Scientometrics manuscript No.
(will be inserted by the editor)

Industry-academia collaborations in software engineering: A bibliometric study

João M. Fernandes · Paulo Cortez ·
Vahid Garousi

Received: date / Accepted: date

Abstract Research collaboration between industry and academia is crucial to boost improvements and innovation in industry and to ensure practical relevance of academic research. This study analyses the research literature of industry-academia collaborations (IAC) in the filed of software engineering (SE), by investigating three objectives: (1) the ratio and publication trend of IAC papers to all SE paper, (2) the SE topics that appeared the most in IACs, and (3) the most active countries in IACs. Our dataset is composed of all papers published in five prestigious journals. We develop a technique to analyse the authors affiliations and to classify them into one of the following three categories: (1) all academic authors, (2) all industrial authors; or (3) collaboration of industry and academia. There has been a gradual growth in the annual number of IAC papers: less than 10 papers in 1970's; about 30 papers per year in 2010's; and, more recently, 53 papers in 2021 and 103 in 2022. However, the annual ratio of IAC papers has stayed quite steady (10–15% in the period 1994–2022, with a very few exceptions). The topics most represented in IAC-papers are testing, modelling, analysis, programming, and development. In terms of the countries, the United States is the leading country, followed by Canada, Germany, and the United Kingdom.

Keywords bibliometrics · software engineering · industry-academia collaborations · research literature

J.M. Fernandes (ORCiD: 0000-0003-1174-1966)
ALGORITMI Research Center / LASI, Dep. Informática, Universidade do Minho, 4710-057 Braga, Portugal. E-mail: jmf@di.uminho.pt.

P. Cortez (ORCiD: 0000-0002-7991-2090)
ALGORITMI Research Center / LASI, Dep. Sistemas de Informação, Universidade do Minho, 4804-533 Guimarães, Portugal. E-mail: pcortez@dsi.uminho.pt.

V. Garousi (ORCiD: 0000-0001-6590-7576)
Queen's University Belfast, Belfast, Northern Ireland, United Kingdom. E-mail: V.Garousi@qub.ac.uk.

1 Introduction

The worldwide Software Engineering (SE) community is generally divided into two communities: industrial SE practitioners and academic researchers. There are some rough estimates on the size of the two communities. According to a report by the Developer Nation (2021), there were about 24.3 million software engineers worldwide in 2021, and that number was estimated to reach 45.0 million by 2030. On the other hand, according to Briand (2012), the SE research community consists of 4,000 individuals who "actively publish in major [SE] journals", which can be used as an approximate measure for its size.

According to several SE researchers (Garousi et al. 2016; Briand et al. 2017; Garousi et al. 2020), the level of collaboration between the two communities (SE practitioners and academic researchers) is relatively insufficient, and much less, compared to the extent of activity in either of the two communities. It should be noted that the need for industry-academia collaborations (IAC) should not be seen as a surprise to the SE community, due to applied nature of the discipline (Garousi et al. 2020). Active IACs could boost improvements and innovation in the software industry and would ensure practical relevance of academic research, which constitute two important objectives. Diverse reasons have been discussed in the community about the low ratio of IACs, such as the difference of objectives between the two communities, industrial problems lacking scientific novelty or challenges, and low applicability or scalability of the approaches/solutions developed in academia (Briand 2012; Garousi et al. 2016, 2020).

As early as in the 1990's, Glass (1994) and Parnas (1998) were among the first pioneers who critiqued and investigated the state of IAC and research relevance in SE. Several more recent papers address various aspects of IACs in SE, e.g., (Carver and Prikladnicki 2018; Runeson 2012; Garousi et al. 2016, 2019). Furthermore, to highlight the importance of IAC and to discuss success stories on how to "bridge the gap", various workshops and panels are regularly organized in the context of international SE conferences. An example is a panel named "What industry wants from research", which was held at the International Conference on Software Engineering (ICSE) 2011. In that event, several insightful talks from companies such as Toshiba, Google and IBM, were presented. Another international workshop on the topic of long-term industrial collaborations on SE (called WISE) was organized in 2014. In 2016, another conference panel, named "the state of software engineering research" (Parnas 1998), was held, in which the panelists discussed the need for more IAC in SE. Similar activities have been continuing up to the present day.

SE celebrated 50 years in an event in 2018 (Ebert 2018), so it is timely to ask and study the following questions in the context of IACs: (1) what is the ratio and publication trend of IAC papers when compared to all SE papers? (2) what are the SE topics that appear the most in IACs?, and (3) what are the most active countries in IAC SE papers? To address the above questions, we report in this paper a bibliometric analysis of papers published in a subset of five top SE journals (selected based on impact factors), by classifying the papers by their author affiliation types (academic or industrial). Similar studies have been published in other fields, such as in oil industry (Gielfi et al. 2017) and fuel cells (Huang et al. 2015). As a related work in SE, a brief analysis of IAC papers and their ratios in the context of the Turkish SE community was reported by Garousi (2015). To the best of our knowledge, the current paper is the first study of its kind in the SE research literature in a more global scale.

The rest of this paper is organized as follows. In Section 2, we review the related work. We then present in Section 3 the research method used in this work. Results are presented and discussed in Section 4. Next, we summarize the findings, implications and limitations of this research in Section 5. Finally, we conclude and discuss the future work in Section 6.

2 Related work

We discuss the related works in two categories: (1) bibliometric analysis of IACs in other research fields; and (2) bibliometric studies in SE in general, and those including IACs.

2.1 Bibliometric analysis of IACs in other research fields

Bibliometric studies that analyse IACs in different fields of science are common, e.g,. in the fuel-cells domain (Huang et al. 2015), water treatment (Butcher and Jeffrey 2005), oil industry (Gielfi et al. 2017), and dental materials (Garrison et al. 1992). Certain studies are focused on specific geographies, e.g, IACs in Brazil (Garcia et al. 2020; Gielfi et al. 2017), Italy (Abramo et al. 2009), Japan (Sun et al. 2007), Latin America (Orduña Malea 2020), comparing China and the USA (Zhou et al. 2016), and New Zealand (Aref et al. 2018). These studies are related to our work in the sense that we also analyse IACs in the context of a specific field (Software Engineering). We review a few of the above works briefly next.

Tijssen et al. (2016) analyse the linkages between universities and industry, without focusing on any specific field. The study analyses how to rank large research-intensive universities worldwide, with an emphasis on the following metric: the share of what they called UICs (university–industry co-authored publications) within the total publication output of a given university. This UIC-based metric has been also used by some other well-known university rankings, such as the Reuters' Top-100 "The World's Most Innovative Universities"[1], Leiden[2], and U-Multirank[3].

The impact of long-term IACs on academic research productivity is analyzed by Garcia et al. (2020). This study, based on empirical evidence regarding collaborative projects involving research groups and companies in Brazil,

[1] www.reuters.com/innovative-universities-2019
[2] www.leidenranking.com
[3] www.umultirank.org

shows that long-term collaboration with industry has a positive impact on the scientific productivity of academic research groups. Research groups that collaborate with industry generally show higher production performance after starting collaboration projects.

The study by Abramo et al. (2009) discusses public-private research collaborations between Italian universities and industry, also using a bibliometric-based approach. The study analyses the co-authored publications in international journals, co-authored by Italian university researchers in collaboration with industrial practitioners. The authors observed that, among all areas of science that were investigated, Information Technology (IT) had the highest percentage of co-authored articles out of all articles.

2.2 Bibliometric studies in SE, including IACs

Bibliometric studies have become quite common in SE, and a number of bibliometric studies are regularly published in SE venues, each having a different focus, e.g., (Fernandes 2014; Fernandes and Monteiro 2017; Freitas and Souza 2011; Garousi and Mäntylä 2016; Garousi and Ruhe 2013; Garousi and Fernandes 2017; Wong et al. 2021). In addition to papers which are purely bibliometric-focused in SE, some other papers have reported additional assessments as part of their overall study, e.g., several Systematic Literature Review (SLR) and Systematic Literature Mapping (SLM), or just called mapping studies, fall in that category, e.g., the following two SLM studies assessed the citation landscapes and top authors of the two specific areas of graphical user interface (GUI) testing (Banerjee et al. 2013) and engineering of scientific software (Farhoodi et al. 2013).

A handful number of SE papers have studied the ratios of IAC papers in certain contexts and sub-fields, e.g,. (Garousi 2015; Banerjee et al. 2013). A brief analysis of IAC papers and their ratios in the context of the Turkish SE community is reported by Garousi (2015). The study indicates that, in general, the involvement of industry in Turkish SE paper is low. In the time-window under analysis (1992-2014), only four and 13 papers out of all the papers in the dataset (corresponding to 1.4% and 4.5% of the total) were solely by industrial authors or by a team of industry-academic authors. The remaining 272 papers (94.1% of the dataset) were authored by only academic authors. In recent years (until 2014), Garousi (2015) observed a small increase in the number of IAC papers, but he recommended that further increase is needed.

Another SLM study (Banerjee et al. 2013) classifies the set of all 230 papers published in the area of GUI (Graphical User Interface) software testing, between years 1991–2011. That SLM study analyses, in addition, the author affiliations of the papers, i.e., whether they belong to academia or to industry. The study classifies the papers as coming from one of the following three categories: academia, industry, and collaborations (papers jointly authored by authors from academia and industry). The ratios are: 74% (academia), 13% (industry), and 13% (collaboration). The authors observe that the number of collaborative papers between academics and practitioners were slowly on the rise, in the last years, in the field of GUI testing.

3 Methodology

As for the methodology that we used to plan, design and conduct our study, we discuss in this section the study goal, the research questions (RQs), and the selection of the publication venues and the sampling of SE papers. We then present the approach and tool that we developed for classifying IAC papers, for the purpose of addressing our RQs.

3.1 Research goal and questions

The goal of this work is to characterize the level and topics of IACs in the field of SE, based on a classification of author affiliations of papers published in five top SE journals. To address this goal, and using the Goal-Question-Metric (GQM) approach (van Solingen et al. 2002), we raised and studied the following research questions (RQs):

- **RQ1-** Ratio and publication trend of IAC papers in SE. We divided this RQ into two sub-RQs as follows:
  - **RQ1.1-** What is the ratio of IAC papers to all SE papers, over the years? Is there an increasing or decreasing trajectory?
  - **RQ1.2-** How does the ratio of IAC papers in SE compare to that ratio in other fields?
- **RQ2-** Which SE topics have been covered the most by IAC papers?
- **RQ3-** Which countries are the most active as measured by the number of IAC papers?

3.2 Approach and tool for classifying papers by author affiliations

To answer the RQs of this study and to distinguish the subset of SE papers which have resulted from IACs, we used a simple but effective heuristic. Our approach was to analyse the author affiliation information of a given paper and classify each author into one of the following three categories: (1) Academic author, (2) Industrial author; (3) Author from the public sector (e.g., government).

Academic authors are those affiliated with academic organizations including: universities, institutes of technology, institutes of science, colleges, etc. Industrial authors, on the other hand, are those affiliated with commercial entities such as: companies, corporations, banks, and alike. Similarly, a third category of authors are from the public sector, e.g, government units such as ministries, municipalities, police departments, tax agencies, etc.

Our classification approach is based on a simple pattern-matching algorithm that searches for certain keywords in the authors' affiliations and their email addresses, and then performs the classification. The predefined keywords for each of the three classifications are defined in a textual "dictionary" file.

We developed and populated the three dictionary files in an empirical and iterative fashion, i.e., by manually analysing the results of the classifications on the papers and expanding or revising the set of terms in each dictionary file. For example, we started the dictionary for academic authors with the following strings (terms): "university" and "college". For instance, by manually analysing the results of the classifications on a subset of the papers, we added the following terms to the academic authors dictionary in our iterative process: "Polytech", "Univ", "Institut", "Faculty" and ".edu" in email addresses. We should clarify that the pattern-matching is done in a sub-string manner. For example, if the affiliations of a given paper includes the string "Institut" in any spelling (e.g., French or English), the pattern-matching procedure returns true. For industrial organizations, the dictionary included terms such as "Corporation", "GmbH", "Inc", "Limited", and "Ltd". To detect governmental affiliations in the papers, we included the following keywords: "government", "ministry", "police", and "municipality".

During the iterative preparation of the dictionaries for the classification, some of the words that were being used were made shorter to match more languages/spellings. For example, "university" was shortened to "univ" to also match with non-English terms, such as "Universidade", "Universidad", "Università", "Université" and "Universität". The same happened, for example, with "Corporation", which was reduced to just "Corp".

Another heuristic that we applied to our classification tool to increase the precision of the matching was the introduction of specific terms. For example, "ETH Zurich", "Fondazione Bruno Kessler", "Fraunhofer", "GMD", "LERO", "M.I.T." were added to the academic dictionary. Similarly, "AEG", "AT&T", "Boeing", "Cisco", "Facebook", "Google", "IBM", "Intel", "Microsoft", "PricewaterhouseCoopers", and "SAP", were added to the industrial dictionary. These specific terms were defined by performing a manual inspection and reviewing on the considered dataset (details in the next sub-section).

To automate our approach, we developed a tool in Python language, using the above specifications of the approach. The latest version of the tool consists of 550 lines of code.

To benefit other researchers who are interested in conducting similar studies and also to make our study reproducible, we have made our tool open-source in www.github.com/vgarousi/IACinSE. We have also provided the entire dataset in the same repository.

3.3 Preparing the papers dataset (sampling a subset of SE papers)

Initially, we intended to develop a tool that would allow us to classify any given SE paper without the need for user validation/intervention. This could have enabled us to run the tool on "all" SE papers. However, in the process of evaluating the tool by running it on a representative subset of all SE papers, we observed that the affiliation information recorded in many papers did not have enough information to enable automated classification. This was mainly due to the "free-text" nature of how author affiliations are written inside the papers and stored in paper databases. For example, the affiliation of an author in a paper was "Waymo, Mountain View, California, United States", which, once checked by a human, appears to be a company, and not an academic or governmental affiliation. Defining rules to classify such ambiguous cases by the tool was not straightforward and it could possibly need some artificial intelligence or other sophisticated approaches, which we decided not to pursue in this initial work of ours in this subject.

In terms of our heuristics-based approach discussed above, we should also mention a caveat that it could be possible that a team of academic researchers and industry practitioners may engage in an IAC, but afterwards only the researchers, without the involvement of the practitioners, write and publish papers from outputs of the their project. In such cases, based on author affiliations on those papers, obviously, our tool is not able to label those papers as IAC papers. This could be seen as a limitation, but obviously there is no automatic mechanism to infer the existence of an IAC from such papers. Also, related to this discussion, there was a keynote talk, entitled "Software Engineering Research under the Lamppost", given by Wohlin (2013), in which he presents five different levels of "closeness" (collaboration) between industry and academia: (1) not in touch, (2) hearsay, (3) sales pitch, (4) offline, and (5) [working as] one team. It is only in level 5 (the most collaborative level) that partners from both sides co-author and co-publish papers from their joint IAC projects. Thus, in our work, we consider and detect such level-5 papers.

If we had to run the classification tool on all SE papers, manual inspection of the classification results, generated by the tool, could have been highly effort-intensive. Therefore, we soon found out that we had to select a subset of all SE papers to be analysed by our tool / study.

It is also worth discussing the estimated number of all SE papers published so far in the field. A study presented by Garousi and Mäntylä (2016) was a bibliometric analysis of all SE papers, published in the period 1969-–2014. The search phase of that study (phrase "software" in the filed "source title") in the Scopus database (www.scopus.com) identified a set of 69,540 SE papers. We repeated the same search method in March 09, 2023 and the Scopus system returned a bigger set of 166,541 SE papers.

For sampling a subset of SE papers, we followed the process depicted in Fig 1 that is next explained in detail.

We chose the following five highly prestigious scientific SE journals, based on their impact factors using data from an online source[4]:

- **TSE**: IEEE Transactions on Software Engineering, ISSN 0098-5589, IEEE, 1975–2022;

[4] www.robertfeldt.net/advice/se_venues

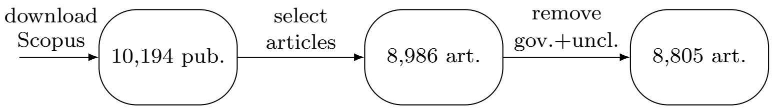

Fig. 1 The data preparation process.

- **EMSE**: Empirical Software Engineering, ISSN 1382-3256, Springer, 1996–2022;
- **TOSEM**: ACM Transactions on Software Engineering and Methodology, ISSN 1382-3256, ACM, 1992–2022;
- **ASE**: Automated Software Engineering, ISSN 1573-7535, Springer, 1994–2022; and
- **IST**: Information and Software Technology, ISSN 0950-5849, Elsevier, 1987–2022.

Sampling a subset of all SE papers is a common practice in bibliometric studies. Wainer et al. (2013) and Fernandes and Monteiro (2017) use the exact above list of journals in their studies.

Data for this bibliometric study were obtained from the Scopus repository, through its online search facility available, for the above five journals, covering the period 1975-2022. Data retrieval from Scopus was done on March 09, 2023. We should add that the metadata downloaded from Scopus includes the following data attributes: Authors, Document title (meaning paper title), Year, Source (meaning: venue / journal name) and document type, DOI (Digital Object Identifier), affiliations, author keywords (keywords chosen by the authors), Index keywords (standard keywords assigned by the journal).

Glass, Wong and their collaborators have published several annual bibliometric reports on SE, aimed at identifying the top scholars and institutions in certain time windows, e.g., (Glass 1995; Wong et al. 2021). The first report (Glass 1995) is focused on the time period of 1993–1994, while the last one (Wong et al. 2021), as of this writing, covered the period 2013-2020. Until 2001-2005, only six journals (IST, TSE, TOSEM, Journal of Systems and Software, Software Practice and Experience, and IEEE Software) were considered, based on their relevance to SE and their reputation in the field as reflected by bibliometric indices, such as the journal impact factor (Wong et al. 2008). For the period 2002–2006, a seventh journal (EMSE) was added to the set (Wong et al. 2009). We particularly note that all five journals of our bibliometric study belong to the set adopted by the above series of studies (Wong et al. 2009).

Our collected data considers publications for a 48-year period (1975–2022). The numbers of papers considered for each journal are presented in Table 1. Our dataset includes metadata related with 10,194 publications in total, which are classified in Scopus under different categories: e.g., articles, reviews, editorials, errata, letters, notes, short surveys, and retracted papers. Publications of type "article" (or article in press) are in fact the technical research papers,

	TSE	EMSE	TOSEM	ASE	IST	Total
# of publications	4,217	1,385	656	417	3,519	10,194
# of articles	3,895	1,229	582	319	2,961	8,986
% of articles	92%	89%	89%	76%	84%	88%
% in our dataset	43.3%	13.7%	6.5%	3.5%	33.0%	100%
# other publications	322	156	74	98	558	1,208

Table 1 Distribution of retrieved SE publications in the five selected journals.

and our rational intention in this work was to consider only those types of publications. We provide the relevant statics in Table 1. For example, out of the 4,217 publications which have appeared in TSE, 3,895 (%92.3) were "articles" and the rest were other publication types.

In summary, as mentioned in Table 1, we consider in this study a total of 8,986 scientific articles, which are further analysed in the rest of this paper. As can be seen in Table 1, for the five selected journals, the percentage of articles among all publication types range between 76% and 92%. TSE (43.3%) and IST (33.0%) are the two journals with the highest contributions to the paper dataset in this study.

In order to characterize our dataset, Figure 2 shows the scale of our sample of SE papers, selected from the top-5 SE journals, versus the total number of SE papers (the "All" curve) through the years (x-axis). The total number of SE papers was gathered from the Scopus database using the methodology proposed in (Garousi and Fernandes 2017). The graph reveals that our sample corresponds to a percentage of all SE papers that ranges from around 27% (years 1987 and 1988) to less than 2% (years of 2010 and 2011). These ratios denote the cumulative set size of all the papers published in the top-5 SE journals, to all the papers published in all SE journals and conferences, which, by themselves, are interesting measures.

From another standpoint, after some initial manual inspections on the dataset, we saw the need for data cleansing. This was done to correct typographical errors whenever identified. For example, in the author affiliations, we changed "Univers ta", "Versity", "Umversity", and "Uiversity" to the correct spelling. These manual modifications were considered important to allow the correct execution of the automatic process of identifying the type of organizations (academia vs. industry) that authors are affiliated with. In addition, other corrections were also made in the author's institutions. For example, "MIT" was replaced with "M.I.T.", since "MIT" is a short string that also occurs in other affiliations, such as "MITRE Corporation". We also performed other manual corrections, to enhance the quality of the author affiliations algorithm. For instance, the following cited paper (Alexandru et al. 2019) only mentioned a research lab's name as its affiliation: "Software Evolution and Architecture Lab - s.e.a.l., Binzmühlestrasse 14, Zürich, CH-8050, Switzerland". In this case, we added the term "University Zürich". As another example, in another paper written (Laukkanen et al. 2018), authors affiliation is indicated as "Department of Computer Science, PO Box 15400, Aalto, 00076, Finland".

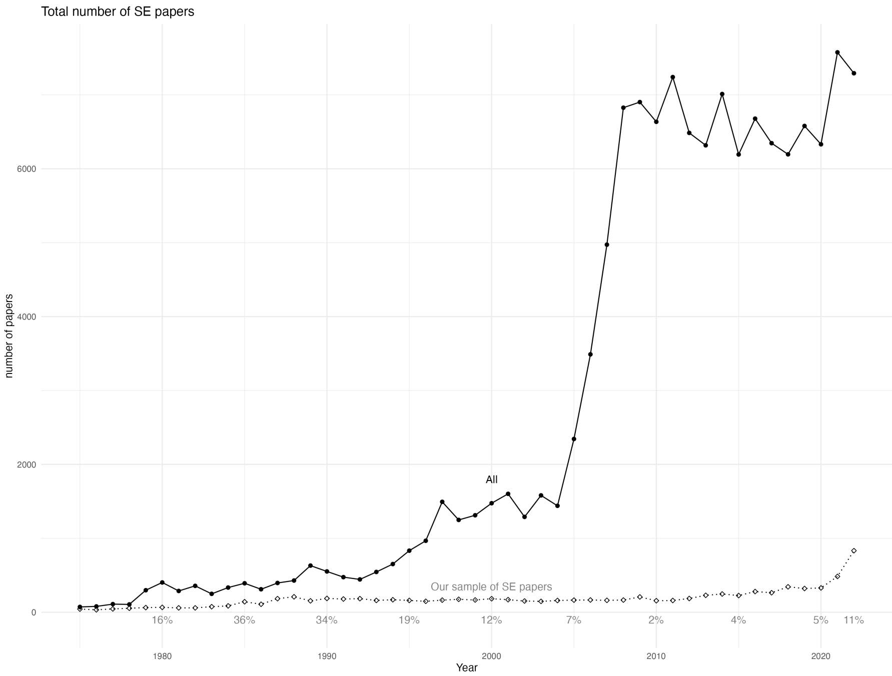

Fig. 2 Comparison between our the annual number of papers in our sampled dataset (gray curve, diamond points) versus the total annual number of SE papers (black curve, circle points).

In this case, we added the phrase “Aalto University”. Similarly, to allow the correct identification of the countries of authors belong to, several changes were performed, such as: “Can” was changed to “Canada”, “Engl” and “Northern Ireland” to “United Kingdom”, “Jpn” to “Japan”, “Russian Federation” to “Russia”, “West Ger” to “Germany”, and so on.

3.4 Applying the classification tool to the SE papers

Once we empirically and iteratively conducted the data cleansing on the dataset (as discussed above) and developed the classification tool, we applied it to the pool of retrieved SE papers.

For the 8,968 papers in the dataset, our tool identified, in total, 18,303 author affiliations. Note that the tool extracts all the different affiliations of each paper and, thus, since an author is often involved in several papers, her/his affiliations could occur in multiple affiliation “instances” across multiple papers. It should also be noted that an affiliation in a given paper can be related to several authors, if they all share it. We shall clarify that we conducted no analysis on author names (such as ranking), but rather we only analysed their affiliation information.

The tool has classified the input set of 18,303 affiliation instances as follows: (1) 15,408 academic, (2) 1,671 as industrial, and (3) only 31 as governmental. The classification tool labels the given affiliation instance as non-classified, whenever none of the pattern-matching rules using the specified terms in the three dictionaries would return true. A total of 1,372 instances were not classified, meaning that it was not possible to identify the type of affiliation based on the respective string that appears in the paper. Upon manual inspection of those records, we found that most of those cases had incomplete or simply null (empty) affiliation information, or had very specific entity names, such as: "Ant Financial", "Transition Technologies", and "Wandoujia". We made some manual corrections (most of them changed to the academic type). However, adding such terms to our dictionaries would have been clearly a non-trivial task and we opted to also exclude the remaining records from the dataset. Since they were making only a reduced fraction of the dataset, we believe the results would not have been altered much with or without having those records (if we had to clean up those records).

So, the number of matches is 18,482 ($15,408 + 1,671 + 31 + 1,372$), higher than 18,303, since 179 affiliations were classified with two types. While this situations could have been indeed the case for some individuals, by manual inspections, we found that in some other cases, there were ambiguous cases. An example was the following affiliation: "Philips Res. Laboratories Eindhoven, Building WL p313, Eindhoven University of Technology", due to the matching of the "Univ" (academic) and "Philips" (industry) terms. Each of those cases were manually analysed and for some and them a specific type (academic, industy of government) was chosen. Whenever, it was not possible to disambiguate we left the type of affiliation as "non-classified".

At the end, our dataset has 15,623 academic, (2) 1,806 industrial, and (3) only 56 governmental affiliations. Only 818 of them remained as as non-classified. We opted to discard the authors with governmental affiliations from our study, given their very minor ratio (constituting only 0.32% of the dataset), compared to the other two classifications, and also since our focus in this work has been on IAC papers. The final dataset includes 17,429 affiliations instances associated with 8,804 published papers, which corresponds to a small reduction from the original dataset of 8,968 papers. In the rest of this paper, we use the above classified dataset, for discussions and analysis.

4 Results

In the next sections, we present and discuss the results for each RQ of this study.

4.1 RQ1.1: Ratio of IAC papers to all SE papers

Figure 3 shows, as a Venn diagram, the distributions of academic and industrial papers in the SE papers dataset.

In total, 7,274 papers (82.6% of the dataset) are authored solely by academics. There are 479 papers (5,4% of the dataset) in the dataset that have all their authors affiliated with industrial corporations. The remaining 11,9% of the dataset (1051 papers) have both industrial and academic authors, and thus according to our understanding, are provable results of IACs. We designated the latter set as the IAC-papers, and some of the analysis in the rest of the paper focus on this particular subset of the dataset.

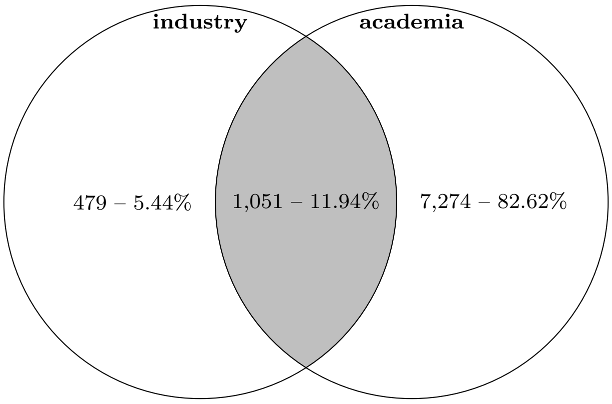

Fig. 3 Distributions of the 8,804 papers in the SE dataset, according to the authors affiliations.

Another important issue to analyse was the distribution of the three classifications of SE papers through time. We show the results of that analysis as a stacked line-chart in Figure 4. The x-axis covers years 1975 to 2022. The graph shows that the overall number of the SE dataset papers has increased. This is similar to what occurred with the timeline of the total number of SE papers (see Figure 2). We have also added the percentage values of each of the three classifications for every five years in the x-axis and for the last year of 2022. Note that these percentage values shall be carefully interpreted as they denote the ratio of each classification for each year only in a cumulative view, and shall not be interpreted across the series.

Furthermore, to get a closer insight into the trend of IAC papers in particular, we show in Figure 5 the volume trend of IAC papers from our SE dataset (black curve). We have also include a linear regression (dashed blue curve).

We discuss the most insightful observations from Figures 2 and 5 in the following:

Papers written solely by academic authors are clearly the majority: It is clear that, in every single year, the majority of published SE papers in the top-5 venues have been by only academic authors. In particular, the

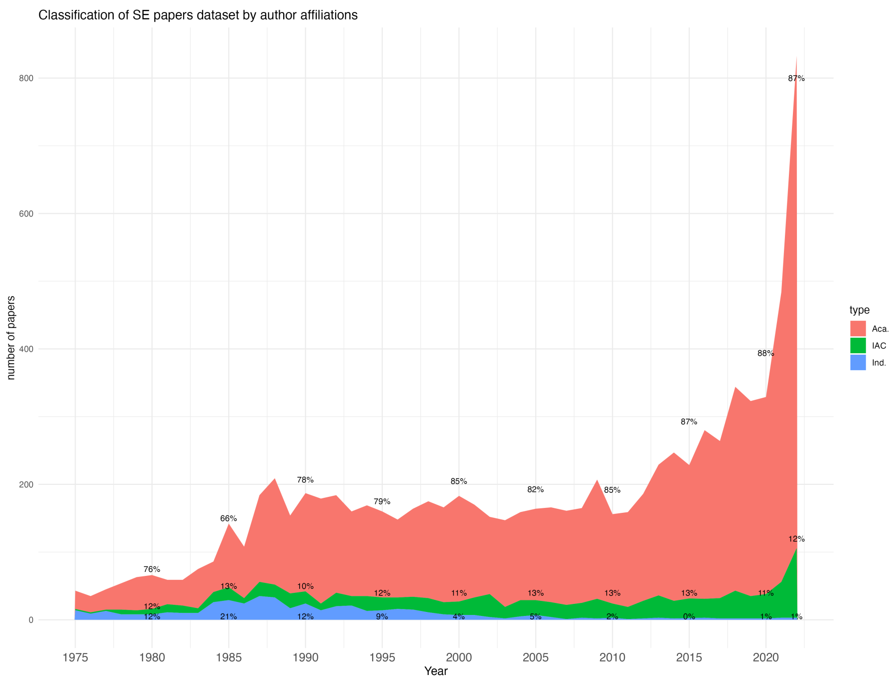

Fig. 4 Breakdown of the sampled SE papers, by their author affiliation classification type (Aca. – papers with all academic authors; IAC – papers with at least one academic and one industry author; Ind. – papers with all industry authors), through time (Stacked line-chart).

purely academic SE paper ratio was around 78% in the 1990–1995 time period and more recently it has increased to around 87% in the period 2019–2022.

Ratio of the papers authored solely by practitioners has declined and has almost reached zero: The ratio and also the number of full industry-authored SE papers have declined, particularly after around year 2007, reaching residual numbers at the end of the considered time period, e.g., there were just three such papers both in 2021 and 2022, or about 0.62% and 0.36 of the papers in each of those two years. This denotes that teams of practitioners, without involvement of researchers, have very little motivation to write and publish papers in the SE research journals. That ratio used to be quite "healthy" in years such as 1984–1986 (about 20–30%), and reasons of such a substantial decrease shall be investigated and addressed by the SE community.

Industry-authored papers versus IAC paper: It is interesting to observe that the ratio of industry-authored papers were significantly higher than IAC papers in the initial years of the bibliometric data (from 1975 to early 1990's). Since 1994, the situation was inverted and the number of IAC papers was always higher then the number of industry-authored papers.

Ratio of the IAC papers has not fluctuated much: The total percentage of IAC papers has reached a stable 10% to 15% range within the period 1994–2022. It was expected that the ratio of the IAC papers would have increased by time, highlighting occurrence of more IAC in the SE community world-wide as more discussions for IACs have been taking places since a few decades ago, but the data are not showing such an increase. However, when we look at the trend in Figure 5, we can see an initial slowly-increasing positive trend, an increase from a handful number of IAC papers in 1970's to about 20-30 IAC papers per year in 1987-2012. In Figure 5, the IAC publication numbers associated with the most recent years of 2021 and 2022 show a substantially different and highly increasing trend. For instance, 103 IAC papers were published in 2022, which deviates much from the overall 0.8× (the publication year) linear trend. Since the IAC recent trend shift is only associated with a two-year period, further IAC publication analyses are needed in the next years, in order to confirm this recent IAC publication pattern.

Number of papers accepted to the top-5 journals are in the rise: This insight is not directly in the scope of our study, but nevertheless we can observe in Figure 2 that the cumulative number of papers accepted each year into the top-5 journals under study has been on the rise, in recent years. This relates to the paper-review and acceptance strategies of each journal and its editorial board.

4.2 RQ1.2: Comparison of the ratio of IAC papers in SE to the ratios in other fields

We compare in Table 2 the ratios of IAC papers in other areas of science that have been presented in the related work (as discussed in Section 2).

Using our approach and the dataset that we compiled in this work, we obtained an overall IAC papers ratio of 9.0% in the analysed SE papers. Comparing this ratio with the two other ratios related to SE in Table 2, we can see that this ratio is twice the number calculated in (Garousi 2015), which also targeted the SE domain but only for the papers originating from Turkey, in which the IAC papers' ratio was only 4.5%. As one can expect and as we discuss in Section 4.4, there are clear differences among different countries with respect to their involvement in IAC papers.

When comparing the IAC papers ratio between our work and the area of GUI testing, we see that our obtained 9.0% value is below the IAC ratio obtained in that field (13.2%). This can possibly be explained as GUI testing, as a subject area under software testing and SE tends to be a rather more applied sub-field, when compared to some other sub-fields of SE, such as formal methods. Therefore, it could be possible that if we were to compare the level of IACs and IAC papers in different sub-fields of SE, more applied sub-fields may have been the subject of more IACs and IAC papers. In fact, RQ2 of our study, to be discussed in Section 4.3, aims at assessing this issue.

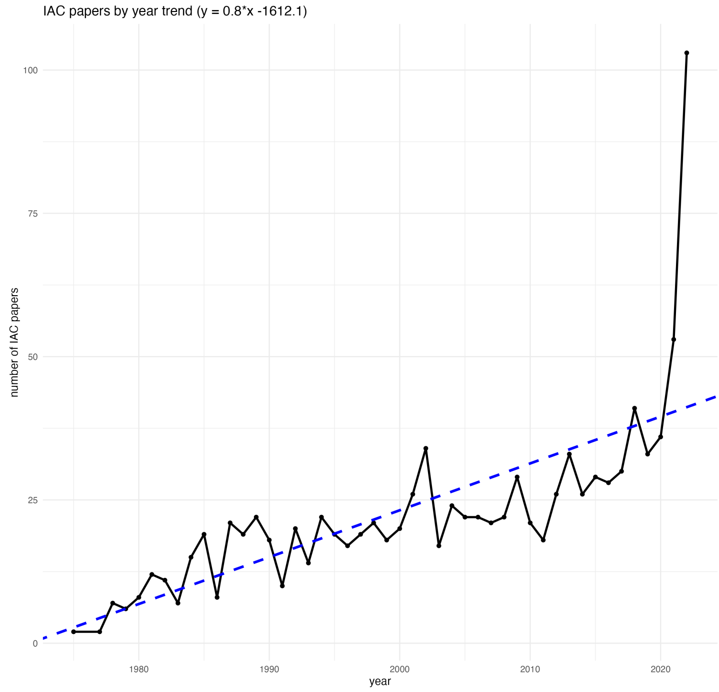

Fig. 5 Number of IAC papers in the dataset and their trend through time.

It is also interesting to compare the IAC papers ratio calculated in our work with the ratios reported in other fields, as listed in Table 2. Those ratios are generally between 2% to about 6%, and we can see that the IAC papers ratio calculated in our work is higher than those other fields or IAC ratios measured for all scientific fields. This could be possibly explained by the fact that SE is often classified among the applied fields of science and thus there are more chances and opportunities for IACs in SE compared to more basic sciences, such as mathematics, etc.

Study	Research field	Region	Period	IAC papers ratio (%)
(Tijssen et al. 2016)	All sciences	World	2009–2012	5.2
(Callaert et al. 2015)	All sciences	USA	2009–2012	6.1
(Callaert et al. 2015)	All sciences	China	2009–2012	2.7
(Abramo et al. 2009)	All sciences	Italy	2001–2003	3.0
(Orduña Malea 2020)	All sciences	Latin America	2009–2012	2.0
(Garousi 2015)	SE	Turkey	1992–2014	4.5
(Banerjee et al. 2013)	GUI testing	World	1991–2011	13.2
This work	SE	World	1975–2020	11.9

Table 2 Ratio of IAC papers presented in different studies.

4.3 RQ2: SE topics covered by IAC papers

To address this RQ, we conducted an automated textual analysis on paper titles and keywords provided in the header of each paper. A suitable method is to derive the word clouds based on those data. We derived two word clouds, one using paper titles and another using keywords, as shown in Figure 6.

Fig. 6 Word-clouds for IAC paper titles (top) and keywords (bottom).

As discussed in Section 3.2, the metadata of each paper provided by Scopus consists of two sets of keywords: author keywords (keywords chosen by the authors), and index keywords, also called index terms, which are a standard set of keywords pre-designed by the publishers (e.g., ACM) that authors can choose from, and mention in their paper header. For obtaining the world clouds in Figure 6, we just use the author keywors, as many index keywords do not appear to include informative terms.

The top word-cloud is based on the titles of all the 1051 IAC papers, while the bottom graph uses the paper keywords. The word clouds were produced by using the `tm` (text mining) package of the widely-used open-source `R` tool (Feinerer et al. 2008). For brevity purposes of the two word clouds, we have set the minimum word frequency to four. Furthermore, the tool we used only works on the syntactic level of terms, so it does distinguish different forms of a single concept, such as"test" and "testing". Thus, we combined such terms into one single form, e.g., by replacing all occurrences of "testing" with the phrase "test".

We can see in Figure 6(top) that model(ling), testing, development, and analysis are among the most frequent terms in terms of titles of the papers. In terms of keywords, Figure 6(bottom) shows that testing, model(ling), analysis, and program(ming) are among the most frequent author keywords, which have some degree of overlap with the most common subjects in the paper titles, as discussed above. Overall, we can infer from the word-clouds that both testing and modelling, followed by analysis, programming, and development, are the topics that have been the most active areas of IAC, as published in the selected journals in the 48-year period under study (1975–2022).

4.4 RQ3: Most active countries measured by the number of IAC papers

For our last RQ, we analysed the set of IAC papers according to the countries of the author affiliations. In total, 65 countries are involved in the IAC pappers that are included in the dataset. Table 3 presents the countries that appear in 10 or more IAC papers.

The top-5 countries in the list are: United States, Canada, Germany, United Kingdom, and China. The other countries are mostly from Europe, plus a few countries from other parts of the world, especially from Asia. The number of IAC papers can be easily normalized, if needed, by the population of each country, to get a more accurate rank of the countries.

Luxembourg, a small country with less that 1M inhabitants, deserves to be cited as it appears in 9 IAC-papers. All IAC-papers with authors based in Luxembourg were published in the period 2012–2022.

We note that these countries are all among the top-10 countries in the world with the highest GDP, and thus, one would expect that the level of R&D expenditure and also enablers for IACs in those countries are quite stronger than in other countries. The case of China is also emerging. In fact, China has three IAC-papers in the period 1975–2001 and 68 in the period 2009–2022. In particular, 30 IAC papers with authors from China were published in 2022. So, it seems that China will reach briefly (i.e., in a short number of years) a position in the top-3.

Country	# papers	Country	# papers
United States	592	Spain	34
Canada	155	India	34
Germany	119	France	29
United Kingdom	109	Singapore	29
China	72	Switzerland	29
Sweden	66	South Korea	27
Italy	50	Brazil	26
Netherlands	49	Austria	21
Australia	40	Israel	15
Finland	37	Ireland	13
Japan	37	Turkey	13
Norway	34	Taiwan	12

Table 3 Countries with 10+ IAC papers.

5 Discussions, implications and limitations

5.1 Discussions and implications

The goal of this work was to characterize the state of IACs in SE in a worldwide scale, from the lens of the author affiliations in papers published in five top SE research journals, venues which are also often selected in other bibliometric studies, e.g., (Wainer et al. 2013; Fernandes and Monteiro 2017).

We focused our analysis on characterising the annual ratio of IAC papers with respect to all SE, the ratio of IAC papers in SE compared to that ratio in other fields, the SE topics covered by IAC papers, and finally the most active countries by the number of IAC papers.

Our analysis shows that there is a slow growth in the number of IAC papers, but the percentage of IAC papers in each year has been around 10%-15% in the period 1994–2022. These results indicate that academic authors still vastly dominate the SE authorship landscape, compared to authors from the SE industry. In fact, 94% of the SE papers have academic authors (85% are exclusively authored by academic authors).

In our opinion, increasing the number (and the percentage) of IAC papers is necessary to make research in SE more relevant and impactful to the industry.

The SE topics most represented in IAC-papers are testing, modelling, analysis, programming, and development. Other important SE topics which seem to have been largely under-represented in IAC papers and projects are software requirements, maintenance, and software management, and thus further focus on those topics in upcoming IACs is encouraged. In terms of the countries, the results show that the United States is the leading country in terms of IAC papers. Canada, Germany, and United Kingdom, which are countries with strong and prestigious research-oriented universities, are also well ranked. China is the fastest growing country in the list, with the majority of their IAC papers published very recently (since 2016).

5.2 Limitations and potential threats to validity

As it is the case of any empirical and bibliometric study, we are aware of the limitations and potential threats to validity of our work. In this section, we discuss the potential threats to the validity of the sturdy in the context of the four types of threats to validity based on a standard checklist (Wohlin et al. 2012). We also discuss the steps that we have taken to minimize or mitigate those potential threats.

Internal validity: Internal validity reflects the extent to which a causal conclusion based on a study is warranted (Wohlin et al. 2012). To address this issue, we discuss in Section 3 our systematic approach for the selection of publication database, SE journals and also our method for classification of papers w.r.t. author affiliation data. In order to ensure transparency and replicability of our analysis, we have provided the tool and all the dataset in an online GitHub repository.

Construct validity: Construct validity is the degree to the scales, metrics and instruments used in a study actually measure the properties they are supposed to measure (Ralph and Tempero 2018). The metrics that we used in this study were rather simple and are related to the number and ratios of three classifications of SE papers: those written by academics, those written by practitioners, and papers authored within IACs. Thus, they indeed measure what we intended to measure.

Conclusion validity: Conclusion validity of a study deals with whether correct conclusions are reached through rigorous and repeatable treatments (Wohlin et al. 2012). Analysis and conclusions that we discuss throughout the paper are based on quantitative measures and statistics on the data extracted from the papers dataset. By following the systematic approach that we have developed for this purpose, if the study is conducted by other researchers, it is expected that results will not have major deviations from our results.

External validity: External validity is concerned with to what extent the results of this secondary study can be generalized (Wohlin et al. 2012). A limitation w.r.t external validity of our work lies in our selection of papers published in top-5 SE journals instead of all 165k SE papers published in the area (as discussed in Section 3.3). Although our tool is an automated tool, the results need to often be validated manually. if we had to run the classification tool on all SE papers, manual inspection of the classification results, generated by the tool, could have been highly effort-intensive. As was discussed, sampling a subset of all SE papers is a common practice in bibliometric studies, e.g.,(Glass 1995; Wong et al. 2021). Additionally, the top-5 SE journals that were selected have been also selected in previous bibliometric studies, e.g., (Wainer et al. 2013; Fernandes and Monteiro 2017).

6 Conclusions and future works

This paper provides a systematic bibliometric study on the patterns of co-authorship of scientific publications in five top research-oriented journals. We retrieved a large bibliometric dataset, covering a 48-year period (from 1975 to 2022) and including almost 9,000 scientific articles related with more than 18,000 affiliations. Our findings indicate that the percentage of IAC papers in each year is around 10-15% in the period 1994–2022. This value indicates that academic authors dominate the authorship landscape in the SE field. In fact, 94% of the SE papers have academic authors (83% are exclusively authored by academic authors). The SE topics most represented in IAC-papers are testing, modelling, analysis, programming, and development. Finally in terms of the countries, the results show that the United states, Canada, Germany, United Kingdom, and China are leading in terms of IAC papers. China is the fastest growing country in the list, with the majority of their IAC papers published very recently (since 2016).

In future work, we intend to perform a similar bibliometric study by analysing the other Computer Science sub-fields (e.g., Artificial Intelligence, Computer Engineering), in order to compare if there are different IAC publication patterns when compared with the SE domain. Moreover, it would be interesting to further analyse certain aspects of IAC, in order to identify factors influencing more productive IAC publications. This could be achieved by further characterizing the selected SE papers (e.g., identification of industry sectors, domains and/or application areas) and by performing multidimensional analyses, e.g., usage of machine learning algorithms to scan a given paper full-text to determine if it is an outcome of an IAC.

Author contributions

J. M. Fernandes performed the conceptualization, methodology, software, investigation, resources, formal analysis, writing – original draft, writing - review and editing. P. Cortez contributed with methodology, software, investigation, validation, formal analysis, writing – review and editing, and visualization. V. Garousi contributed in refining the methodology, validation, writing – review and editing.

Acknowledgements

This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the R&D Units Project Scope: UIDB/00319/2020.

References

Abramo G, D'Angelo C, Di Costa F, Solazzi M (2009) University-industry collaboration in Italy: A bibliometric examination. Technovation 29(6–7):498–507, DOI 10.1016/j.technovation.2008.11.003

Alexandru CV, Panichella S, Proksch S, Gall HC (2019) Redundancy-free analysis of multi-revision software artifacts. Empirical Software Engineering 24(1):332–380, DOI 10.1007/s10664-018-9630-9

Aref S, Friggens D, Hendy S (2018) Analysing scientific collaborations of New Zealand institutions using Scopus bibliometric data. In: Australasian Computer Science Week Multiconference (ACSW 2018), DOI 10.1145/3167918.3167920

Banerjee I, Nguyen B, Garousi V, Memon A (2013) Graphical user interface (GUI) testing: Systematic mapping and repository. Information and Software Technology 55(10):1679–1694, DOI 10.1016/j.infsof.2013.03.004

Briand L (2012) Embracing the engineering side of software engineering. IEEE Software 29(4):96–96, DOI 10.1109/MS.2012.86

Briand L, Bianculli D, Nejati S, Pastore F, Sabetzadeh M (2017) The case for context-driven software engineering research: Generalizability is overrated. IEEE Software 34(5):72–75, DOI 10.1109/MS.2017.3571562

Butcher J, Jeffrey P (2005) The use of bibliometric indicators to explore industry–academia collaboration trends over time in the field of membrane use for water treatment. Technovation 25(11):1273—-1280, DOI 10.1016/j.technovation.2004.06.003

Callaert J, Landoni P, Van Looy B, Verganti R (2015) Scientific yield from collaboration with industry: The relevance of researchers' strategic approaches. Research Policy 44(4):990–998, DOI 10.1016/j.respol.2015.02.003

Carver JC, Prikladnicki R (2018) Industry–academia collaboration in software engineering. IEEE Software 35(5):120–124, DOI 10.1109/MS.2018.3571250

Developer Nation (2021) State of the Developer Nation, 20th edition. https://wwwdevelopernationnet/developer-reports/de20 Last accessed: Mar. 08, 2023

Ebert C (2018) 50 years of software engineering: Progress and perils. IEEE Software 35(5):94–101, DOI 10.1109/MS.2018.3571228

Farhoodi R, Garousi V, Pfahl D, Sillito J (2013) Development of scientific software: A systematic mapping, a bibliometrics study, and a paper repository. International Journal of Software Engineering and Knowledge Engineering 23(4):463–506, DOI 10.1142/S0218194013500137

Feinerer I, Hornik K, Meyer D (2008) Text mining infrastructure in R. Journal of Statistical Software 25(5):1–54, URL `https://www.jstatsoft.org/v25/i05/`

Fernandes JM (2014) Authorship trends in software engineering. Scientometrics 101(1):257–271, DOI 10.1007/s11192-014-1331-6

Fernandes JM, Monteiro MP (2017) Evolution in the number of authors of computer science publications. Scientometrics 110(2):529–539, DOI 10.1007/s11192-016-2214-9

Freitas FG, Souza JT (2011) Ten years of search based software engineering: A bibliometric analysis. In: International Symposium on Search Based Software Engineering (SSBSE 2011), Springer, pp 18–32, DOI 10.1007/978-3-642-23716-4_5

Garcia R, Araújo V, Mascarini S, Santos E, Costa A (2020) How long-term university-industry collaboration shapes the academic productivity of research groups. Innovation 22(1):56–70, DOI 10.1080/14479338.2019.1632711

Garousi V (2015) A bibliometric analysis of the turkish software engineering research community. Scientometrics 105(1):23–49, DOI 10.1007/s11192-015-1663-x

Garousi V, Fernandes JM (2017) Quantity versus impact of software engineering papers: A quantitative study. Scientometrics 112(2):963–1006, DOI 10.1007/s11192-017-2419-6

Garousi V, Mäntylä MV (2016) Citations, research topics and active countries in software engineering: A bibliometrics study. Computer Science Review 19:56–77, DOI 10.1016/j.cosrev.2015.12.002

Garousi V, Ruhe G (2013) A bibliometric/geographic assessment of 40 years of software engineering research (1969–2009). International Journal of Software Engineering and

Knowledge Engineering 23(9):1343–1366, DOI 10.1142/S0218194013500423
Garousi V, Petersen K, Ozkan B (2016) Challenges and best practices in industry-academia collaborations in software engineering: A systematic literature review. Information and Software Technology 79:106–127, DOI 10.1016/j.infsof.2016.07.006
Garousi V, Pfahl D, Fernandes JM, Felderer M, Mäntylä MV, Shepherd D, Arcuri A, Coşkunçay A, Tekinerdogan B (2019) Characterizing industry-academia collaborations in software engineering: evidence from 101 projects. Empirical Software Engineering 24(4):2540–2602, DOI 10.1007/s10664-019-09711-y
Garousi V, Borg M, Oivo M (2020) Practical relevance of software engineering research: Synthesizing the community's voice. Empirical Software Engineering 25(3):1687–1754, DOI 10.1007/s10664-020-09803-0
Garrison HH, Herman SS, Lipton JA (1992) Collaborative relationships in dental materials research measuring the volume and outcomes. Evaluation Review 16(2):184–197, DOI 10.1177/0193841X9201600205
Gielfi GG, Furtado AT, Campos A, Tijssen R (2017) University-industry research collaboration in the Brazilian oil industry: the case of Petrobras. Revista Brasileira de Inovação 16(2):325–350
Glass RL (1994) The software-research crisis. IEEE Software 11(6):42–47, DOI 10.1109/52.329400
Glass RL (1995) Editor's corner an assessment of systems and software engineering scholars and institutions, 1993 and 1994. Journal of Systems and Software 31(1):3–6, DOI 10.1016/0164-1212(95)00058-9
Huang MH, Yang HW, Chen DZ (2015) Industry–academia collaboration in fuel cells: A perspective from paper and patent analysis. Scientometrics 105(2):1301–1318, DOI 10.1007/s11192-015-1748-6
Laukkanen E, Paasivaara M, Itkonen J, Lassenius C (2018) Comparison of release engineering practices in a large mature company and a startup. Empirical Software Engineering 23(6):3535–3577, DOI 10.1007/s10664-018-9616-7
Orduña Malea E (2020) Do Latin American universities engage industry in the scientific publication?: A bibliometrics approach through Scopus. Palabra Clave 10(1):e100, DOI 10.24215/18539912e100
Parnas DL (1998) Successful software engineering research. ACM SIGSOFT Software Engineering Notes 23(3):64–68, DOI 10.1145/279437.279464
Ralph P, Tempero E (2018) Construct validity in software engineering research and software metrics. In: 22nd International Conference on Evaluation and Assessment in Software Engineering (EASE 2018), pp 13–23, DOI 10.1145/3210459.3210461
Runeson P (2012) It takes two to tango – An experience report on industry – Academia collaboration. In: 5th IEEE International Conference on Software Testing, Verification and Validation (ICST 2012), IEEE, pp 872–877, DOI 10.1109/ICST.2012.190
Sun Y, Negishi M, Nishizawa M (2007) Coauthorship linkages between universities and industry in Japan. Research Evaluation 16(4):299–309, DOI 10.3152/095820207X263619
Tijssen RJW, Yegros-Yegros A, Winnink JJ (2016) University–industry R&D linkage metrics: validity and applicability in world university rankings. Scientometrics 109(2):677–696, DOI 10.1007/s11192-016-2098-8
van Solingen R, Basili V, Caldiera G, Rombach HD (2002) Goal Question Metric (GQM) Approach, John Wiley & Sons. DOI 10.1002/0471028959.sof142
Wainer J, Eckmann M, Goldenstein S, Rocha A (2013) How productivity and impact differ across computer science subareas. Communication of the ACM 56(8):67–73, DOI 10.1145/2492007.2492026
Wohlin C (2013) Software engineering research under the lamppost. In: International Joint Conference on Software Technologies, pp IS–11
Wohlin C, Runeson P, Höst M, Ohlsson MC, Regnell B, Wesslén A (2012) Experimentation in software engineering. Springer
Wong WE, Tse TH, Glass RL, Basili VR, Chen TY (2008) An assessment of systems and software engineering scholars and institutions (2001-2005). Journal of Systems and Software 81(6):1059–1062, DOI 10.1016/j.jss.2007.09.018
Wong WE, Tse TH, Glass RL, Basili VR, Chen TY (2009) An assessment of systems and software engineering scholars and institutions (2002-2006). Journal of Systems and Soft-

ware 82(8):1370–1373, DOI 10.1016/j.jss.2009.06.018

Wong WE, Mittas N, Arvanitou EM, Li Y (2021) A bibliometric assessment of software engineering themes, scholars and institutions (2013–2020). Journal of Systems and Software p 111029, DOI 10.1016/j.jss.2021.111029

Zhou P, Tijssen R, Leydesdorff L (2016) University-industry collaboration in China and the USA: A bibliometric comparison. PLoS ONE 11(11):e0165277, DOI 10.1371/journal.pone.0165277