Investigating the Relation between Authors' Academic Age and their Citations

Rand Alchokr[1][0000−0003−0112−5430], Sanket Vikas Joshi[1], Gunter Saake[1][0000−0001−9576−8474], Thomas Leich[2][0000−0001−9580−7728], and Jacob Krüger[3][0000−0002−0283−248X]

[1] Otto-von-Guericke University, Magdeburg, Germany
{rand.alchokr,Sanket.Joshi,saake}@ovgu.de
[2] Harz University & METOP GmbH, Wernigerode, Germany
tleich@hs-harz.de
[3] Eindhoven University of Technology, Eindhoven, The Netherlands
j.kruger@tue.nl

Abstract. The increasing number of authors and consequent publications in computer science can cause some pitfalls, such as understanding the use and fairness of quality indicators for assessing research. In this preliminary work, we aim to examine whether there is a correlation between the citation count and the number of authors contributing to a paper as well as their academic ages. Additionally, we shed light on highly cited papers and compare their authors. For this purpose, we investigate authors' characteristics by conducting data analyses based on a dataset of four prestigious software-engineering-related conferences comprising 5,143 papers and their authors. Our results indicate that the number of authors does not connect to the citation count, but the current academic age of the authors does. We also found that 98 % of the highly cited main-track papers had a contribution from at least one senior researcher, whereas none of these papers was written by a junior researcher alone. These first results are a step towards more in-depth research concerning the fair evaluation of computer-science researchers—specifically regarding juniors and their inclusion.

Keywords: Software engineering · Publications · Scientific collaboration · Junior researchers

1 Introduction

In recent years, computer science has undergone rapid evolution, with a notable shift from solitary to collaborative efforts [3,7,10]. Working in teams is generally thought of as a way to benefit from the experiences of researchers from different disciplines, thereby improving knowledge sharing and easing access to resources [4,15,19]. However, the growing trend of scientific collaboration has also raised concerns regarding research assessments. While researchers are increasingly working in teams to publish papers, there is often a lack of transparency regarding their individual contributions, which challenges a fair evaluation. Some

journals require disclosure of each researcher's unique contributions, but there is currently no standardized framework to precisely measure and assess the contributions of individual authors [22].

Typically, researchers are distinguished based on their expertise, for instance, as junior, mid-level, or senior researchers. However, there is an issue, since quality indicators used for evaluating researchers are the same regardless of their career stage. This has raised questions about the fairness and impact of such indicators on different groups of researchers. Generally, a researcher is assessed based on different measures, the most famous being the *citation count*, which ranks researchers according to their citations, besides other metrics like the *h-index*, *G-index*, or *W-index* [12,14,27] and the rather new *Altmetrics* [23–25]. Analyzing what factors impact such metrics is essential to derive fairer assessments of individual researchers, for instance, for funding agencies and tenure committees [22].

Moving into this direction, *our goal in this paper is to examine whether there are connections between a publication's citation count and the number of authors or their academic ages*. We choose the citation count as a popular assessment method and concentrate on the two variables pertaining to authors' characteristics that may impact citation counts. To the best of our knowledge, researchers' academic age has not been analyzed in depth before. So, we report an analysis on the relationships between these variables and look to find patterns or trends favoring a specific group of researchers if such exist. We defined the following two research questions (RQs) and answer them using a dataset of main-track papers and the corresponding authors' information of four reputable conferences, namely the 1) International Conference on Automated Software Engineering Conference (ASE); 2) Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE); 3) International Conference on Software Engineering (ICSE); and 4) Joint Conference on Digital Libraries (JCDL), that we extracted from dblp:[4]

RQ_1 Does the number of authors contributing to a paper affect its citation count?
RQ_2 Is there a correlation between authors' academic age and the citation count?

Our work is an initial step for a more comprehensive analysis to identify and develop fair metrics and quality indicators to evaluate researchers.

2 Background

There have been various proposals for how to assess research. For instance, Hirsch [14] suggested the h-index, which attempts to calculate a researcher's output and influence over time using the number of papers receiving citations. However, this metric has limitations, since it assigns equal importance to all papers and ignores their age [26]. To address this limitation to some extent, variants of the h-index have been proposed, like the contemporary or trendy h-index—which consider a paper's publication year or age [26]. The citation count is the most widely used metric to assess researchers, and has often been relevant

[4] https://dblp.uni-trier.de/

Table 1: Overview of our dataset.

conference	period	# unique papers	# unique authors
ASE	1991–2020	1,070	2,465
ESEC/FSE	1987–2020	1,193	2,530
ICSE	1976–2020	2,300	4,357
JCDL	2001–2020	580	1,390
total		5,143	8,730

in career decisions [18, 20]. However, in a collaborative scientific environment in which researchers combine their knowledge and contribute to muti-author papers, crediting the authors fairly becomes a challenging task that needs to be tackled [3, 15, 21]. Multiple studies emphasize the role of collaboration specifically with highly cited scholars, as it benefits early career researchers to gain experience and improve their careers [7, 9]. This opinion is also shared among early career researchers themselves [4, 20]. Nonetheless, different opinions exist when it comes to citations and collaborative papers. A study shows that an increase in the number of co-authors has a definite impact on productivity in terms of the number of papers published [16], but this does not always mean more citations [2, 17]. However, a contradicting result indicates that co-authored publications receive more citations because collaboration improves the transfer and synthesis of knowledge [1]. Noteworthy, we found that the assessment metrics reported in such studies are the same for all researchers, albeit their expertise or academic age. Academic age is an important characteristic that we have investigated more deeply in recent research [3–6]. Yet, the impact of ignoring such characteristics when utilizing metrics or the correlations between these characteristics and the citation count is unknown.

3 Methodology

We performed a retrospective study in which we examined a dataset extracted from dblp based on our RQs. Note that we studied papers published at conferences, because computer science focuses more on conferences than journals [11]. Namely, we examined the research tracks of three major software-engineering conferences (ASE, 1991; ESEC/FSE, 1987; ICSE, 1976) and one software-related conference (JCDL, 2001). We gathered all papers since the first edition (years in previous hyphens) of each conference until 2020 by crawling paper and author data from dblp using Python scripts. Note that each author has a website on dblp that acts as an identifier to distinguish authors with the same name. For validation, we compared the data manually against the official data from the ACM.[5] If we could not clearly identify research-track papers due to missing data, especially for older conferences, we used a proxy by excluding papers with fewer than seven pages. In Table 1, we summarize the number of unique authors and papers for all four conferences. Since dblp does not provide citation counts, we fetched this data from Scopus—a permissive citation database by

[5] https://dl.acm.org/

Elsevier.[6] Note that the total number of unique authors (8,730) is not the sum of the last column, since we counted each author only once across all conferences. For a more robust analysis, we used a subset of our dataset comprising highly-cited papers only (≥ 25 citations) and papers with a publishing year until 2010. Furthermore, we divided the resulting dataset (808 papers) into the following subsets: (i) PRE-2000 (114); (ii) POST-2000 (694); and (iii) PRE- and POST-2000 combined (808). We used regression analysis and in-depth data exploration via Python and KNIME [8] to analyze our data.

Academic age is the number of years an author has actively published until a particular paper, and must be calculated individually for each researcher based on a paper's publication year and the author's first paper's publication year (not restricted to the four conferences). For the *current academic age*, we replace the paper publication year with the author's last paper publication year:

$$Age_{academic} = Year_{paper} - Year_{firstPaper} + 1 \quad (1)$$

To distinguish researchers based on their academic age, we classified them as: *Juniors* (academic age ≤ 3) have up to three years of academic experience and only recently started working in research [18]. For *mid-level* researchers ($3 <$ academic age ≤ 15), we used the upper limit of 15, since we identified it as the "Golden Age" of software-engineering researchers [6]. Lastly, we labeled researchers with an academic age above 15 years as *seniors* (academic age > 15).

4 RQ_1: Number of Authors and Citations

First, we investigated whether more authors contributing to a paper impacts the citation count of that paper. This direction is inspired by research on public health indicating that the number of citations decreases as the number of authors increases [2]. Using qualitative data analysis, we checked for connections between the *number of authors* and the *citation count* across all three datasets. For all three, we found that most papers have been written by a team of two, three, or four authors. The paper with the highest number of citations (more than 1,000) was authored by three researchers. Moreover, we observe that the number of citations does not increase with the number of authors. So, we conclude that having more authors seems to have no bearing on citations, despite more authors likely increasing the visibility and dissemination of the paper—since it is exposed to more networks and personal contacts While we require further research in this direction, it seems that other factors like the topic of the paper and its quality or the author's reputation and academic age may be more important [18, 20].

5 RQ_2: Academic Age and Citations

Regarding the age of a paper's authors, we inevstigated the **null hypothesis:** "The authors' ages (current author age, academic age) do not impact the citation

[6] https://www.scopus.com

Table 2: Coefficients and Statistics.

data	variable	coefficient	std. error	p-values	R^2	adjusted R^2
PRE-2000	Current author age	0.398	0.845	0.637	0.005	-0.004
	Academic age	-0.456	1.557	0.769	0.005	-0.004
POST-2000	Current author age	1.507	0.445	0.001	0.005	0.004
	Academic age	-2.028	0.592	0.001	0.005	0.004
PRE-POST	Current author age	0.863	0.357	0.016	0.002	0.001
	Academic age	-1.235	0.502	0.014	0.002	0.001

count." For this purpose, we first used statistical inference (regression modeling) to determine what kind of relationship exists between the dependent *citation count* and the independent variables *current author age* as well as *academic age* [13]. The null hypothesis generally rejects the theory that independent variables do not impact the result, while the alternate hypothesis is precisely the opposite. In Table 2, we can see the values for the coefficients for the *current author age* and *academic age*. For the **PRE-2000** dataset, we see positive coefficients for the *current author age*, which means that a positive effect on the citation count exists. The *academic age* has a negative coefficient, which means that the citation count decreases as the academic age increases. However, the academic age has a higher error rate than the current author's age. Seeing the significance values, we cannot make a strong inference from these findings, since the values suggest that the dataset is too small to draw any conclusions. R-squared or R^2 is a metric that measures the proportion of the dependent variable's variance that the independent variables account for collectively. According to its value, only 0.5 % of citation-count fluctuations can be attributed to our dependent variables. Adjusted-R^2 is a more accurate version of R^2. The R^2 value increases as we add the independent variables, but the adjusted-R^2 value increases only when the independent variable strongly influences the dependent variable. A negative value signifies that the impact of the independent variables is very low on the dependent variable for this dataset, at the least.

In the two other datasets, we had more data. Interestingly, we found that the explanatory variables *current author age* and *academic age* influenced the citation count. The former positively, and the latter negatively. The p-values are less than the significance level of 0.05. Thus, we could partly reject the null hypothesis, for one variable, which eventually means our alternative hypothesis partly holds in this scenario. So, the age features seem to impact the citation count. The R^2, and adjusted-R^2 values are positive, because the explanatory variable impact the response variable.

To further explore the data, we compared the three categories of researchers as illustrated in Figure 1. According to (a), the seniors' percentage surpasses other researchers with a gradual increase for juniors. In (b), we can see that seniors comprise the highest percentage of first authors in multi-authored papers. After investigating our datasets deeply, we found that juniors are on average third authors, whereas mid-levels and seniors are on average second authors. Via (c), we mainly checked how many papers researchers have written without collaboration across different groups (e.g., seniors with juniors) We can see

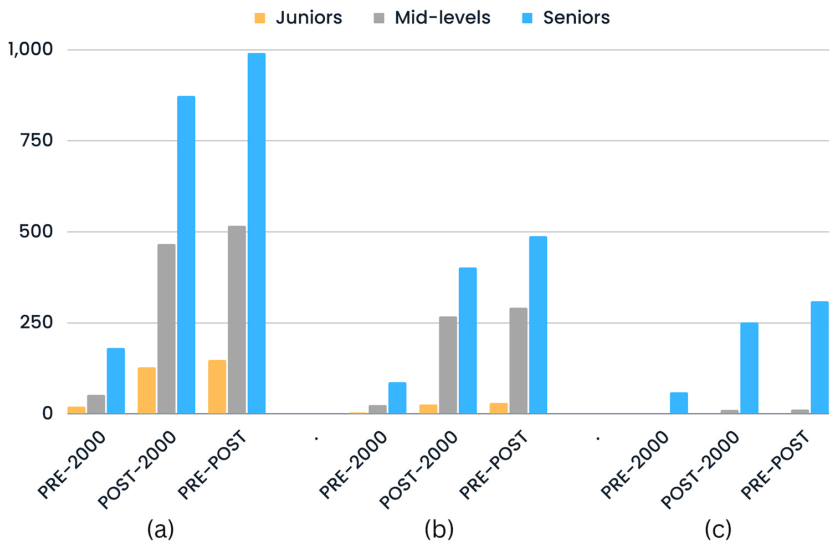

Fig. 1: In-depth data exploration. (a): Researchers contributing to papers with ≥25 citations; (b): Papers with ≥citations & first author from a specific group; (c): Papers with citations ≥25 & single-authored by a specific group.

that for **PRE-2000** no paper was authored exclusively by one or more junior researchers, whereas senior researchers wrote 57 papers alone, compared to one paper written by mid-level researchers only. Moving to **POST-2000**, again none of the junior researchers wrote a publication alone. Mid-level researchers published 11 single-author papers, while seniors wrote 250 papers. Consequently, in the **PRE and POST-2000** dataset, there are also no papers written solely by juniors. These insights suggest that for a paper to be cited frequently, it must have a contribution from a senior researcher.

6 Conclusion

In this paper, we have reported an initial analysis of the relationship between citation count and two features (number of researchers contributing and their academic age) using data of 808 highly-cited papers from the software-engineering community. Overall, the results indicate that the number of authors contributing does not relate to a paper being cited highly, but that the current age of authors is an influential factor. We also found that 98 % of these highly-cited papers had a contribution from at least one senior researcher with around 60 % as first authors and no paper written solely by a junior researchers. Therefore, our results indicate that comparing two groups of researchers based on citation-related indicators is unfair because it is highly influenced by the *age* factor. Consequently, we believe that researchers should be compared based on their actual contribution and there is a need for a consistent framework with which the contribution of every researcher can be determined. The results also emphasize the role of collaboration in helping early-career researchers.

References

1. Adams, J.: Collaborations: The Rise of Research Networks. Nature **490** (2012)
2. Ahmed, A., Mastura, A., Ghafar, N.A., Muhammad, M., Ebrahim, N.A.: Impact of Article Page Count and Number of Authors on Citations in Disability Related Fields: A Systematic Review Article. Iranian Journal of Public Health **45**(9) (2016)
3. Alchokr, R., Krüger, J., Shakeel, Y., Saake, G., Leich, T.: A Closer Look into Collaborative Publishing at Software-Engineering Conferences. In: International Conference on Theory and Practice of Digital Libraries (TPDL). Springer (2022)
4. Alchokr, R., Krüger, J., Shakeel, Y., Saake, G., Leich, T.: Peer-Reviewing and Submission Dynamics Around Top Software-Engineering Venues: A Juniors' Perspective. In: International Conference on Evaluation and Assessment in Software Engineering (EASE). ACM (2022)
5. Alchokr, R., Krüger, J., Shakeel, Y., Saake, G., Leich, T.: Understanding the Contributions of Junior Researchers at Software-Engineering Conferences. In: Joint Conference on Digital Libraries (JCDL). IEEE (2021)
6. Alchokr, R., Krüger, J., Shakeel, Y., Saake, G., Leich, T.: On Academic Age Aspects and Discovering the Golden Age in Software Engineering. In: International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE). ACM (2022)
7. Aubert Bonn, N., Pinxten, W.: Advancing Science or Advancing Careers? Researchers' Opinions on Success Indicators. PLOS One **16** (2021)
8. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz Information Miner. In: Data Analysis, Machine Learning and Applications. Springer (2008)
9. van den Besselaar, P., Sandström, U.: Measuring Researcher Independence using Bibliometric Data: A Proposal for a New Performance Indicator. PLOS One **14** (2019)
10. Bukvova, H.: Studying Research Collaboration: A Literature Review. All Sprouts Content (2010)
11. Chen, J., Konstan, J.A.: Conference Paper Selectivity and Impact. Communications of the ACM **53**(6) (2010)
12. Egghe, L.: An Improvement of the h-Index: The g-Index. ISSI Newsletter **2**(1) (2006)
13. Freedman, D.A.: Statistical Models: Theory and Practice. Cambridge University Press (2009)
14. Hirsch, J.E.: An Index to Quantify an Individual's Scientific Research Output. Proceedings of the National Academy of Sciences **102**(46) (2005)
15. Katz, J.S., Martin, B.R.: What is Research Collaboration? Research Policy **26**(1) (1997)
16. Lee, S., Bozeman, B.: The Impact of Research Collaboration on Scientific Productivity. Social Studies of Science **35**(5) (2005)
17. Levitt, J., Thelwall, M.: Long Term Productivity and Collaboration in Information Science. Scientometrics **108** (2016)
18. Li, W., Aste, T., Caccioli, F., Livan, G.: Early Coauthorship with Top Scientists Predicts Success in Academic Careers. Nature Communications **10**(1) (2019)
19. Melin, G.: Pragmatism and Self-Organization: Research Collaboration on the Individual Level. Research Policy **29**(1) (2000)
20. Nicholas, D., Herman, E., Xu, J., Boukacem-Zeghmouri, C., Abrizah, A., Watkinson, A., Świgoń, M., Rodriguez-Bravo, B.: Early Career Researchers' Quest for Reputation in the Digital Age. Journal of Scholarly Publishing **49** (2018)

21. Qi, M., Zeng, A., Li, M., Fan, Y., Di, Z.: Standing on the Shoulders of Giants: The Effect of Outstanding Scientists on Young Collaborators' Careers. Scientometrics **111** (2017)
22. Sauermann, H., Haeussler, C.: Authorship and Contribution Disclosures. Science Advances **3**(11) (2017)
23. Shakeel, Y., Alchokr, R., Krüger, J., Leich, T., Saake, G.: Are Altmetrics Useful for Assessing Scientific Impact? A Survey. In: International Conference on Management of Digital EcoSystems (MEDES). ACM (2022)
24. Shakeel, Y., Alchokr, R., Krüger, J., Saake, G., Leich, T.: Altmetrics and Citation Counts: An Empirical Analysis of the Computer Science Domain. In: Joint Conference on Digital Libraries (JCDL). IEEE (2022)
25. Shakeel, Y., Alchokr, R., Krüger, J., Saake, G., Leich, T.: Are Altmetrics Proxies or Complements to Citations for Assessing Impact in Computer Science? In: Joint Conference on Digital Libraries (JCDL). IEEE (2022)
26. Sidiropoulos, A., Katsaros, D., Manolopoulos, Y.: Generalized Hirsch h-Index for Disclosing Latent Facts in Citation Networks. Scientometrics **72** (2007)
27. Wu, Q.: The w-Index: A Significant Improvement of the h-Index. arXiv preprint arXiv:0805.4650 (2008)