[k] Formerly Conference on Software Maintenance.
[l] http://conferences.computer.org/icsm/PC-Guidelines.pdf.
[m] A charter is in preparation and is expected for MSR 2014.

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 introversion. 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 *identity merging* [16–22], and manually checked the results in a post-processing step.

2.2. Metrics

Table 2 shows all metrics we have used. Some of these coincide with those used in [13]. The basic metrics count the number of authors *#A* and PC members *#C* as well as how many papers have been submitted *#SP* to a conference for a particular year. For the *#C* metric we also considered PC chairs and General Chairs as PC members.

We quantify the review ***workload*** *RL* experienced by PC members as the ratio between the number of submitted papers and the size of the PC.

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 New Authors or programme Committee members over 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*.

By ***introversion*** we understand influence of PC membership on paper co-authorship. Systä, Harsu and Koskimies [13] call this feature "inbreeding" to reflect the negative phenomenon of favouritism of the PC towards papers co-authored by PC members. However, high share of papers co-authored by PC members among the accepted papers might stem from higher quality papers or higher publishing activity of more experienced researchers, who are also more likely to be invited to join the PC. Therefore, we use a more neutral term for this aspect of the conference health. Similarly to [13], we quantify introversion by calculating *RAC*, the ratio of accepted papers co-authored by programme committee members who served at least once in recent years.

We assess the ***representativeness of a PC for the community*** by *RCnA*, the ratio of PC members that have never co-authored a paper in several preceding editions of the same conference.

To assess the ***sustainability of the PC-candidates pool*** (its rejuvenation capacity), we measure the sustainability ratio *SR*, 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.

We define a *core author* for a given conference as 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

Table 2
Metrics used to assess conference health. If applicable, the acronym used by Systä et al. [13] is mentioned between square brackets.

Acronym	Definition
Basic metrics	
$\#A(c, y)$	number of distinct Authors for conference c in year y
$\#C(c, y)$ [*#PCmem*]	number of PC members for conference c in year y
$\#SP(c, y)$ [*#subm*]	number of Submitted Papers for conference c in year y
Workload	
$RL(c, y)$ [*revCoeff*]	Review Load for conference c in year y, i.e., $\frac{\#SP(c,y)}{\#C(c,y)}$
Stability	
$\#NA(c, y, n)$	number of New Authors for conference c in year y that were not author in years $y-n$ to $y-1$
$\#NC(c, y, n)$ [*#(real)newPCmem*]	number of New PC members for conference c in year y that were not PC member in years $y-n$ to $y-1$
$RNA(c, y, n)$	author turnover = Ratio of New Authors for conference c in year y w.r.t. years $y-n$ to $y-1$, i.e., $\frac{\#NA(c,y,n)}{\#A(c,y)}$
$RNC(c, y, n)$ [*(real)newPCprop*]	PC turnover = Ratio of New programme Committee members for conference c in year y w.r.t. years $y-n$ to $y-1$, i.e., $\frac{\#NC(c,y,n)}{\#C(c,y)}$
Openness	
$\#PNA(c, y, n)$	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 $y-1$
$RPNA(c, y, n)$	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 $y-1$, i.e., $\frac{\#PNA(c,y,n)}{RA(c,y)*\#SP(c,y)}$
Introversion	
$RAC(c, y, n)$ [*PCaccProp*]	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
Representativeness	
$RCnA(c, y, n)$	Ratio of PC members for conference c in year y that have never co-authored a paper at preceding instances of c between $y-n$ and $y-1$
Sustainability	
$SR(c, y, n)$	Sustainability Ratio = ratio between the number of core authors that have not served on the PC in years $y-n$ to y and $\#C(c, y)$
Prestige	
$RA(c, y)$ [*accRate*]	Ratio of accepted papers for conference c in year y
$CI(c)$	Conference Impact of conference c = SHINE h-index for c between 2000 and 2012

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.

Finally, we use two accepted ***prestige*** measures, the conference impact *CI* and the acceptance ratio *RA*. We compute *CI* based on the Simple H-INdex Estimator (*SHINE*) [23], a conference-specific variant of Hirsch's h-index for quantifying an individual's scientific research output [24,25]. For a given conference c, $CI(c) = SHINE(c, 2000, 2012) = 40$ means that conference c has 40 papers published between 2000 and 2012, each with at least 40 citations in the same period. 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. We have also observed high degree of agreement between *CI* and the conference rankings published by the CORE ERA ranking[2] (Computing Research and Education Association of Australasia): ICSE ($CI = 117$) is the only conference in our list ranked A^*, ASE, FSE and ICSM ($49 \leqslant CI \leqslant 59$) got the A rank, FASE (42), GPCE (37) and WCRE (43) were ranked B, and finally, CSMR (40), ICPC (41), MSR (32) and SCAM (15) were ranked C.

2.3. Data analysis

Using the R project for statistical computing [26] we visualise and statistically analyse the data to detect patterns and trends, with the aim to detect (counter-)evidence of conference health, e.g., signs of introversion or openness. The visualisation consists of two components, cf. discussion in [27]: (i) a simple graph with all the time series (conferences) overlaid, to facilitate comparisons over smaller visual spans, and (ii) small multiples for each of the time series, to facilitate assessing trends visually.

To quantify *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) [28]. Similarly, to verify presence of *linear* trends (between different metrics rather than between a metric and the time axis) we compute Pearson correlation r.

To verify claims such as "conference A tends to have higher values for metric m than conference B", we compare multiple distributions of m (one for each of the 11 conferences). Traditionally, comparison of multiple groups follows a

[2] http://core.edu.au/index.php/categories/conference%20rankings/1.

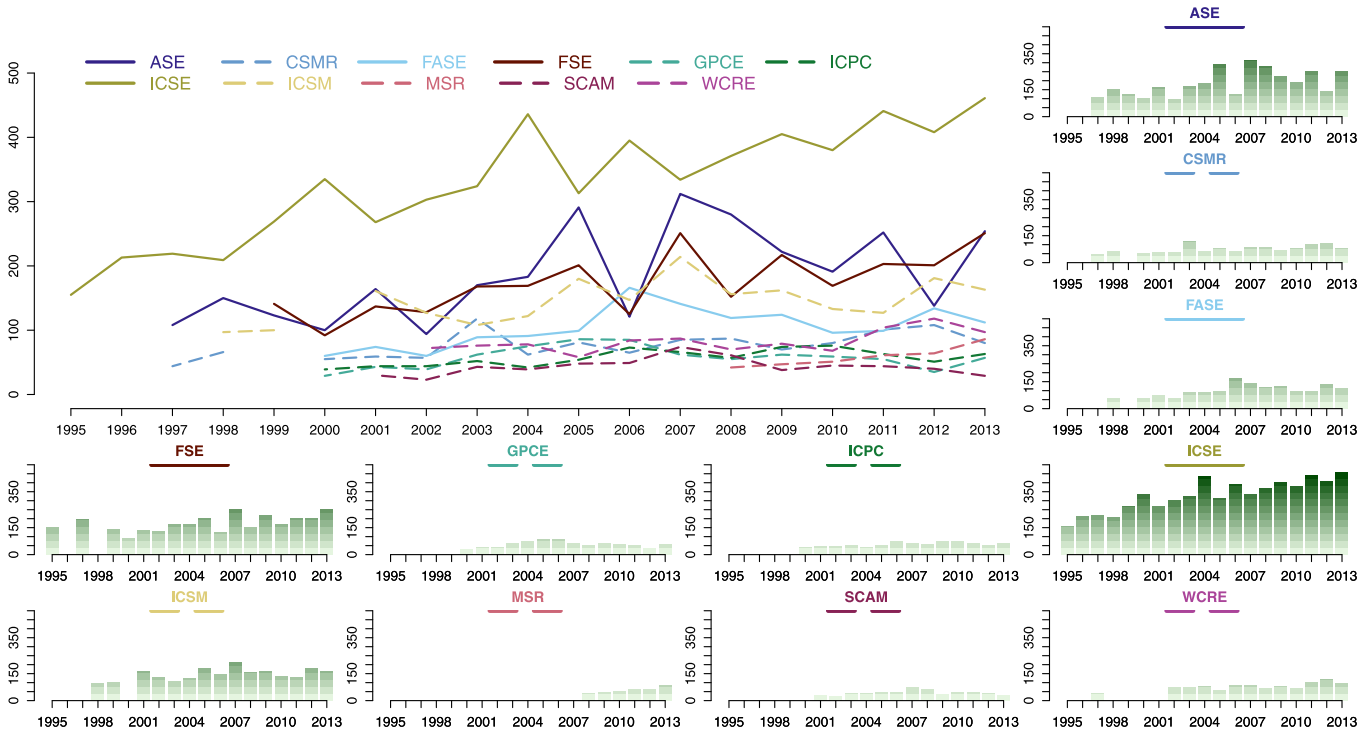

Fig. 1. Variation of the number of submitted papers per year *#SP*. "Gaps" correspond to older editions of the conferences, for which we could not retrieve the data. Wide-scoped conferences (solid lines) tend to receive more submissions than narrow-scoped ones (dashed lines): $p(\text{narrow}, \text{wide}) = 0.911^{**}$. Relations between individual conferences are visualised in the $\widetilde{\mathbf{T}}$-graph from Fig. 3.

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 by ranks [29]. The second step uses the t-test or the rank-based Wilcoxon–Mann–Whitney test [30], with Bonferroni correction [31,32]. Unfortunately, the global test null hypothesis may be rejected while none of the sub-hypotheses are rejected, or vice versa [33]. Moreover, simulation studies suggest that the Wilcoxon–Mann–Whitney test is not robust to unequal population variances, especially in the unequal sample size case [34,35]. 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 $\widetilde{\mathbf{T}}$ [36] using the traditional 5% family-wise error rate. $\widetilde{\mathbf{T}}$ is robust against unequal population variances. A more comprehensive discussion of $\widetilde{\mathbf{T}}$ goes beyond the scope of this article and can be found in [36]. Small examples detailing the application of $\widetilde{\mathbf{T}}$ can be found in [21,36] and additional scientific applications of the technique in [37–39].

For ease of presentation (given 11 conferences, one has to report the results of $\frac{11\times 10}{2} = 55$ comparisons per metric), we use the $\widetilde{\mathbf{T}}$-graphs proposed in [21] 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 $\widetilde{\mathbf{T}}$ (as opposed to, e.g., the traditional pairwise Wilcoxon–Mann–Whitney tests [40]), 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., all wide-scoped conferences versus all 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 [41,42] is not robust against differences in variance [34,35], and the $\widetilde{\mathbf{T}}$ procedure as described above cannot be used to compare two distributions [36]. We therefore use the two-distributions equivalent of the $\widetilde{\mathbf{T}}$ procedure, i.e., we perform two sample tests for the non-parametric Behrens–Fisher problem [35], and compute confidence intervals for the *relative effect* of the two samples (using the R package `nparcomp` [43]). If the relative effect of samples A and B, which is traditionally denoted $p(A, B)$ [35], exceeds 0.5 then B tends to be larger than A. Therefore, we accompany visualisations of the evolution of different metrics by $\widetilde{\mathbf{T}}$-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, and to avoid the possible confusion between the relative effect p and p-values, we superscript the test result using the following convention: no superscript corresponds to p-value $\geqslant 0.05$, * corresponds to $0.01 \leqslant$ p-value < 0.05, and ** corresponds to p-value < 0.01.

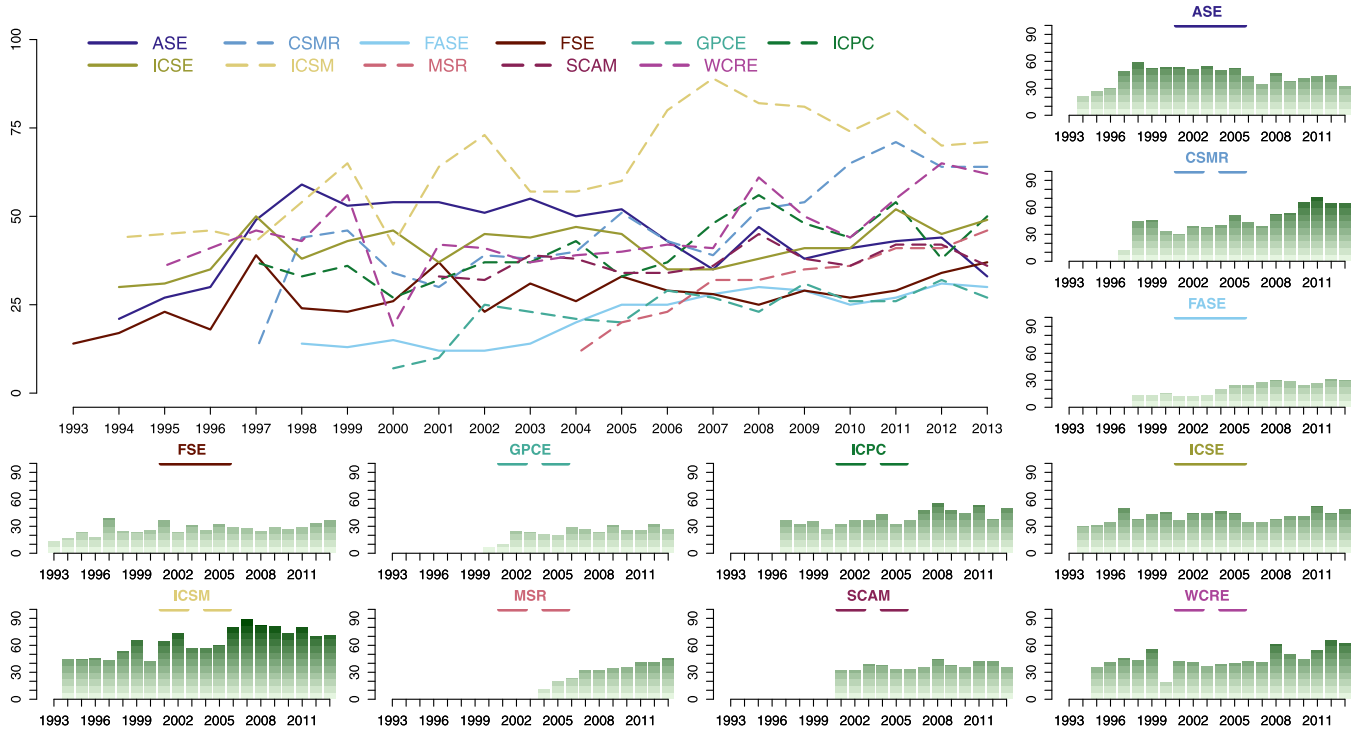

Fig. 2. Variation of the number of programme committee members $\#C(c, y)$ per year. Wide-scoped conferences (solid lines) have smaller PCs than narrow-scoped ones (dashed lines): $p(\text{narrow}, \text{wide}) = 0.352^{**}$. Relations between individual conferences are visualised in the $\widetilde{\mathbf{T}}$-graph from Fig. 4.

3. Results

3.1. Workload

We start our analysis of conference health 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?

Fig. 1 displays the variation of the number of submitted papers per year $\#SP(c, y)$, for each of the conferences. The most popular conference (i.e., the one receiving the most submissions) is ICSE, followed by ASE, FSE and ICSM (Fig. 3). It is interesting to note that the narrow-scoped ICSM received over the years comparable (in terms of the $\widetilde{\mathbf{T}}$ procedure) numbers of submissions as the wider-scoped FSE, ASE and FASE. Increasing trends are confirmed for MSR ($\rho = 1^{**}$), ICSE ($\rho = 0.89^{**}$) and FASE ($\rho = 0.73^{**}$): these conferences tend to receive more submissions each year, since their first editions. The other conferences exhibit less clear (increasing or decreasing) trends. Overall, wide-scoped conferences (solid lines) tend to receive more submissions than narrow-scoped ones (dashed lines): $p(\text{narrow}, \text{wide}) = 0.911^{**}$.

With the exception of ICSE, conferences receiving increasingly more submissions resort to increasing the size of their programme committees. Inspection of Fig. 2 reveals increasing trends for $\#C$ for MSR ($\rho = 0.99^{**}$), FASE ($\rho = 0.87^{**}$), CSMR ($\rho = 0.79^{**}$), ICSM ($\rho = 0.77^{**}$), ICPC ($\rho = 0.76^{**}$) and GPCE ($\rho = 0.75^{**}$): these tend to increase their PC size over the years. We notice that this includes 5 out of 7 of the narrow-scoped conferences. The other conferences exhibit less clear trends. Out of all the conferences in our sample, ICSM consistently has the largest PC (Fig. 4).

Overall, when comparing wide-scoped conferences to narrow-scoped ones, we observe that the former receive more submissions but have smaller PCs, hence higher review load than the latter. Of course, some PC members engage external reviewers, implying that the actual review load might be lower than the ratio *RL* of the number of submissions and the number of PC members. Moreover, the actual review load might be higher than *RL* if some of the PC members, e.g., the PC chairs, do not review submissions at all or review less submissions than other PC members.

Fig. 6 displays the variation of the review load *RL*. It is interesting to observe the increasing trend for ICSE ($\rho = 0.80^{**}$), resulting from increasingly more submissions each year and a PC size that does not tend to grow accordingly. ICSE and to a lesser extent FSE are also the conferences with the highest values of *RL* (Fig. 5): since 2007, ICSE has stabilised to a ratio of 9 submissions per PC member (if each submission is reviewed on average by 3 PC members, this implies approximately 27 submissions being assigned to each PC member), followed by FSE with a ratio of 7 submissions per PC member (i.e. 21 submissions assigned to each PC member). Other trends are visible for MSR (increasing, $\rho = 1^{**}$), ASE (increasing, $\rho = 0.78^{**}$), and GPCE (decreasing, $\rho = -0.65^{*}$).

Recognition 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 for its 2014 edition. 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

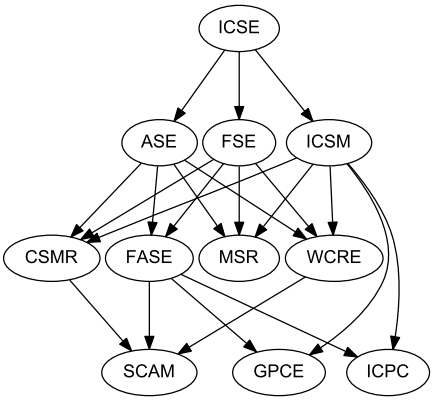

Fig. 3. $\widetilde{\mathbf{T}}$-graph for the number of submitted papers *#SP*.

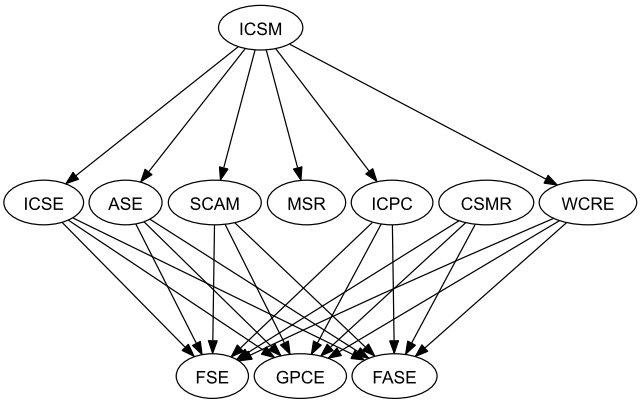

Fig. 4. $\widetilde{\mathbf{T}}$-graph for the number of PC members *#C*.

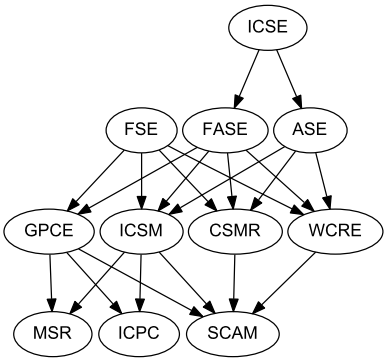

Fig. 5. $\widetilde{\mathbf{T}}$-graph for the review load *#RL*.

model has recently been used by the International Requirements Engineering Conference (RE) as well as the International Conference on Model-Driven Engineering Languages and System (MoDELS).

> Wide-scoped conferences tend to receive more submissions, but typically have smaller PCs than narrow-scoped ones, resulting in higher review load. A potential approach to lowering review loads is the Program Board model used recently by conferences such as RE and MoDELS, and adopted by ICSE for its 2014 edition.

3.2. Stability

To study stability of the conference community we analyse the PC turnover and author turnover for the considered conferences. We also assess the influence of the availability of a PC charter on the PC turnover. Indeed, since conferences are frequently subject to charters or guidelines that require PC renewal, they must strike a balance between PC turnover and continuity.

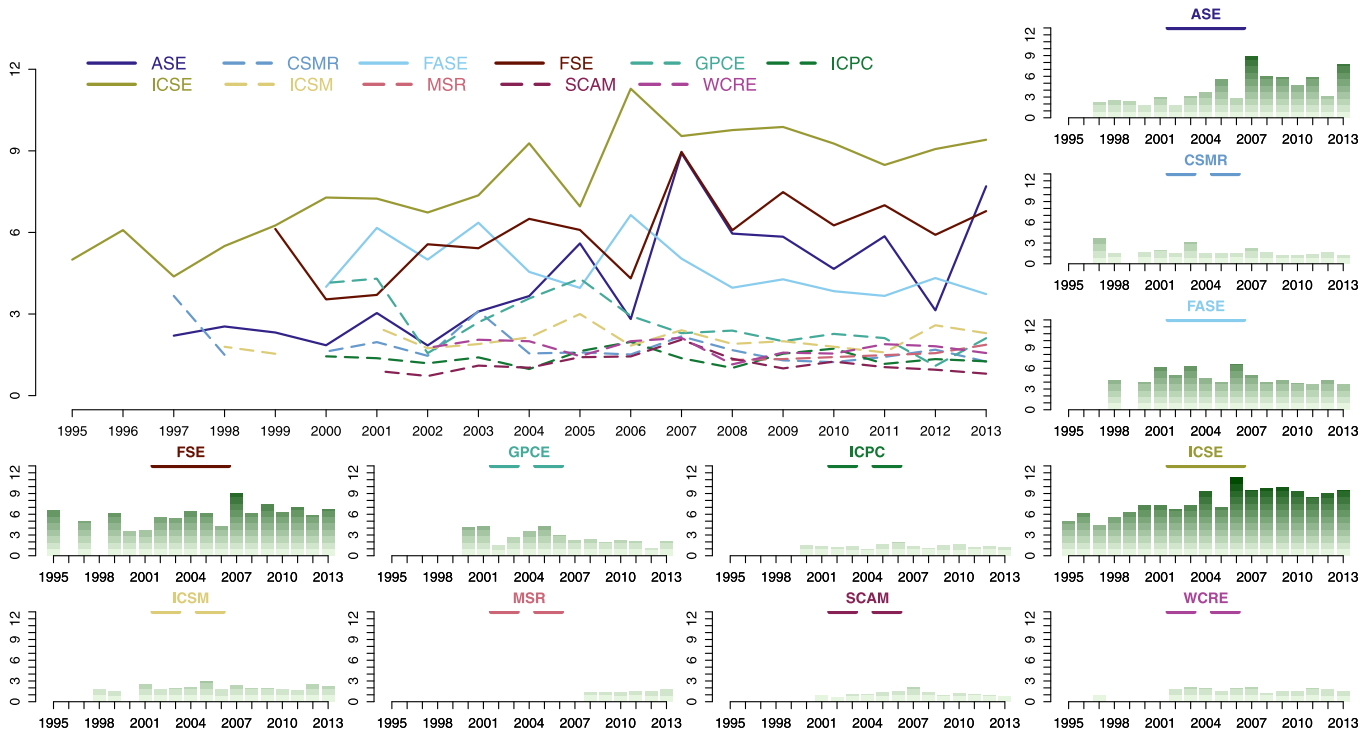

Fig. 6. Variation of the review load *RL* per year. Wide-scoped conferences (solid lines) have higher review load than narrow-scoped ones (dashed lines): $p(\text{narrow}, \text{wide}) = 0.974^{**}$. Relations between individual conferences are visualised in the $\widetilde{\mathbf{T}}$-graph from Fig. 5.

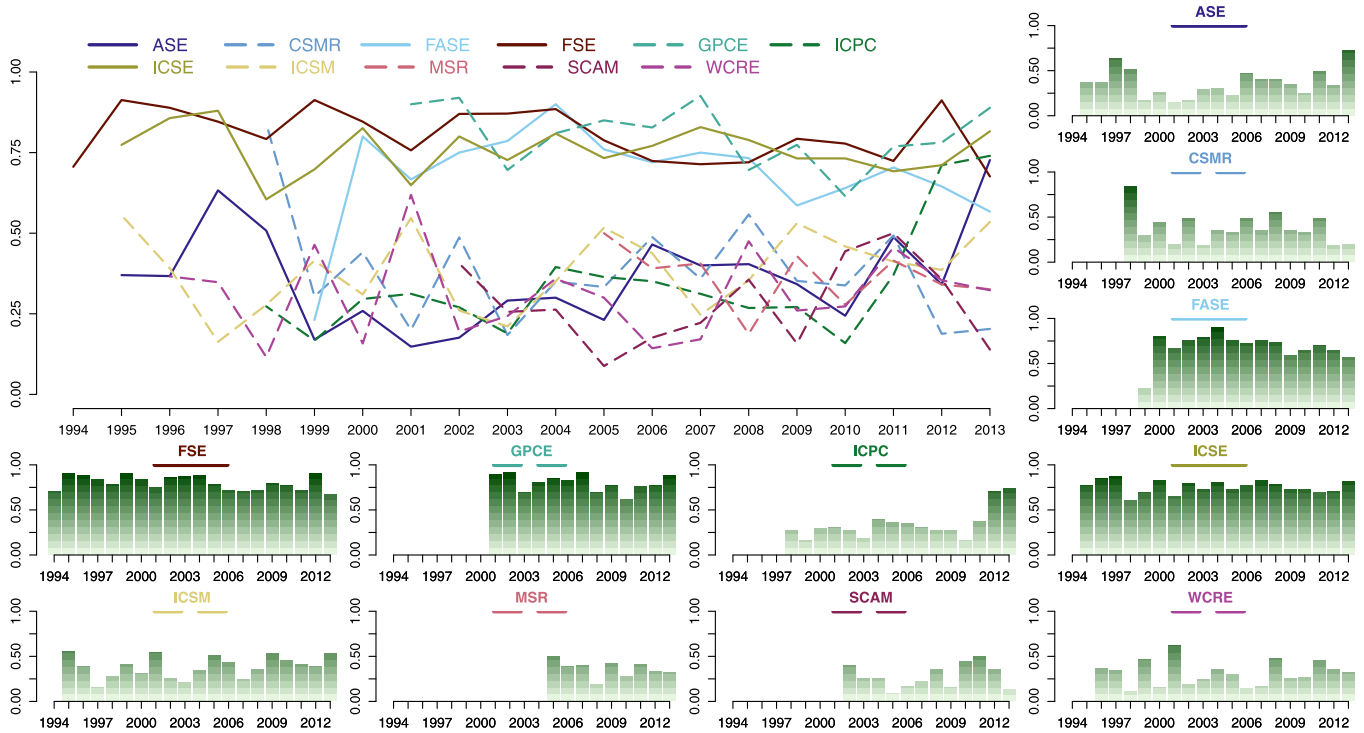

Fig. 7. Variation of PC turnover w.r.t. previous year $RNC(c, y, 1)$. Wide-scoped conferences (solid lines) usually have higher values than narrow-scoped ones (dashed lines): $p(\text{narrow}, \text{wide}) = 0.776^{**}$. The $\widetilde{\mathbf{T}}$-graph in Fig. 8 confirms this, except for GPCE (which behaves like a wide-scoped conference) and ASE (which behaves like a narrow-scoped conference).

3.2.1. PC turnover

Inviting PC members from previous editions helps to ensure continuity and coherence. When studying $RNC(c, y, 1)$, the PC turnover rate w.r.t. the previous year, we observe a wide variation (Fig. 7): the lowest observed PC renewal ratio is 8.8% for SCAM 2005, and the highest is 93% for GPCE 2007. Using Spearman rank correlation, no clear monotonic trends could be inferred for $RNC(c, y, 1)$ for any of the conferences. To check whether the chosen width of the sliding window (one

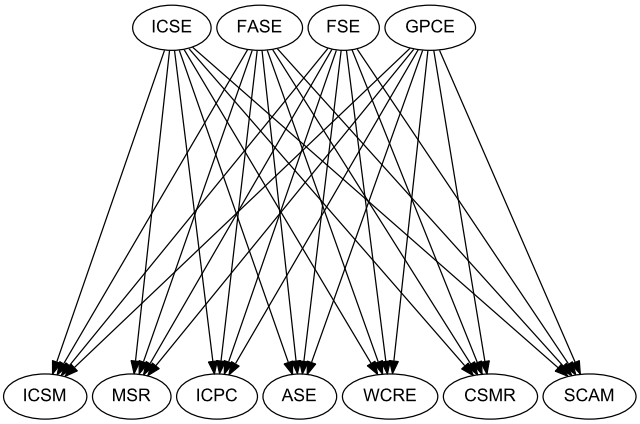

Fig. 8. $\widetilde{\mathbf{T}}$-graph for the PC turnover w.r.t. previous year $RNC(c, y, 1)$.

year) affects our results, we repeated the analysis by looking at a longer period of 4 previous years. The PC turnover rate $RNC(c, y, 4)$ shows similar behaviour: no clear trends, and a big range of values for the metric (from 8.8% for SCAM 2005 to 85% for GPCE 2007).

Confidence intervals for the relative effect reveal that wide-scoped conferences usually have higher values than narrow-scoped ones: $p(\text{narrow}, \text{wide}) = 0.776^{**}$. One can conjecture that PC turnover might be correlated with the number of potential PC members, i.e., senior researchers active in research areas covered by the conference. The first step towards estimating the number of potential PC members, and, hence, sustainability of the conference, is the Sustainability Ratio $SR(c, y, n)$ discussed in Section 3.6.

We use $\widetilde{\mathbf{T}}$-graphs to compare distributions of RNC-metrics for different conferences. Two groups of conferences become apparent in Fig. 8: the wide-scoped ICSE, FSE and FASE and the narrow-scoped GPCE, have consistently higher values of $RNC(c, y, 1)$ than the other considered conferences.

We conjecture that presence of GPCE in the high-turnover group of conferences might be due either to relevance of the GPCE topics to a broader scientific community, or failure to establish a core community. One could also argue that the small PC size of GPCE plays a role: 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 $\widetilde{\mathbf{T}}$ to the values of $\#C$, visualised in Fig. 4). A more detailed investigation of the reasons for high PC turnover in GPCE goes beyond the scope of this article.

> Wide-scoped conferences ICSE, FSE and FASE, and narrow-scoped GPCE have consistently higher PC turnover than the other conferences. ASE (wide-scoped) appears to be an outlier with respect to the other wide-scoped conferences.

3.2.2. 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, ASE and SCAM, 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. Similarly, SCAM adheres to this policy in only 5 out of 12 editions considered, with the most recent noncompliance being in 2013.

> ICSE and FSE always conform to the “at least one third” PC renewal policy, while ASE and SCAM do 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.

Finally, the PC charter of ICSM, applied since 2004, requires between 10% and 30% of the members to be new with respect to the preceding year’s PC. Surprisingly, in nine years following the application of the charter (2004–2012) only ICSM 2007 adhered to this renewal policy (24,7%). All other ICSM editions in this period *exceed* the required renewal percentage reaching 53% in 2007.

ICSM almost always exceeds the percentage of new PC members prescribed by the charter.

3.2.3. Author turnover

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 introversion.

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 98% for GPCE 2000 or SCAM 2012, 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.

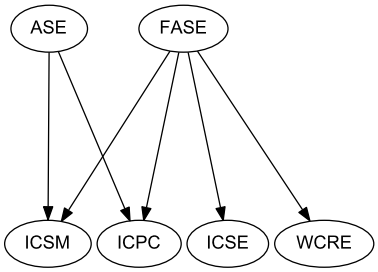

Fig. 9. $\widetilde{\mathbf{T}}$-graph for the author turnover w.r.t. four previous years $RNA(c, y, 4)$. Statistically indistinguishable conferences are not displayed.

Fig. 10 visualises $RNA(c, y, 4)$ that takes 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 $\widetilde{\mathbf{T}}$-graph of Fig. 9 further reveals differences between the wide-scoped ASE and FASE, and the narrow-scoped ICPC, ICSM and WCRE. Considered together, the wide-scoped conferences tend to have higher author turnover ($p(\text{narrow}, \text{wide}) = 0.684^{**}$). 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 [44].

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.

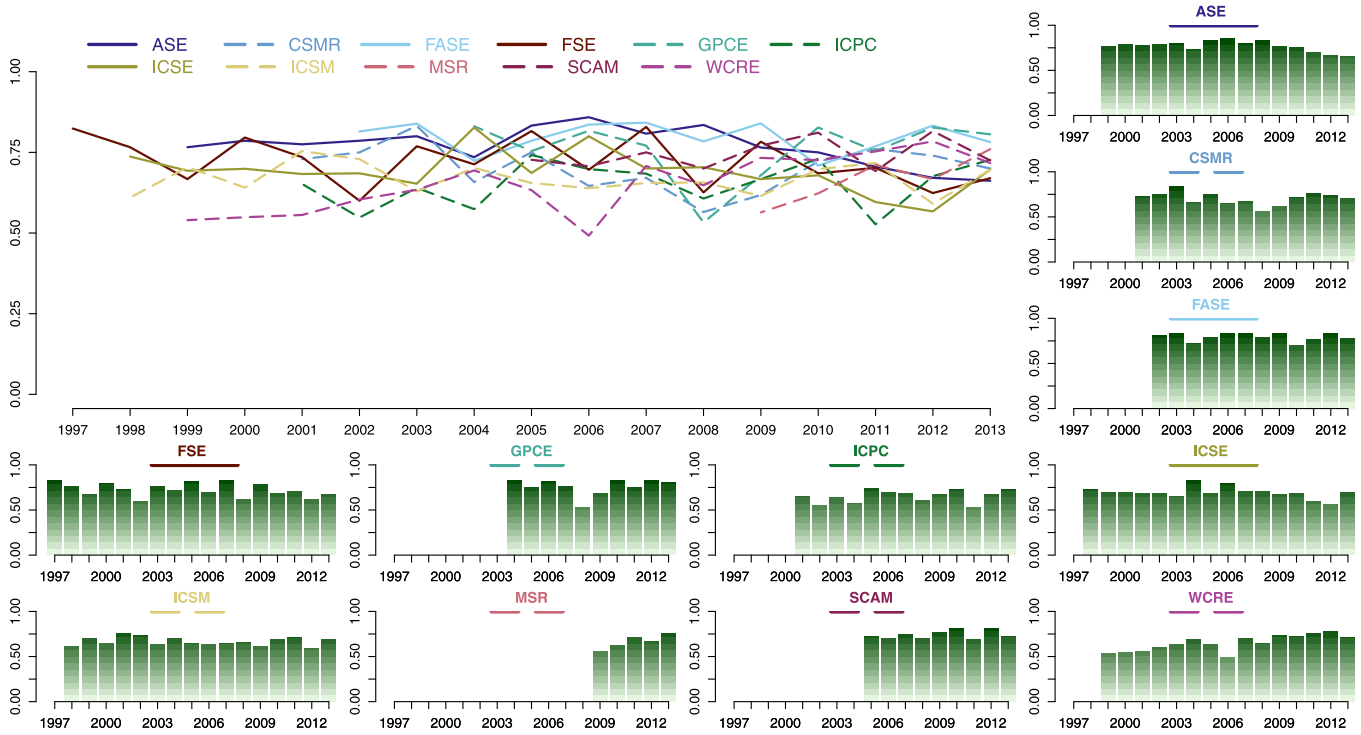

Fig. 10. Variation of the author turnover w.r.t. four previous years $RNA(c, y, 4)$. Wide-scoped conferences (solid lines) tend to have higher values than narrow-scoped ones (dashed lines): $p(\text{narrow}, \text{wide}) = 0.684^{**}$. Relations between individual conferences are visualised in the $\widetilde{\mathbf{T}}$-graph from Fig. 9.

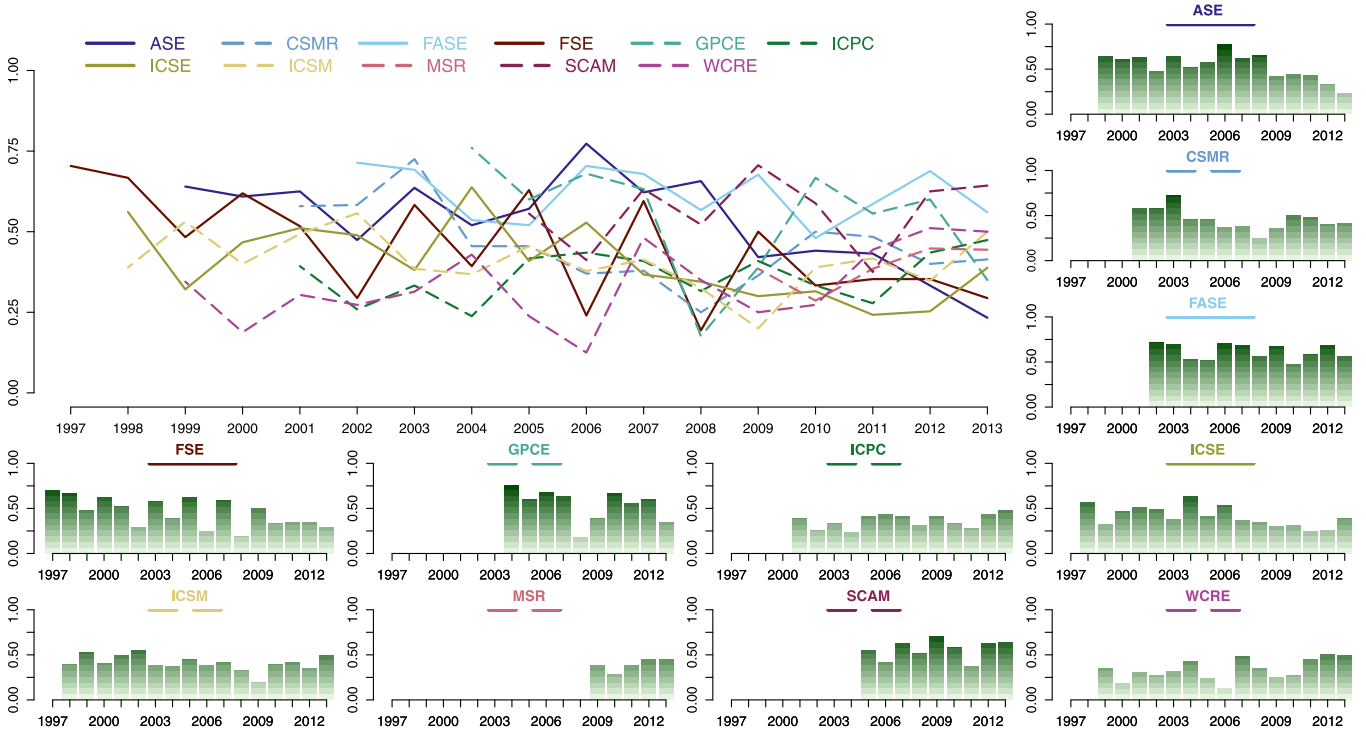

Fig. 11. Variation of $RPNA(c, y, 4)$—the fraction of papers for which none of the co-authors has previously published at conference c in the 4 preceding years to y. Wide-scoped conferences (solid lines) usually have higher values (are more open) than the narrow-scoped ones (dashed lines): $p(\text{narrow}, \text{wide}) = 0.627^{**}$. Relations between individual conferences are visualised in Fig. 13.

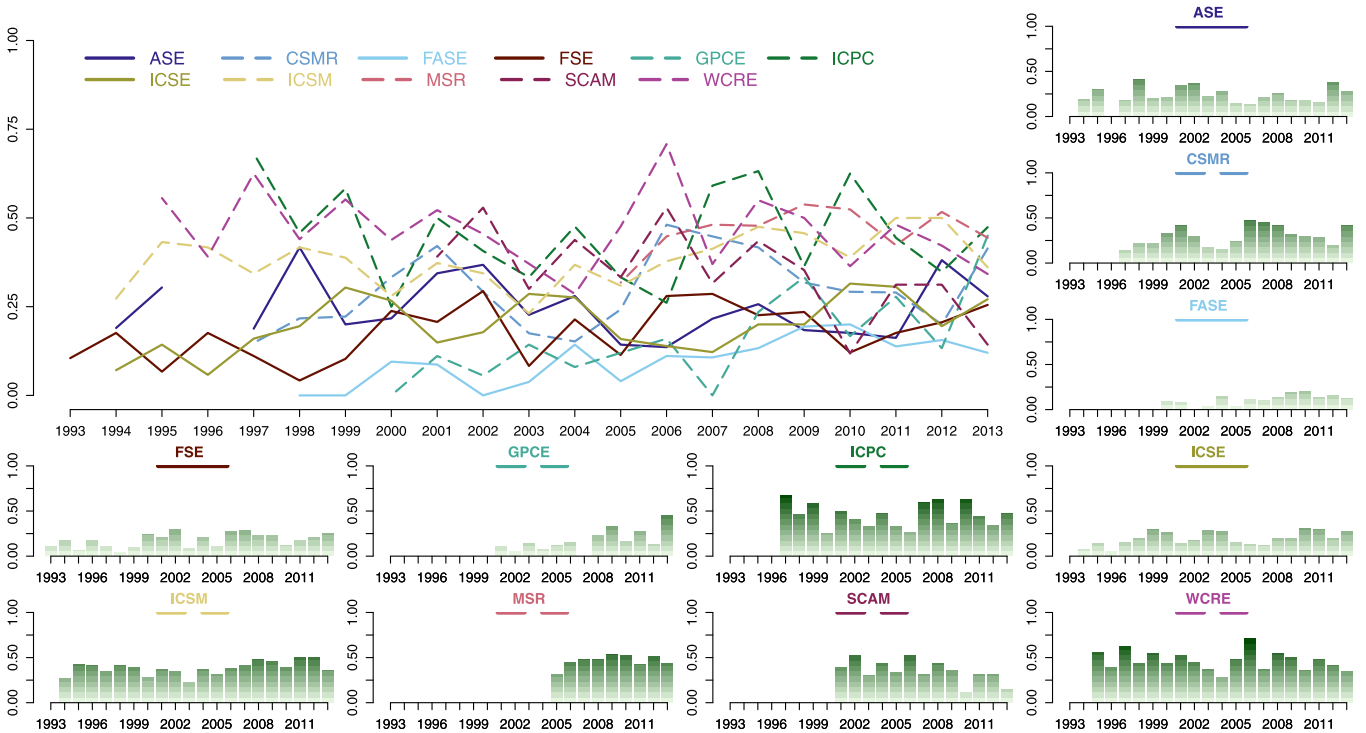

Fig. 12. Variation of $RAC(c, y, 0)$—the fraction of papers co-authored by PC members in the same year. Wide-scoped conferences (solid lines) tend to have lower values (are less introvert) than the narrow-scoped ones (dashed lines): $p(\text{narrow}, \text{wide}) = 0.144^{**}$. Relations between individual conferences are visualised in Fig. 14.

We emphasize that attracting new authors and renewing the PC do not necessarily prevent introversion in wide-scoped conferences (e.g., the new authors could also be PC members). We therefore further investigate openness and introversion 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 published at conference c in the 4 preceding years (Fig. 11). 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 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.

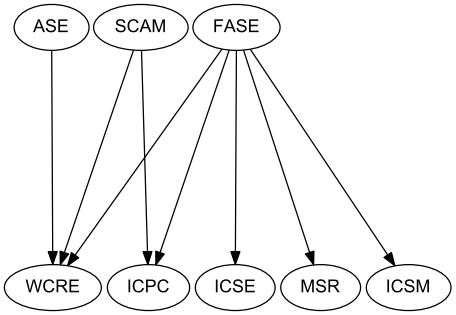

Fig. 13. $\widetilde{\mathbf{T}}$-graph for the ratio of papers by new authors w.r.t. four previous years $RPNA(c, y, 4)$. Statistically indistinguishable conferences are not displayed.

At the high end of the scale (more open communities), the $\widetilde{\mathbf{T}}$-graph of Fig. 13 reveals ASE, FASE and SCAM: ASE and SCAM tend to be more open than ICPC and WCRE. Similarly, FASE tends to be more open than ICPC, WCRE, ICSM, MSR 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 trends are exhibited by ICSE and FSE, and in recent years ASE: over the years, the percentage of papers for which none of the co-authors has ever published at these conferences in the preceding four editions is decreasing (ASE: $\rho = -0.629^{*}$; FSE: $\rho = -0.613^{**}$; ICSE: $\rho = -0.612^{*}$). In other words, ICSE, FSE and ASE are becoming decreasingly open.

> ASE, FASE and SCAM are among the most open communities (have a low entrance barrier). ICSE, FSE and ASE are becoming increasingly less open over the years.

3.4. Introversion

To evaluate introversion we study the evolution of $RAC(c, y, 0)$, the fraction of papers co-authored by PC members in the same year (Fig. 12). When studying introversion it is essential to consider the PC chairs and the General chairs as PC members because they are still influential enough within the community. Again, the $\widetilde{\mathbf{T}}$-graph reveals differences between the wide-scoped ICSE, FSE, FASE, and ASE, and the narrow-scoped ICSM, CSMR, WCRE, ICPC, MSR, SCAM and GPCE when considered as two groups: the wide-scoped conferences tend to be less introvert ($p(\text{narrow}, \text{wide}) = 0.144^{**}$).

The values range between 0% (no introversion at all) for FASE 2002 or GPCE 2007 and 71% for WCRE 2006 (high introversion). Overall, WCRE, ICPC, MSR, SCAM and ICSM tend to be the most introvert (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 introvert. However, both FASE ($\rho = 0.79^{**}$) and GPCE ($\rho = 0.73^{**}$) are becoming increasingly introvert. Repeating the analysis with a longer time window using $RAC(c, y, 4)$ concurs with the previous ranking.

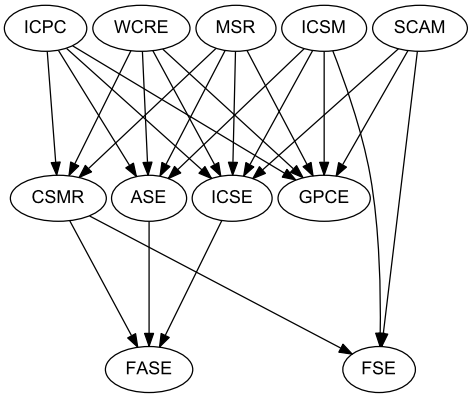

Fig. 14. $\widetilde{\mathbf{T}}$-graph for the ratio of papers co-authored by PC members in the same year $RAC(c, y, 0)$.

> WCRE (on average 47% of the papers accepted each year are co-authored by PC members), MSR (46%), ICPC (46%), ICSM (38%) and SCAM (35%) tend to be the most introvert conferences. In contrast, FASE (10%), GPCE (14%) and FSE (18%) tend to be the least introvert. Overall, wide-scoped conferences tend to be less introvert than narrow-scoped ones.

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 (Fig. 15): $RAC(c, y, 0)$ and $RNC(c, y, 1)$ show a moderately-strong negative linear correlation ($r = -0.58^{**}$). The moderate correlation is applicable more to the narrow-scoped ($r = -0.54^{**}$) rather than the wide-scoped ($r = -0.45^{**}$) conferences. At the level of individual conference series, none of the conferences confirms this trend. This finding means that inherent features of conferences (rather than differences between individual editions of the same conference) influence the association between a lower PC turnover and a greater proportion of PC papers among the accepted papers.

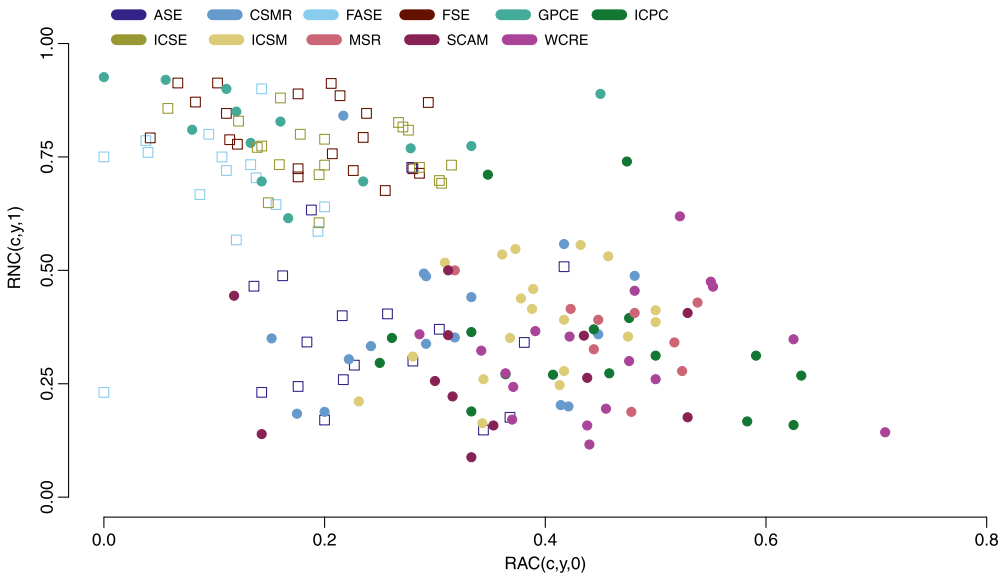

Fig. 15. For narrow-scoped (filled circles) rather than wide-scoped (empty squares) conferences, higher PC turnover is associated with smaller fractions of PC papers among the accepted papers.

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

3.5. Representativeness

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. *A fortiori* this should also hold for the PC chairs and General chairs.

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 often. 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.

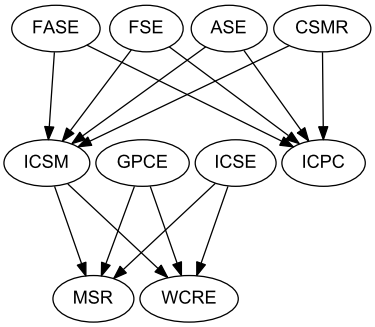

Fig. 16. $\widetilde{\mathbf{T}}$-graph for $RCnA(c, y, 4)$.

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 (Fig. 17): 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 $\widetilde{\mathbf{T}}$-graph of Fig. 16 reveals that values of $RCnA(c, y, 4)$ for WCRE are lower than for all other conferences except ICPC, MSR and SCAM; ICSM and ICPC are both lower than any of CSMR, FSE, ASE and FASE. When viewed together, narrow-scoped conferences have more representative PCs than wide-scoped ones ($p(\text{narrow}, \text{wide}) = 0.801^{**}$).

> 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, MSR and SCAM. ICSM and ICPC also have representative PCs.

MSR ($\rho = 0.9^{*}$) and ICPC ($\rho = 0.731^{**}$) exhibit the clearest increasing trends, suggesting that their PCs are becoming increasingly less representative of their respective communities. However, one should also take into account the fact that the time series for MSR and ICPC started from values that were much lower than for the other conferences, and that the highest values (2013 for MSR and 2012 for ICPC) still fall within the same range as for the other conferences.

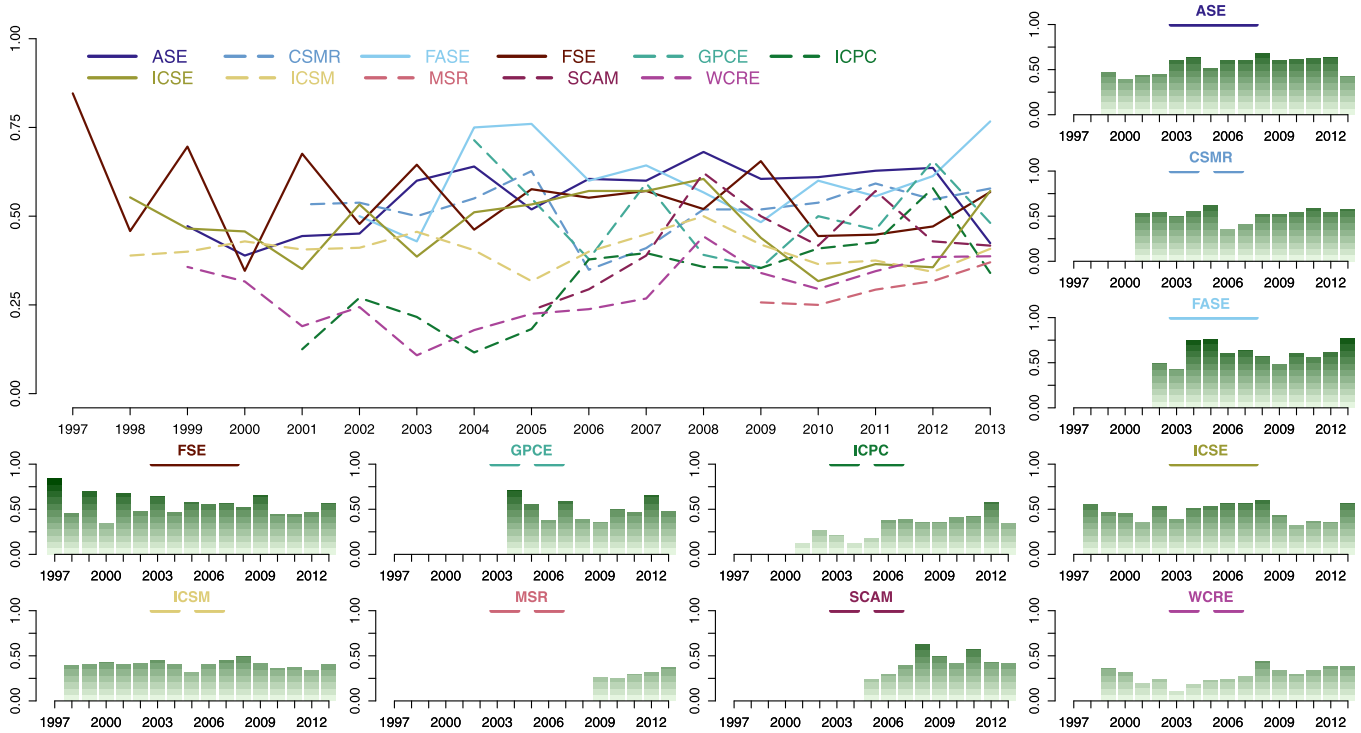

Fig. 17. Variation of $RCnA(c, y, 4)$—the ratio of PC members for conference c in year y that have never co-authored a paper at any of the four preceding instances of c. The PCs of narrow-scoped conferences (dashed lines) are more representative of their communities than those of wide-scoped conferences (solid lines): $p(\text{narrow}, \text{wide}) = 0.801^{**}$. Relations between individual conferences are visualised in Fig. 16.

> 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. Sustainability

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 the 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.

ICSE and FSE stand out as the conferences with the most sustainable pools of PC candidates in recent years (Figs. 19 and 18): since 2010 there are on average 1.6 potential replacements (core authors that did not serve on the PC of any of the preceding four editions) for each PC member for ICSE, and 1.2 for FSE. Using the Spearman correlation we also confirm (Fig. 18) increasing trends for ASE ($\rho = 0.907^{**}$), FSE ($\rho = 0.832^{**}$) and ICSE ($\rho = 0.754^{**}$)—they have increasingly more sustainable pools of PC candidates; in contrast, GPCE ($\rho = -0.929^{**}$), CSMR ($\rho = -0.773^{*}$) and WCRE ($\rho = -0.721^{**}$) exhibit decreasing trends—finding qualified PC candidates is becoming more challenging. When viewed together, wide-scoped conferences have more sustainable pools of PC candidates than narrow-scoped ones ($p(\text{narrow}, \text{wide}) = 0.804^{**}$).

> 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, ICPC 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 [45]. 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 and is computed for the 2000–2012 interval, we contrast it against the mean number of submissions during the same period. Computing the mean is meaningful since for

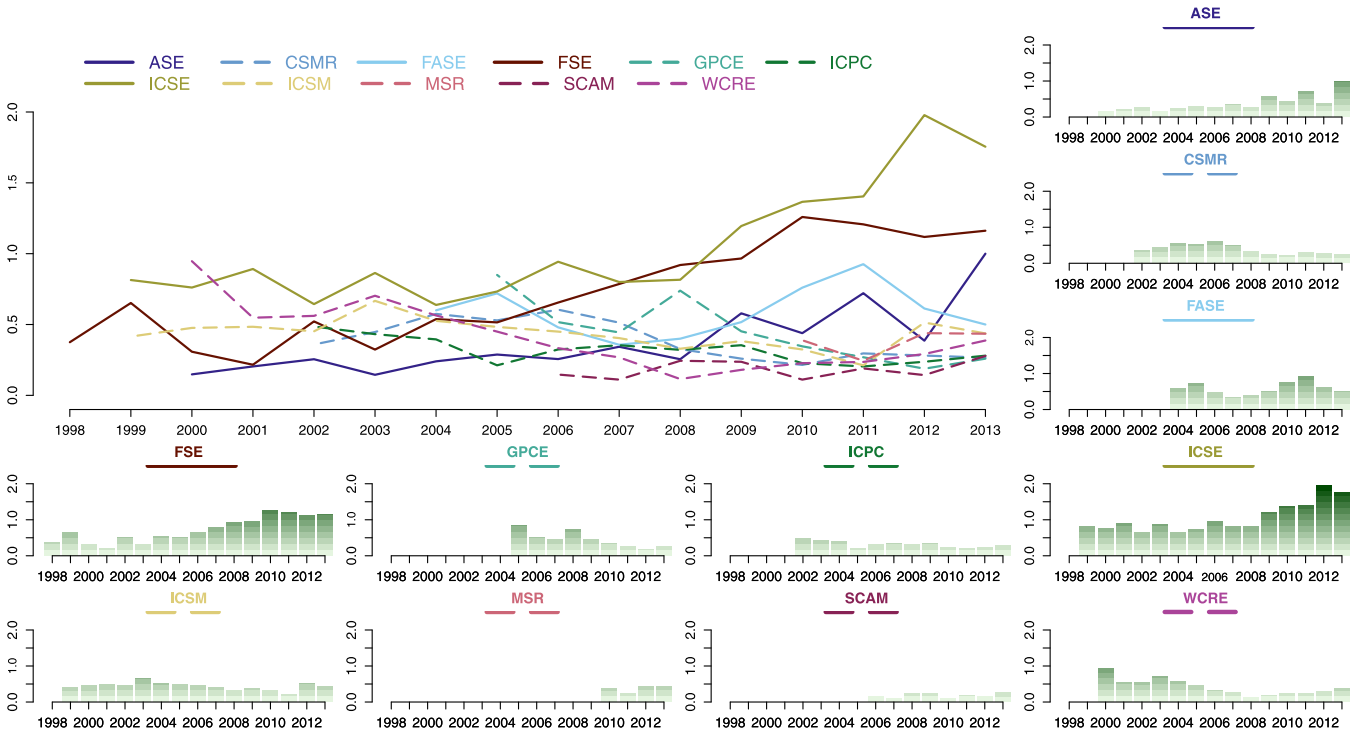

Fig. 18. Variation of $SR(c, y)$—the average number of *unsung heroes* per PC member. Wide-scoped conferences (solid lines) have more sustainable pools of PC candidates than narrow-scoped ones (dashed lines): $p(\text{narrow}, \text{wide}) = 0.804^{**}$. Relations between individual conferences are visualised in Fig. 19.

each conference series the distribution of $\#SP(c, y)$ over the years is close to normal (the Shapiro–Wilkinson test fails to reject the normality hypothesis at 95% confidence level). We observe very strong positive and statistically significant linear correlation ($r = 0.95^{**}$), confirming that more prestigious conferences attract more submissions.

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

The acceptance ratio has been related to conference prestige, e.g., by Manolopoulos [46], and further debated by Laplante et al. [47]. It is also generally believed that the acceptance rate of a conference is inversely proportional to its scientific impact [45]. To verify this intuition, we study the relation between the conference impact factor $CI(c)$ and the mean acceptance rate $RA(c, y)$ for the 2000–2012 period, following the same reasoning as above. We observe strong negative linear correlation ($r = -0.77$), 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, and with a study of a database conference ADBIS by Manolopoulos [46].

> Higher impact conferences tend to have lower acceptance rates.

Taken together our conclusions further concur with the rules "reverse engineered" by Küngas et al. [48] from the CORE ERA ranking: lower acceptance rate (< 0.363) *in combination with* higher number of citations (< 706) corresponds to A-ranked conferences (ICSE, ASE, FSE and ICSM in our list), while higher acceptance rate or lower acceptance rate in combination with lower number of citations correspond to B-rated conferences.

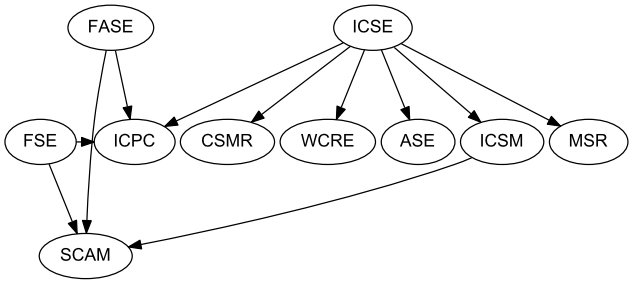

Fig. 19. $\widetilde{\mathbf{T}}$-graph for $SR(c, y, 4)$.

4. 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.

We stress that our analysis does not suggest "optimal values" for conference health metrics. Based on our experience with software metrics, as well as with different software engineering conferences, we do not believe in "one size fits all" thresholds. Consensus on this kind of "optimal values" is hard to achieve, and once achieved it will miss the specific context of each individual conference. Moreover, the following discussion shows that a relative comparison of different conferences is still possible even in absence of "optimal values".

CSMR has a low author turnover. Together with average openness, this suggests a relatively stable community. Moreover, the PC turnover is low, the introversion 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 is decreasing, 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 there is also high introversion. 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 introversion and closedness, these conferences may consider changing the traditional 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 [49]. 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 introversion, 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, the failure to establish a relatively stable and sufficiently large core group ready to serve on the PC. 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 have a mentoring program for new authors and, moreover, decided to adopt a Program Board model as of 2014.

Similarly to ICSE, FSE has high author turnover, high PC turnover and low introversion. 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 have stabilised at 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 introversion 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). Therefore, similarly to CSMR we suggest inviting some of the "unsung heroes" of ASE to join the PC. Unlike CSMR, the sustainability of the PC candidates pool is higher for ASE.

MSR seems to be a conference at cross-roads: it has a relatively low but increasing workload, caused by the increasing number of submissions and not adequately balanced by the increasing size of the PC. Its PC is representative of the author community but is becoming less so due to a relatively low PC turnover and increasing author turnover. Finally, it does not yet have a charter, but the charter is in preparation and is expected to be presented at MSR 2014. While high author turnover and workload are more typical for wide-scoped conferences, low PC turnover is more common for narrow-scoped conferences. Sustainability of the conference is comparable to the more sustainable narrow-scoped conferences such as ICSM.

SCAM is an introvert conference, highly open to new authors and exhibiting high author turnover. Sustainability of the PC candidates pool is low, i.e., finding qualified PC candidates is challenging. This suggests that while SCAM succeeds in attracting new authors, it does not succeed in retaining those authors for a number of subsequent editions. SCAM has the lowest Conference Impact $CI = 15$ among the conferences considered, which might be the reason why only few authors would explicitly target SCAM on a regular basis.

4.1. Towards MCDM

The above discussion of the relative strengths and weaknesses of each conference based on the considered health metrics can be seen as a first step of a multiple-criteria decision making/analysis process (MCDM or MCDA) [50,51], a well-studied area of operations research. Based on our data, a prospective conference author with no preceding experience in this particular conference might apply MCDM techniques to help decide whether or not to submit a paper using the openness,

introversion and prestige metrics as decision criteria. Similarly, a conference steering committee interested in the health assessment of their conference relatively to other software engineering conferences, can consider health assessment as a hierarchical MCDM problem [51, p. 2] with workload, stability, openness, introversion, representativeness, sustainability and prestige as criteria (cf. Table 2), and individual metrics as the corresponding subcriteria.

5. Related work

In this article, we built upon the work by Systä, Harsu and Koskimies [13], replicating and extending their introversion study of 6 software engineering conferences (ICSE, ICSM, ICPC, CSMR, WCRE, and GPCE) observed during the 2004–2009 period.[3] Our study takes into account more conferences, uses a wider range of metrics, considers longer time periods, and uses $\widetilde{\mathbf{T}}$-graphs to perform pairwise comparison of metrics across conferences. We assess the health of software engineering conferences with respect to several criteria (community stability, openness to new authors, introversion, 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.

Similar in spirit, although not focusing on software engineering conferences,[4] Biryukov and Dong [52] 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 3.2.3) 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 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 [53], Zhuang et al. [54] and Sakr and Alomari [55], who recognise the impact of PC quality on the conference quality. Elmacioglu and Lee [53] extract information about PC composition for a number of conferences from Calls for Papers published on DBWorld,[5] and construct a collaboration graph for authors of these conferences from the ACM Guide.[6] By dividing the set of analysed conferences into two groups (reputable and questionable[7]), 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 representativeness, 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 labelled as "reputable". However, our results using the $\widetilde{\mathbf{T}}$-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 [55] 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.[8] Nonetheless, both their analysis and ours suggest that from the PC composition viewpoint, the considered conferences are relatively healthy.

Inbreeding, related to introversion, has been studied by Inanc and Tuncer [56], albeit with a different meaning. While we consider 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 [23]) 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., [45,48,57–63]), but they fall beyond the scope of this paper. For example, da Silva et al. [61] propose a ranking scheme for scientific conferences based on ranking the PC

[3] The GPCE data is collected over the 2002–2009 period.
[4] The only conference in common with our study is FSE.
[5] http://research.cs.wisc.edu/dbworld/.
[6] Currently the ACM Digital Library.
[7] Distinction based on personal experience of Elmacioglu and Lee.
[8] 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.

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 [24] at least x) as well as the inequality (spread) of the h-indices of the PC members, computed using the Gini index [64].

6. Threats to validity

As any empirical study our work is subject to a number threats to validity. We distinguish between three categories of threats: construct validity, external validity and internal validity. Construct validity is related to the question whether the measures proposed, i.e., the metrics in Table 2, constitute an adequate operationalisation of the concept we would like to measure, i.e., software engineering conference health. Internal validity is related to validity of the inferences made based on application of our entire research methodology. External validity discusses generalisation of our findings beyond the data we have collected.

6.1. Construct validity

While assessing conference health has been the subject of a number of studies [13,65–67], operationalisation of the concept has never been studied on its own. Based on their experience as conference organisers, these authors used the number of submissions, the acceptance rate at the conference and the program committee's workload as indicators of the conference health [65]. They then formulated best practices for conference organisation such as hierarchical program committees, double blind reviewing and acceptance rate between 15% and 40% [66], and debated policy issues such as travel reduction and decoupling publication from presentation, as well as evaluation issues such as "do PCs tend to favour PC-authored papers?" [67]. From this perspective we observe that most of the metrics we proposed in Table 2 are closely related to the metrics used, best practices formulated and issues debated in the aforementioned publications. It seems, therefore, that the research community is converging to a shared understanding of the concept of conference health.

Nevertheless, in the future we aim to assess the robustness of some of our metrics, such as the *PNA* metric or the *SR* metric that relies on a particular definition of the notion of *core author*. If we would use another variant of these metrics, how would this affect the results we obtained?

Further threats to construct validity are related to the concepts of "prestige" and "review load". While a number of citation-based measures have been proposed in the literature to operationalise "prestige", these measures do not distinguish between different contexts of citations. Moreover, a smaller, focused conference will necessarily have fewer submissions than a large wide-scope conference, and yet, the smaller conference may carry more prestige in its area than a wide-scoped one. In a similar way, we define the review load without taking into account the help of subreviewers, presence of non-reviewing PC members, e.g., PC chairs, or characteristics of individual submissions, such as the number of pages or quality of writing, that can facilitate or hinder the reviewing process. Moreover, while we have not found evidence for application of more elaborate reviewing models such as the Program Board model (Section 3.1) among the conferences we have studied, such a model might have been used informally. This suggests that a more fine-grained distinction between different PC members might be needed to better quantify the review load.

6.2. Internal validity

Recall that the three steps of our methodology (Section 2) are data extraction, metrics calculation, and data analysis and interpretation. Most internal validity threats related to the *data extraction* step stem from possible incompleteness of the data sources we have used: e.g., while the SHINE database contains information about circa 1800 conferences,[9] as any such collection it is inherently incomplete, and therefore the $CI(c)$ values might underestimate the true value of the conference h-index. Information about composition of the program committees has been taken from conference websites, proceedings volumes, the Wayback machine archive as well as announcements posted by conference organisers in Usenet newsgroups. Each one of these sources might be incomplete due to, e.g., editorial oversight. In the same way, information about the conference authors has been taken from DBLP that might have been incomplete as well. Finally, the number of submissions $\#SP(c, y)$ has been extracted from the prefaces in the proceedings volume. While many conferences employ a two-phase submission process (first an abstract is submitted, then, usually a week later, a full paper), in a number of cases the prefaces ambiguously referred to the "number of submissions" which can refer either to number of received abstracts or number of received full papers.

Identity merging across multiple data sources has been explicitly recognised as a reliability threat [68]. Automatic merging techniques might have missed multiple representations of the same individual or considered multiple individuals to be one and the same person. To counteract this threat we have consulted DBLP pages, that frequently indicate different names of the same researcher,[10] and also reviewed the results of the identity merging and corrected them manually.

To ensure validity of statistical inferences we have paid particular attention to the use of appropriate statistical techniques.

[9] http://shine.icomp.ufam.edu.br/about.php.

[10] E.g., http://www.informatik.uni-trier.de/~ley/pers/hd/a/Andrews:Anneliese_Amschler.

6.3. External validity

We believe that the eleven conferences we have studied are a representative sample of well-established software engineering conferences. As described in Section 2.1 our study covers six conferences studied in [13], three additional wide-scoped conferences and two additional young conferences. Our results might not be generalisable to workshops, regional conferences, non-academic conferences or conferences outside of the software engineering domain.

7. Future work

The analysis carried out in this article can be extended in many different ways. In this section we present what we believe to be the most interesting future research directions.

7.1. More objects of study

The first group of future work directions pertains to enriching the data collected by including more objects in our study.

For instance, throughout our analysis, we restricted ourselves to research papers submitted to the main conference track only. This was a deliberate choice since we did not want to make the analysis overly complex. However, some conferences allow for both long and short papers in the main track. Both types of papers are quite different in nature and should perhaps be accounted for differently. Many conferences also offer different kinds of parallel tracks, in order to facilitate presentation of, e.g., early research achievements, PhD research, industrial results, research tools, and research projects.

A straightforward but labour-intensive extension would consist in replicating our study for other computer science research domains (e.g. databases, artificial intelligence, theoretical computer science). Such replications would also allow one to relate the conference health factors to the considered domain.

While we focused on software engineering *conferences*, it would be useful to apply a similar approach to software engineering *journals*. Many of the metrics proposed would need to be modified to take into account the specificities of journal publication. For example, there is no such thing as a "journal programme committee"[11] since, for every submitted paper, reviewers are assigned "ad hoc" based on the topic of the paper and the expertise and availability of reviewers. It is probably also much harder to obtain historical data about the reviewers for papers that have been published in a particular journal. We are not aware of such data being freely available. Another main difference is that journals do not tend to work with submission deadlines.[12] Papers can be submitted at any time, and the time and process required for reviewing, revising, and resubmitting papers is also much more flexible than for conference papers. We would also have to come up with new metrics that reflect characteristics that are important for journal publications such as, for example, the average time from submission to final acceptance of the paper.

7.2. More information about the study objects

Rather than extending the study to include more study objects, one can consider enriching the dataset by including additional information about the study objects (conferences) we have already considered. First of all, while we excluded short-paper tracks from consideration, we have included in our data submissions to the main conference track that have been accepted as short papers. Distinction between main track submissions accepted as full papers and accepted as short papers would allow us to investigate impact of the differences between the two kinds of papers on, e.g., openness and introversion.

We can also record additional information about authors and PC members, e.g., their gender (which could be inferred automatically based on their names, e.g., as described in [69,70]), geographical location, seniority, and research expertise. Presence of this information would allow us to obtain additional insights in representativeness of the PC with respect to the author community in terms of gender, geographical location, seniority, and research expertise (cf. Section 3.5). Moreover, availability of this additional information about the PC members, would allow us to extend the charter adherence study initiated in Section 3.2. For instance, the ICSM charter explicitly states that the "Program Chairs should make every effort to achieve diversity on the PC with respect to gender, geographic distribution, experience, and industry versus academic experience" and requires PC to include members "whose areas of expertise sufficiently cover" different sub-areas of ICSM-related research. Furthermore, we can also check whether evidence can be found for PC chairs/General conference chairs favouring committee members of a certain gender, geographical location, seniority, and research expertise.

Information about the PC members' research expertise would make it feasible, with some effort, to come up with a fully objective measure of the conference scope, allowing us to refine the distinction between wide-scope and narrow-scope conferences considered in the current article. To achieve this, in addition to the PC members' research expertise one would need to extract the solicited research topics as found in the call for papers and the websites of each conference, as well as the topics of the actually accepted and published papers. Most software engineering conferences publish their proceedings

[11] The so-called editorial board is not the same, since it does not contain the full list of reviewers of journal papers.

[12] We exclude here the so-called journal "special issues".

through a digital library (IEEE, ACM, or Springer) that requires to follow a strict classification scheme of the paper topic and subject area. Based on this, tracking and analysis the different software engineering topics that are covered by each conference, should be possible, which in their turn can form the basis of a new metric that objectively quantifies the narrowness of scope of the conference.

7.3. Using the data

The final group of future work directions is related to possible applications and extensions of our work. To start with, we could consider studying the probability of paper acceptance. While the acceptance ratio $RA(c, y)$ can be used as a first rough estimate, better techniques should include information about the authorship (e.g., is the paper more likely to be accepted if the author is also member of the PC or a frequent co-author of some PC members?) and topic(s) of the paper. Based on a more refined acceptance probability estimate, one can also consider modelling the prospective author behaviour by incorporating the effort she needs to put in preparing a submission, and potential benefits such as increased scientific visibility. Similarly to positive emotions affecting work engagement mediated by hope [71], we expect that positive emotions associated with a paper accepted at a prestigious conference in the preceding year give rise to hope, making the author to be more inclined to submit the paper in the year afterwards.

As mentioned in Section 4, deciding whether to submit a paper to a conference can be seen as an MCDM problem [50] with openness, introversion and prestige metrics as decision making criteria. Similarly, conference organisers or steering committees can use MCDM to assess conference health, using any of the considered health criteria and associated health metrics of their liking. Using these kinds of analyses, conference organisers can furthermore obtain additional insights in the impact of different policies recommended in the literature, such as increasing the acceptance rate, on future submissions (cf. [47]) and on evolution of the conference communities ("how will the GPCE author community look like in 2020?"). We believe that this kind of customized analysis is far better than trying to come up with a single index of conference health or a "one size fits all" threshold. Nevertheless, the application of MCDM techniques is, in our case, challenged by the temporal dimension of the data, as well as by the limited number of conferences considered compared to the number of criteria.

Another promising research direction pertains to collaboration between researchers, both at intra-conference level (cohesion) and inter-conference level (coupling). *Conference cohesion* can be evaluated by studying topics of the papers accepted and cooperation between the authors submitting to the same conference: if the author community can be clearly separated into distinct groups not linked by common publications or research topics, the conference is essentially an amalgamation of several disconnected communities. Lack of cohesion might also explain high PC turnover. *Coupling* between different conferences arises if the same persons are PC members or authors of different conferences. The amount of coupling between conferences may play a role in how some of the conference health indicators evolve over time. For example, when renewing the PC, chairs can easily look for new qualified and willing PC candidates in "coupled" conferences. Improved knowledge of conference coupling would allow one to replicate the study on *groups* of related conferences. Indeed, if there is a high overlap in the author and PC communities of two conferences, these conferences could be aggregated into one since an author that has frequently published at one of the venues cannot be reasonably considered as a "new face" for the other. Finally, empirical evidence of high coupling between conferences may help the conference organisers to adapt conference organisation, e.g., by merging or co-locating the events or by trying to differentiate them more. Significant overlaps are likely to be present for some of the narrow-scoped conferences we have studied (in particular, ICSM, WCRE and CSMR that target more or less the same subdomains of software engineering). This is one of the main reasons why CSMR and WCRE have decided to merge their conferences into a single event, called CSMR-WCRE in 2014.

Finally, one can study "migration" of researchers from one conference to another, from one research topic to another, akin to migration of software developers between different projects within the GNOME ecosystem [72].

8. Conclusions

The goal of this article was to assess how the health of software engineering conferences evolves over time, in terms of a variety of criteria: stability and representativeness of the PC with respect to the authors' community, openness to new authors, introversion, availability of PC candidates and scientific prestige. To this extent, we proposed a suite of metrics aiming to measure these health indicators. The used metrics and methodology are applicable to other scientific conferences as well.

Our analysis, covering 11 software engineering conferences over a period of more than 10 years, indicate that these conferences are relatively healthy: balanced PC turnover (high enough to avoid introversion, 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 introversion (in terms of fraction of papers co-authored by PC members). Nonetheless, we observed important differences across conferences according to the aforementioned criteria, suggesting different strengths, weaknesses, opportunities and threats. We also observed important differences between wide-scoped and narrow-scoped software engineering conferences.

We do not believe in identifying a "holy grail" of conference health assessment, i.e., a small set of metrics or optimal values that are universally indicative of conference health. Based on our previous experience with software metrics, as well as our experience as author, PC member or PC chair of different software engineering conferences, we are convinced that

consensus on optimal metrics values is hard to achieve and, once achieved, will miss conference-specific context. Instead, we presented a range of comparative analyses and countermeasures that can be useful for steering committees, programme committees, prospective authors and researchers in different ways. A range of strategies could be used to increase conference health, such as: 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, reducing reviewer load by resorting to a Program Board review model or a "rolling deadline" model (such as the one followed by the VLDB conference series), merging conferences together or differentiating them better.

Acknowledgements

We thank Dave Binkley, Prem Devanbu, Vladimir Filkov, Leon Moonen, David Rosenblum and the anonymous reviewers for their very useful feedback on earlier versions of this paper. We are grateful to Dr. Frank Konietschke for providing us with the implementation of the $\widetilde{\mathbf{T}}$ procedure. We also thank Andrei Jalba and Michel Westenberg for insightful comments on how to improve the visualisation. 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), and by F.R.S-FNRS research grant BSS-2012/V 6/5/015 (Fonds de la Recherche Scientifique) during the second author's stay at the Université de Mons. None of the funding agencies was involved in the 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

[1] J. Chen, J.A. Konstan, Conference paper selectivity and impact, Commun. ACM 53 (6) (2010) 79–83.
[2] M. Franceschet, The role of conference publications in CS, Commun. ACM 53 (12) (2010) 129–132.
[3] B. Meyer, C. Choppy, J. Staunstrup, J. van Leeuwen, Viewpoint – Research evaluation for computer science, Commun. ACM 52 (4) (2009) 31–34.
[4] D. Patterson, L. Snyder, J. Ullman, Best practices memo: evaluating computer scientists and engineers for promotion and tenure, Comput. Res. News 11 (4) (August 1999) A–B.
[5] L. Fortnow, Viewpoint – time for computer science to grow up, Commun. ACM 52 (8) (2009) 33–35.
[6] M.Y. Vardi, Conferences vs. journals in computing research, Commun. ACM 52 (5) (2009) 5.
[7] J. Crowcroft, S. Keshav, N. McKeown, Viewpoint: Scaling the academic publication process to internet scale, Commun. ACM 52 (1) (2009) 27–30.
[8] K. Birman, F.B. Schneider, Viewpoint – program committee overload in systems, Commun. ACM 52 (5) (2009) 34–37.
[9] J. Freyne, L. Coyle, B. Smyth, P. Cunningham, Relative status of journal and conference publications in computer science, Commun. ACM 53 (11) (2010) 124–132.
[10] M. Franceschet, The skewness of computer science, Inf. Process. Manag. 47 (1) (2011) 117–124.
[11] M. Eckmann, A. Rocha, J. Wainer, Relationship between high-quality journals and conferences in computer vision, Scientometrics 90 (2) (2012) 617–630.
[12] H.V. Jagadish, The conference reviewing crisis and a proposed solution, SIGMOD Rec. 37 (3) (2008) 40–45.
[13] T. Systä, M. Harsu, K. Koskimies, Inbreeding in software engineering conferences, http://www.cs.tut.fi/~tsysta/, accessed November 2013.
[14] B. Vasilescu, A. Serebrenik, T. Mens, A historical dataset of software engineering conferences, in: Int'l Conf. Mining Software Repositories (MSR), IEEE, 2013, pp. 373–376.
[15] The DBLP computer science bibliography, http://dblp.uni-trier.de, accessed October 2012.
[16] G. Robles, J. González-Barahona, Developer identification methods for integrated data from various sources, in: MSR, ACM, 2005, pp. 1–5.
[17] C. Bird, A. Gourley, P. Devanbu, M. Gertz, A. Swaminathan, Mining email social networks, in: MSR, ACM, 2006, pp. 137–143.
[18] W. Poncin, A. Serebrenik, M.G.J. van den Brand, Process mining software repositories, in: CSMR, IEEE Computer Society, 2011, pp. 5–14.
[19] E. Kouters, B. Vasilescu, A. Serebrenik, M.G. van den Brand, Who's who in GNOME: using LSA to merge software repository identities, in: Int'l Conference Software Maintenance – ERA Track, IEEE, 2012, pp. 592–595.
[20] M. Goeminne, T. Mens, A comparison of identity merge algorithms for software repositories, Sci. Comput. Program. 78 (8) (2013) 971–982.
[21] B. Vasilescu, A. Serebrenik, M. Goeminne, T. Mens, On the variation and specialisation of workload—a case study of the Gnome ecosystem community, Empir. Softw. Eng. (2013), http://dx.doi.org/10.1007/s10664-013-9244-1.
[22] B. Vasilescu, A. Serebrenik, P.T. Devanbu, V. Filkov, How social Q&A sites are changing knowledge sharing in open source software communities, in: CSCW, 2014, pp. 342–354.
[23] Simple *h*-index estimation, http://shine.icomp.ufam.edu.br/index.php, accessed November 2013.
[24] J.E. Hirsch, An index to quantify an individual's scientific research output, Proc. Natl. Acad. Sci. USA 102 (46) (2005) 16569–16572.
[25] A. Capiluppi, A. Serebrenik, A. Youssef, Developing an h-index for OSS developers, in: Int'l Conf. Mining Software Repositories (MSR), 2012, pp. 251–254.
[26] R Development Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2010.
[27] W. Javed, B. McDonnel, N. Elmqvist, Graphical perception of multiple time series, IEEE Trans. Vis. Comput. Graph. 16 (6) (2010) 927–934.
[28] P. Cowpertwait, A. Metcalfe, Introductory Time Series with R, Springer, 2009.
[29] M. Holander, D.A. Wolfe, Nonparametric Statistical Methods, Wiley, 1973.
[30] F. Wilcoxon, Individual comparisons by ranking methods, Biom. Bull. 1 (6) (1945) 80–83.
[31] O.J. Dunn, Multiple comparisons among means, J. Am. Stat. Assoc. 56 (1961) 52–64.
[32] D.J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, 4th edition, Chapman & Hall, 2007.
[33] K.R. Gabriel, Simultaneous test procedures—some theory of multiple comparisons, Ann. Math. Stat. 40 (1) (1969) 224–250.
[34] D.W. Zimmerman, B.D. Zumbo, Parametric alternatives to the Student t test under violation of normality and homogeneity of variance, Percept. Mot. Skills 74 (3(1)) (1992) 835–844.
[35] E. Brunner, U. Munzel, The nonparametric Behrens–Fisher problem: asymptotic theory and a small-sample approximation, Biom. J. 42 (1) (2000) 17–25.
[36] F. Konietschke, L.A. Hothorn, E. Brunner, Rank-based multiple test procedures and simultaneous confidence intervals, Electron. J. Stat. 6 (2012) 738–759.
[37] T. Jaki, L. Hothorn, Statistical evaluation of toxicological assays: Dunnett or Williams test—take both, Arch. Toxicol. 87 (11) (2013) 1901–1910, http://dx.doi.org/10.1007/s00204-013-1065-x.
[38] Y. Dajsuren, M.G.J. van den Brand, A. Serebrenik, S. Roubtsov, Simulink models are also software: modularity assessment, in: Int'l ACM Sigsoft Conf. Quality of Software Architectures, QoSA '13, ACM, 2013, pp. 99–106.
[39] B. Vasilescu, V. Filkov, A. Serebrenik, StackOverflow and GitHub: Associations between software development and crowdsourced knowledge, in: SocialCom/PASSAT, IEEE, 2013, pp. 188–195.

[40] B.M. Brown, T.P. Hettmansperger, Kruskal–Wallis, Multiple comparisons and efron dice, Aust. N. Z. J. Stat. 44 (4) (2002) 427–438.
[41] G. Antoniol, M. Di Penta, M. Harman, Search-based techniques applied to optimization of project planning for a massive maintenance project, in: Int'l Conf. Software Maintenance (ICSM), IEEE, 2005, pp. 240–249.
[42] F. Khomh, M. Di Penta, Y.-G. Guéhéneuc, An exploratory study of the impact of code smells on software change-proneness, in: Working Conf. Reverse Engineering, IEEE, 2009, pp. 75–84.
[43] F. Konietschke, nparcomp. Reference Manual, 2012.
[44] C. Bird, E.T. Barr, A. Nash, P.T. Devanbu, V. Filkov, Z. Su, Structure and dynamics of research collaboration in computer science, in: SDM, 2009, pp. 826–837.
[45] W.S. Martins, M.A. Gonçalves, A.H.F. Laender, G.L. Pappa, Learning to assess the quality of scientific conferences: a case study in computer science, in: Joint Conf. Digital Libraries (JCDL), 2009, pp. 193–202.
[46] Y. Manolopoulos, A statistic study for the ADBIS period 1994–2006, in: ADBIS Research Communications, in: CEUR Workshop Proceedings, vol. 215, CEUR-WS.org, 2006.
[47] P. Laplante, J. Rockne, P. Montuschi, T. Baldwin, M. Hinchey, L. Shafer, J. Voas, W. Wang, Quality in conference publishing, IEEE Trans. Prof. Commun. 52 (2) (2009) 183–196.
[48] P. Küngas, S. Karus, S. Vakulenko, M. Dumas, C. Parra, F. Casati, Reverse-engineering conference rankings: what does it take to make a reputable conference?, Scientometrics (2013) 1–15.
[49] R. Snodgrass, Single- versus double-blind reviewing: an analysis of the literature, SIGMOD Rec. 35 (3) (2006) 8–21.
[50] S. Zionts, MCDM—if not a roman numeral, then what?, Interfaces 9 (4) (1979) 94–101.
[51] E. Triantaphyllou, Multi-Criteria Decision Making Methods: A Comparative Study, Applied Optimization, Springer, 2000.
[52] M. Biryukov, C. Dong, Analysis of computer science communities based on DBLP, in: European Conf. Research and Advanced Technology for Digital Libraries (ECDL), 2010, pp. 228–235.
[53] E. Elmacioglu, D. Lee, Oracle, where shall I submit my papers?, Commun. ACM 52 (2) (2009) 115–118.
[54] Z. Zhuang, E. Elmacioglu, D. Lee, C.L. Giles, Measuring conference quality by mining program committee characteristics, in: Joint Conf. Digital Libraries (JCDL), 2007, pp. 225–234.
[55] S. Sakr, M. Alomari, A decade of database conferences: A look inside the program committees, Scientometrics 91 (1) (2012) 173–184.
[56] O. Inanc, O. Tuncer, The effect of academic inbreeding on scientific effectiveness, Scientometrics 88 (3) (2011) 885–898.
[57] J.P.C. Kleijnen, W.J.H.V. Groenendaal, Measuring the quality of publications: new methodology and case study, Inf. Process. Manag. 36 (4) (2000) 551–570.
[58] M.A.M. Souto, M. Warpechowski, J.P.M. Oliveira, An ontological approach for the quality assessment of computer science conferences, in: Advances in Conceptual Modeling – Foundations and Applications, in: Lecture Notes in Computer Science, vol. 4802, Springer, 2007, pp. 202–212.
[59] W.S. Martins, M.A. Gonçalves, A.H.F. Laender, N. Ziviani, Assessing the quality of scientific conferences based on bibliographic citations, Scientometrics 83 (1) (2010) 133–155.
[60] R. Klamma, M.C. Pham, Y. Cao, You never walk alone: Recommending academic events based on social network analysis, in: Int'l Conf. Complex Sciences, 2009, pp. 657–670.
[61] R. da Silva, J.P.M. de Oliveira, J.V. de Lima, V. Moreira, Statistics for ranking program committees and editorial boards, CoRR abs/1002.1060.
[62] M.C. Pham, R. Klamma, M. Jarke, Development of computer science disciplines: a social network analysis approach, Soc. Netw. Anal. Min. 1 (4) (2011) 321–340.
[63] S. Yan, D. Lee, Toward alternative measures for ranking venues: a case of database research community, in: Joint Conf. Digital Libraries (JCDL), 2007, pp. 235–244.
[64] C. Gini, Measurement of inequality of incomes, Econ. J. 31 (1921) 124–126.
[65] D.A. Patterson, The health of research conferences and the dearth of big idea papers, Commun. ACM 47 (12) (2004) 23–24.
[66] M.D. Hill, J.-L. Gaudiot, M. Hall, J. Marks, P. Prinetto, D. Baglio, A wiki for discussing and promoting best practices in research, Commun. ACM 49 (9) (2006) 63–64.
[67] J.C. Mogul, T. Anderson, Open issues in organizing computer systems conferences, SIGCOMM Comput. Commun. Rev. 38 (3) (2008) 93–102.
[68] J. Howison, K. Crowston, A. Wiggins, Validity issues in the use of social network analysis with digital trace data, J. Assoc. Inf. Syst. 12 (2011), article 2.
[69] B. Vasilescu, A. Capiluppi, A. Serebrenik, Gender, representation and online participation: A quantitative study of StackOverflow, in: ASE International Conference on Social Informatics, IEEE, 2012, pp. 332–338.
[70] B. Vasilescu, A. Capiluppi, A. Serebrenik, Gender, representation and online participation: A quantitative study, Interact. Comput. (2013), http://dx.doi.org/10.1093/iwc/iwt047.
[71] E. Ouweneel, P.M. Le Blanc, W.B. Schaufeli, C.I. van Wijhe, Good morning, good day: A diary study on positive emotions, hope, and work engagement, Hum. Relat. 65 (9) (2012) 1129–1154.
[72] T. Mens, M. Claes, P. Grosjean, A. Serebrenik, Studying evolving software ecosystems based on ecological models, in: T. Mens, A. Serebrenik, A. Cleve (Eds.), Evolving Software Systems, Springer, 2013, pp. 297–326.