12-2013

Factors influencing research contributions and researcher interactions in software engineering: An empirical study

Subhajit DATTA
Singapore Management University, subhajitd@smu.edu.sg

A. S. M. Sajeev
University of New England

Santonu SARKAR
Infosys Technologies Ltd India

Nishant KUMAR
Appnomic Systems

Published in 2013 20th Asia-Pacific Software Engineering Conference (APSEC): Bangkok, December 2-5: Proceedings, pp. 34-41.
https://doi.org/10.1109/APSEC.2013.16

Factors Influencing Research Contributions and Researcher Interactions in Software Engineering: An Empirical Study

Subhajit Datta*, A. S. M. Sajeev†, Santonu Sarkar‡, Nishant Kumar§
*Singapore University of Technology & Design, Singapore, Email: subhajit.datta@acm.org
†University of New England, Australia, Email: sajeev@une.edu.au
‡Infosys Labs, Bangalore, India, Email: santonus@gmail.com
§Appnomic Systems, Bangalore, India, Email: kumar.nishant1@gmail.com

***Abstract*—Research into software engineering (SE) education is largely concentrated on teaching and learning issues in coursework programs. This paper, in contrast, provides a meta-analysis of research publications in software engineering to help with *research education* in SE. Studying publication patterns in a discipline will assist research students and supervisors gain a deeper understanding of how successful research has occurred in the discipline. We present results from a large scale empirical study covering over three and half decades of software engineering research publications. We identify how different factors of publishing relate to the number of papers published as well as citations received for a researcher, and how the most successful researchers collaborate and co-cite one another. Our results show that authors with high publication rates do not concentrate on a few selected venues to publish, researchers with high publication rates behave differently from researchers of high citation rates (with the latter group co-authoring and citing their peers to a much lesser extent than the former), and collaborators citing each other's works is not a significant phenomenon in SE research.**

I. Introduction

Software Engineering (SE) education as a field of research is involved in developing techniques and methods to help improve the education of software engineers. Generally publications in this field involve a wide variety of topics varying from different models of teaching SE [1], [2], [3] to the role of other disciplines in SE education [4] to tools and methodologies for use in classrooms [5], [6], [7] among others. However, almost all of these topics are on coursework education. *Research* education in SE, on the other hand, has been largely ignored. Part of the reason for this could be that research students are traditionally expected to read papers in their field and pick up on their own (or with assistance from their research supervisors) the state of research. The few publications in research education in software engineering include papers on how to publish (for example, see [8]) and how to choose the right methodology (for example, see [9]). We believe, it is also an important aspect of research education to gain sufficient meta-level understanding of publications in the discipline in order to be able to appreciate how impactful publications have occurred in SE and how they are related to various factors of publications; this paper aims to make a significant contribution towards this understanding.

Software Engineering, being a relatively young field, is influenced on the one hand by the rapidly growing demands of software industry and on the other by the need to gain recognition as a discipline by the application of scientific and engineering rigour in its research. More than 35 years have passed since the first dedicated venue for publishing software engineering research started in 1975 with the introduction of *IEEE Transactions on Software Engineering* (TSE). Analysing the publication history in software engineering can provide valuable lessons on how publications in general, and successful software engineering researchers in particular, were influenced by the various factors of publishing research outcomes. The scientific identity of a discipline is defined to a large extent by i) the productivity and ii) the interaction patterns of researchers in the discipline. In this paper, we examine these two aspects based on a large corpus of software engineering publications from prominent journals and conferences.

The first aspect, the productivity and the contribution of individual researchers to the SE discipline is a key aspect for examination because a discipline derives its vitality from the fecundity of established researchers, as well as the influx of new ones. Identifying the influences on highly published and highly cited authors in software engineering offers unique insights into what drives the research agendas of SE researchers, as well as the factors critical to their success. In our first research question (*RQ-01*), we examine: *how different factors of publishing relate to citation and publication records of researchers.*

The second aspect is about collaboration among eminent researchers. In this day and age, collaboration is taken as a cornerstone of scientific research. However, different disciplines have different collaboration styles. In mathematics and theoretical sciences the number of joint authors for a paper is usually low, whereas in empirical sciences the authorship traditions are more "generous"; "it is common, for example, for a researcher to be made a coauthor (sic) of a paper in return for synthesizing reagents used in an experimental procedure"! [10]. Software engineering emerged as a sub-discipline of computer science – which has distinctly mathematical roots – but SE research has been observed to become increasingly empirical [11]. How the most highly contributing SE researchers (the "prolifics") collaborate and whether collaboration facilitates research impact reflects on the essential characteristics of research in SE. We address this question in: *RQ-02: How do the prolifics interact amongst themselves?*

In order to find answers to the above research questions, we have inspected a corpus of 19,000+ papers involving 21,000+ authors, derived from 16 major SE publication venues from 1975 to 2010. Since this data-set extends back to the very beginning of organized publication in SE research, *we are able to preclude many of the "missing past" problems that beset similar studies in other disciplines*[12]. Research in a discipline represents a continuum of ideas and engagement of researchers. Thus, studies limited by specific time-windows are prone to incomplete conclusions (see the Related Work section).

The rest of the paper is organised as follows. In Section II, we review the related literature. Section III explains the research method. This is followed by Section IV which discusses our first research question. Similarly, Section V examines the second research question. Section VI gives the limitations of the research by addressing the threats to validity. Finally Section VII gives the conclusions.

II. Related Work

Barabasi et al. report the earliest results from a study of the evolution of scientific collaboration [12]. They examine the co-authorship networks in mathematics and neuro-science between 1991-1998 and conclude scale-free properties of the networks and preferential attachment being the mechanism of network evolution. However, since collaboration is studied for a limited time-window, the authors are unable to explain some of the observations – such as dramatic growth of the largest connected cluster – and ascribe them to the "missing past", that is, the period of time since the beginning of organized publication in the discipline outside the purview of the study. As we remarked earlier, the expanse of our data-set precludes such difficulties. Newman has studied the structure of scientific collaborations in depth to establish that these collaboration networks form small-worlds and these networks show non-trivial clustering [13]. Newman extends his work in following papers, by studying the statistical properties of these networks, and other parameters such as closeness and betweenness [10], [14]. Newman's work highlights the subtly different patterns of scientific collaboration in different disciplines.

Boerner et al. analyze the impact of co-authorship teams by studying a set of 614 articles by 1,036 authors between 1974 and 2004 [15]. They observe a trend towards deepening global collaboration in the production of scientific knowledge. Bettencourt et al. study publication data from six different fields and infer that, while each field develops differently over time, population contagion models adapted from epidemiology can generally explain their development [16]. The dynamics and evolution of scientific disciplines is studied by Herrera et al. [17]. They build an idea network of American Physical Society Physics and Astronomy Classification Scheme (PACS) numbers as nodes representing scientific concepts and use a community finding algorithm to understand the evolution of these fields between 1985-2006. Gerrish and Blei introduce a dynamic topic model for identifying the most influential documents in a corpus, and validate their model on three corpora – fraction of publications from the Association for Computational Linguistics (ACL) anthology, the Proceedings of the National Academy of Sciences (PNAS) and the journal *Nature* [18].

Huang et al. have studied the evolution of research collaboration networks for computer science between 1980 to 2005 [19]. They examine six sub-categories within the discipline and conclude that the database community is the best connected, while the artificial intelligence community is most assortative, and computer science resembles mathematics more than biology. Software engineering has not been recognized and studied as a sub-category within computer science in this work. Bird et al. define 14 sub-areas (including software engineering) and build a collaboration network for computer science in [20]. They use topological measures to study individual behavior and collaboration patterns across these sub-areas. The authors have only considered seven conference venues for software engineering; we do not think this offers a representative sample of SE publication data. A study of collaboration networks based on a very limited data-set – the proceedings of the Working Conference on Reverse Engineering (WCRE) for the period 1993-2002 has been conducted by Hassan and Holt; they reach the conclusion that these have small-world properties [21]. Glass, Vessey, and Ramesh examine 369 papers in six software engineering publication venues and conclude that software engineering research is "... diverse regarding topic, narrow regarding research approach and method, inwardly-focused regarding reference discipline, and technically focused ... regarding level of analysis" [22]. The same set of authors have also compared methods and topics between what they call the "three major subdivisions of the computing realm" – computer science, software engineering, and information systems – and conclude that each field has preferred research approach and methods, which is not necessarily "respected" by the other fields [23]. Our earlier work builds a social network of software engineering research and examines its changing parameters over time [24]. We found an average separation of seven degrees, non-trivial clustering and assortativity, and evidence of increasing collaboration over time. The data-set of the current paper is significantly expanded from that of [24] with more measures adopted for data cleansing and validation.

As evident from the discussion above, understanding the dynamics of scientific research is a rich area of research by itself. However, existing studies of software engineering research collaboration have been of limited scope.

III. Research Method

A. Metrics

To address the research questions, we first need to decide on the metrics that reflect on parameters of our interest.

For measuring the contribution of a researcher, we use two basic measures – publication count and citation count. We assume that the former reflects the amount of research published by a researcher, while the latter indicates the extent to which the researcher's work has been recognised. These two measures are taken to provide distinct yet complementary points of view – the quantitative and qualitative aspects of research contribution. We recognise that merely "counting" the number of papers and citations by way of measuring a researcher's contribution is not without controversy [25]. However as much of academic and industrial research evaluation continues to rely on these measures, we believe our approach is aligned with the state of practice. Thus publication and citation counts are our measures of researcher "success".

In order to identify interactions among highly successful researchers, we use their co-authorship information. Even though researchers can interact among themselves in other ways – for example, being friends or serving on the same committees – such interactions, if research-related, are also likely to result in co-authorships.

B. Research Framework

Next we define a framework for our study, the software engineering research corpus (SERC):

$$SERC =< \mathcal{V}, \mathcal{P}, \mathcal{A}, \mathcal{T}, \mathcal{C}ref >$$

where $\mathcal{V}$ denotes the set of software engineering publication venues we consider (see Table I), $\mathcal{P}$ denotes the set of papers published in these venues from 1975 to 2010, $\mathcal{A}$ denotes the set of authors of these papers. The relation $\mathcal{C}ref \subset \mathcal{P} \times \mathcal{P}$ captures citation information between the publications. In SERC, the list of publication years is denoted by $\mathcal{T} = \{t_0 \cdots t_n\}$ where $t_0 = 1975$ is the starting year and $t_n = 2010$ is the ending year of our measurement-period. A time-step is denoted by $(t_i : t_j)$ which implies a time period from t_i till t_j, both years inclusive.

A paper (or publication) $p \in \mathcal{P}$ has the following attributes:

- $p.a \subset \mathcal{A}$ denotes the set of authors for the paper p,
- $p.y \in \mathcal{T}$ denotes its year of publication, and
- $p.v \in \mathcal{V}$ denotes the publication venue for p.
- $\mathcal{P}(t_i : t_j) \subset \mathcal{P}$ is the set of all papers published in the time-step t_i-t_j.

An author (or researcher) $a \in \mathcal{A}$ has the following attributes:

- $a(t_i : t_j).p$ (or p, for short) is the count of all papers published by a from t_i till t_j.
- $a(t_i : t_j).c$ (or c, for short) is the count of unique co-authors with whom a has published papers from t_i till t_j.
- $a(t_i : t_j).v$ (or v, for short) is the count of unique venues where a has published from t_i till t_j.
- $a(t_i : t_j).s$ (or s, for short) is the count of difference in years (span) between t_i and t_j, that is, the duration between a's first and last publication.

We implemented a tool as shown in Figure 1 that illustrates our approach to instantiate the SERC framework. We have collected publication data published in the list of venues described in Table I. Information around papers published in these venues is available at DBLP[1]. The citation cross indexing module, shown in Figure 1, builds a citation cross reference database between papers in SERC using publicly available information from ACM Digital Library[2], IEEE Xplore[3], and Microsoft Academic Search[4]. Once the cross referencing for all papers—whose citation information could be accessed—is constructed, the citation count for authors is computed. Paper abstracts were also extracted from these bibliographic repositories.

We implemented specialized web crawlers to search each source in turn and store the data in a MySQL database[5]. A set of Java based components was developed to further process and analyse the data.

C. Statistical Analysis

In order to analyse the relationship between the number of publications (and similarly the number of citations) and the different factors of publishing, we used Poisson regression modelling. This was done for two reasons: (1) Poisson analysis is applicable to *count* data, and our dependent variable (that is number of publications and, similarly number of citations) satisfies that criterion, and (2) as shown in Figure 2, the distribution of the variables are highly skewed and is similar to a Poisson distribution. Assumptions of Poisson regression such as control of over-dispersion [26] were also addressed. The significance level, α was set at 0.05 which is common.

The Poisson models obtained were cross-validated using the 80-20 random split method[27]. In this method, we:

1) Split the data-set into two *random* groups where the first group has approximately 80% of the cases and the second group has the remaining cases.
2) Generate the model using the larger set.
3) Validate the model using the smaller set (called the test set) using the following approach:
   a) Use the model to calculate the predicted value for each case in the test set.
   b) Check how the predicted value correlates with the actual value in the test set.

SPSS software was used for statistical analysis.

IV. RQ-01: How do different factors of publishing relate to citation and publication records of researchers?

Typically, a paper is the outcome of significant research by one more authors with varying levels of research experience and published in a journal or refereed conference. It is interesting to explore the relationships between publication and citation rates and the choice of venues, forming of co-authorship relations and research experience. Understanding factors that influence a researcher's contribution can lead us to interesting insights around questions like: Do more productive researchers concentrate on a few venues, or do they prefer to "spread out"? Does enhanced collaboration bring about more publications? While it seems plausible that longer a researcher has been publishing, the larger would be his/her number of papers; actually how strong is the influence of publishing age?

With reference to the earlier discussion, we consider the number of papers published by a researcher, and aggregate citation count as two measures representing his/her contribution. While the number of papers indicate the extent of a researcher's output; citation count – the basis for widely used measures such as *H-Index* and *G-Index* – reflects how much

[1]http://www.informatik.uni-trier.de/ ley/db/
[2]http://dl.acm.org
[3]http://ieeexplore.ieee.org
[4]http://academic.research.microsoft.com
[5]htp://www.mysql.com

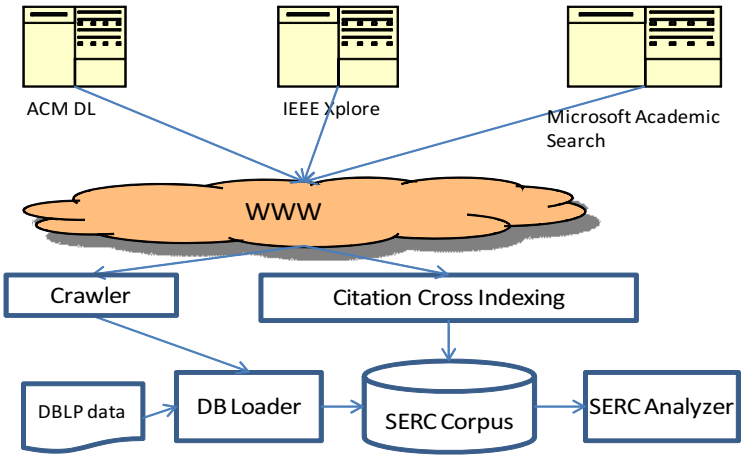

Fig. 1. Schematic diagram of the SERC Tool

TABLE I. SERC: PUBLICATION VENUES AND OTHER DETAILS

TSE - IEEE Transactions on Software Engineering
TOSEM - ACM Transactions on Software Engg. & Methodology
JSS - Journal of Systems and Software
IEEE SW - IEEE Software
ICSE - Intl. Conference on Software Engineering
OOPSLA/SPLASH - Object-Oriented Progg, Systems, Lang. & App.
FSE - Intl. Symposium on the Foundations of Software Engg.
ECOOP - European Conference on Object-Oriented Programming
FASE - Intl. Conf on Fundamental Approaches to Software Engg.
ASE - Intl. Conference on Automated Software Engineering
APSEC - Asia-Pacific Software Engineering Conference
ISSTA - Intl. Conference on Software Testing and Analysis
KBSE - Knowledge-Based Software Engineering Conference
WICSA - Working Conference on Software Architecture
CBSE - Component-Based Software Engineering
ISSRE - Intl. Symposium on Software Reliability Engineering
Total number of years (1975 to 2010, both inclusive) - 36
Total number of venues - 16
Total number of papers - 19,731
Total number of authors - 21,282

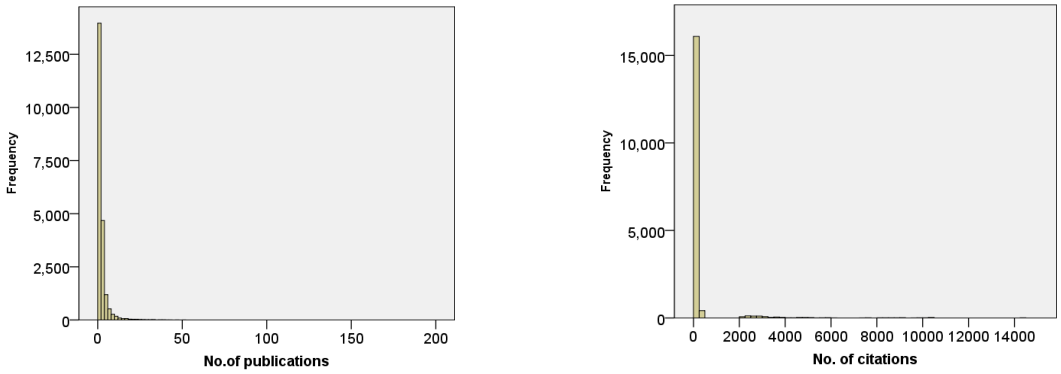

Fig. 2. Histograms of the number of publications and the number of citations for authors

Group	#Citations
1	0-2
2	3-10
3	11-50
4	51-100
5	101-250
6	251-500
7	501-1000
8	1001-2500
9	2501-5000
10	5001-7500
11	7501-10000
12	$>$ 10000

TABLE II. CITATION GROUPS

a researcher's work is being recognized by his/her peers. A researcher's H-Index h is the number of papers that have at least h citations each. And, given a set of publications by a researcher ranked in decreasing order of their citations, his/her G-Index is the largest number such that the top g articles received together at least g^2 citations.

A. Factors Related to the Number of Publications

We first investigate the relationship between the number of papers published by an author (dependent variable), and the number of different co-authors (s)he had, the span of his/her publications, and the number of different venues (s)he published in (independent variables).

As mentioned before, the histograms of the number of publications for authors (as well as citations) in Figure 2 show highly skewed distributions. The distributions of the number of co-authors, venues, and span are also similarly skewed.

Based on the notation introduced in Section III, let $a.p$ be the count of all papers published by the author a, $a.c$ be the count of unique co-authors with whom a has published so far, $a.v$ be the count of unique venues that the author a has published so far, and $a.s$ be the span of publication of a. Poisson regression analysis resulted in the equation:

$$\ln(a.p) = 0.049 + 0.013a.c + 0.236a.v + 0.055a.s$$

All independent variables have a positive relationship with the dependent variable and all three are significant (p-value$<$0.001). We will discuss the implications of this result in Section IV-D.

B. Factors Related to the Number of Citations

We investigated the relationship between a researcher's aggregate citation count with the number of publications, number of different venues of publications, number of co-authors and span of publication (independent variables). Papers were grouped in 12 categories on a sliding scale based on their citation counts as shown in Table II; this was done to control over-dispersion in regression analysis as mentioned in Section III-C.

The relationship is expressed in the following regression equation:

$$\ln(g) = 0.676 + 0.003a.c + 0.159a.v + 0.032a.s - 0.01a.p$$

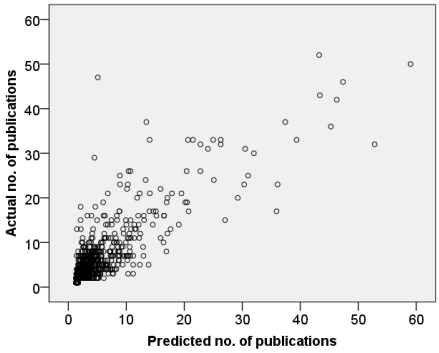

Fig. 3. Scatter diagram of actual versus predicted #publications

where g is the citation group where an author belongs, $a.c$ is the number of co-authors, $a.v$ is the number of venues, $a.s$ is the span of publication and $a.p$ is the number of publications, for the author a. All parameters except that for c were significant (p-value$<$0.001); for c, the significance was 0.051 which is very close to 0.05, our α level. Therefore we did not discard it from the regression equation. However, as demonstrated by the high p-value, the coefficient—and thus influence—of c in the equation is very small.

C. Model validation

The two models were validated by applying them on the test cases and checking the correlation between the predicted and actual values. For the publications model, the correlation was very strong (Spearman's ρ = 0.818; p-value$<$ 0.001) (Figure 3 shows the scatter diagram.) For the citations model, the correlation was medium (Spearman's ρ = 0.403; p-value$<$ 0.001).

D. Discussion

We can make the following observations based on the results of the regression analysis.

- While it is reasonable to expect that the number of co-authors, venues, and publishing years of an author will influence his/her total publication count, the coefficients of these variables in the above equation help us understand the weights of these factors in the overall relationship. Accordingly, we see that the number of venues where an author publishes has a stronger relationship on his/her number of publications, compared to the number of co-authors and the number of years (s)he has been publishing. This is counter-intuitive, as one usually expects the number of years of publishing or number of co-authors as stronger influencing factors on the number of papers published. But from the empirical evidence, we see that there is less commonality among highly published authors in terms of co-authorships or span of publishing, compared to spreading their publications over a larger number of venues. Each publication venue generally has particular focus. So authors who publish in many

different venues are likely to have a wider arc of research interests, which in turn appears to lead them to publish more.

- The number of publications has a negative influence on the citation group. That is, authors who are in the high citation categories have relatively lesser number of publications. Assuming that citation count is a reliable proxy for the quality of research output, this observation aligns with the traditional *quality versus quantity* argument, where increasing quantity does not necessarily correspond to enhanced quality.
- Those who are in higher citation categories publish in a larger number of venues. While the impact factor of venues can differ, it seems that widely cited authors do not restrict their publications to fewer high impact venues.
- The influence of reciprocity (that is, *you scratch my back and I will scratch yours* principle, in common parlance) has been investigated in different social science settings [28] [29]. One could argue that collaboration through co-authorship could potentially lead to formation of a friendly club where authors cite each others' work, thus increasing the citation count of the club members. Our analysis shows that this is *not* a factor in software engineering publications. The coefficient of co-authorship is very low and its p-value, as mentioned earlier, is close to it being not statistically significant.
- The number of years an author has been publishing has a positive influence on citations. This is expected, as it has been observed that researchers need a period of time to come across papers, study and cite them in their own work.

V. RQ-02: How do the prolifics interact amongst themselves?

As observed, the number of publications and aggregate citation counts have highly skewed distributions with long right tails (Figure 2). This implies few researchers dominate the field in terms of the number of papers published, as well as the number of citations attracted. In the preceding section we also discerned the influences on high publication counts vis-a-vis high citation counts for researchers. Let us now examine how a select group of researchers – the *prolifics* – interact.

A. Analysis Approach

From our pool of authors $\mathcal{A}$, we isolate two sets of prolifics: the top 100 researchers ranked by the number of papers published ($\mathcal{A}_p$), and the top 100 researchers ranked by the aggregate number of citations received ($\mathcal{A}_r$). Out of the 200 researchers across these sets, 192 are unique (96%). We note that the members of $\mathcal{A}_p$ have published 9,006 unique papers between them, while the members of $\mathcal{A}_r$ have published 6,406 unique papers between them. Out of the total papers and authors in our corpus (Table I), it is significant that the 100 odd researchers in each of $\mathcal{A}_p$ and $\mathcal{A}_r$ represent less than 5% of the author pool but have contributed between 31-47% of the papers published.

We are interested in finding how the members of $\mathcal{A}_p$ and $\mathcal{A}_r$ interact amongst themselves within each group. In the following discussion, "peer(s)" of some researcher a_i, will mean other member(s) of $\mathcal{A}_p$ or $\mathcal{A}_r$ to which a_i belongs.

B. Results and Discussion

To understand the interaction between peers, we explore two perspectives. Table III presents statistics around the co-authorship and citation patterns of prolifics and Table IV highlights aspects of their interactions in pairs. As we observe in Table III, the mean number of peers who is a co-author of a researcher in $\mathcal{A}_p$ is 6.40 vis-a-vis 1.16 in $\mathcal{A}_r$. We also see that in $\mathcal{A}_p$, a researcher cites around 27 of his/her peers, while a member of $\mathcal{A}_r$ cites only around 4 peers, on average. Going by Cohen's criteria, the correlations (as given by the Pearson correlation coefficients) between the number of peers collaborated with vis-a-vis the number of peers cited is strong for $\mathcal{A}_p$ (0.689), while moderate for $\mathcal{A}_r$ (0.443) in Table III [30]. From the above, it is evident that prolifics by citations interact significantly less with their peers when compared to prolifics by publications.

To get a sense of the extent to which members of a peer group collaborate amongst themselves, we define the *Co-authoring Index*:

Definition 1: Co-authoring Index of a peer group of researchers is the ratio of the number of pairs of researchers in a peer group who have co-authored at least one paper, to the maximum number of such pairs in that peer group.

Thus higher co-authoring index in a peer group indicate a higher extent of collaboration among members of the group and vice-versa. From Table IV we see that the value of Co-authoring Index is 0.065 for $\mathcal{A}_p$ and 0.011 for $\mathcal{A}_r$. Thus members of the latter group collaborate amongst themselves almost six times less than those of the former.

Writing a paper together is likely to make researchers more familiar with one another's research interests. It is interesting to examine whether such familiarity makes it more probable for researchers to cite co-authors' work and also whether citing someone leads to being cited back as a return of compliment. To help us discern these trends, in Table IV, $C(a_x \Rightarrow a_y)$ denotes the number of times author a_x cites a_y, $J(a_x + a_y)$ denotes the number of papers in which a_x and a_y are co-authors, and $S(a_x \Leftrightarrow a_y)$ is the sum of the number of times a_x and a_y have cited each other, that is, $S(a_x \Leftrightarrow a_y) = C(a_x \Rightarrow a_y) + C(a_y \Rightarrow a_x)$.

We observe that $C(a_x \Rightarrow a_y)$ and $C(a_y \Rightarrow a_x)$, as well as $J(a_x + a_y)$ and $S(a_x \Leftrightarrow a_y)$ are strongly correlated for $\mathcal{A}_p$ (0.585 and 0.602), and weakly correlated for $\mathcal{A}_r$ (0.315 and 0.215) in Table IV [30]. Thus there is evidence that highly published researchers tend to reciprocate to their peers the compliment of being cited, to a much larger extent than highly cited researchers. Additionally, highly published researchers are much more prone to citing their collaborators than the highly cited researchers.

VI. Threats to Validity

As in any research, there are limitations that could affect the validity of the outcomes.

TABLE III. CHARACTERISTICS OF PROLIFICS ACROSS INDIVIDUALS ($^* p - value < 0.001$)

		$\mathcal{A}_p$	$\mathcal{A}_r$
#Peers as co-authors (P)			
	Mean	6.40	1.16
	SD	4.73	1.41
	95% CI	5.47 to 7.33	0.88 to 1.44
#Peers cited (R)			
	Mean	26.75	3.77
	SD	17.11	4.66
	95% CI	23.40 to 30.10	2.86 to 4.68
Pearson coeff: P & R		0.689^*	0.443^*

TABLE IV. CHARACTERISTICS OF PROLIFICS ACROSS PAIRS OF INDIVIDUALS; ($^* p - value < 0.001$)

	$\mathcal{A}_p$	$\mathcal{A}_r$
Co-authoring Index	0.065	0.011
Pearson coeff: $C(a_x \Rightarrow a_y)$ & $C(a_y \Rightarrow a_x)$	0.585^*	0.315^*
Pearson coeff: $J(a_x + a_y)$ & $S(a_x \Leftrightarrow a_y)$	0.602^*	0.215^*

A. Construct validity

Construct validity implies that variables are measured correctly. In areas where there is considerable theoretical work, it usually involves establishing that the measurements are constructed in accordance with theoretical foundations in the area. Measuring impact and importance of a publication by counting the number of citations is well accepted, however, there are several other measures of impact of individuals such as *H-Index* and *G-Index*, that we have not used.

B. Internal validity

A study shows internal validity if it is free from systematic errors and biases. Since our data set is derived from all accessible publications in a predefined set of venues, issues that can affect internal validity such as mortality (that is, subjects withdrawing from a study during data collection) and maturation (that is, subjects changing their characteristics during the study outside the parameters of the research) do not arise in our case. A threat to validity in our context is selection bias which occurs when the sample chosen in not representative of the population it is expected to represent. We have chosen 16 major publication venues that focus on software engineering research. Although we believe our data-set covers a major portion of the discipline's research publication corpus, we cannot claim to have captured *all* published software engineering papers in 1975-2010, . Whether or not a particular venue included in SERC exclusively focuses on software engineering is a matter of judgement, as is the question of what is truly a software engineering paper. Thus it is likely SERC consists of some papers which relate to software engineering only in a broad sense, and SERC has missed out some software engineering related papers published in other venues. As mentioned earlier, the citation counts were based on citation cross indexing between papers that we constructed across several of our data sources. For the papers for which citation information is not available in the public domain, could not be included in our analysis. A common problem in studies of scientific publication comes from the ambiguity of author names. If the same author chooses to be identified as "John Doe" and "J Doe", it is difficult to reconcile their identities. Conversely, if there are multiple individuals with the exact same moniker "John Doe" it is difficult to distinguish them. Such inconsistencies are minimized in DBLP through significant human intervention [31]. To further enhance accuracy, we manually verified the names all the authors considered in RQ-02.

C. External validity

External validity indicates the generalisability of the results of the study. The population for our study is all software engineering publications. Our sample size and the sampling method are unlikely to be a threat to external validity.

VII. CONCLUSIONS

Software engineering education themes tend to involve coursework education rather than research education. This paper, in contrast, makes a contribution to research education in software engineering by conducting a meta-analysis of research publications in software engineering. The results presented are from a large scale empirical study of software engineering research publications across three and half decades, starting from the very inception of organised research publication in the discipline.

We investigated two research questions. The first question (RQ-01) showed that authors with high publication rates did not concentrate on a few selected venues to publish their papers; instead, the number of venues had a stronger relationship with publication numbers than citations or co-authorships or even how many years an author has been publishing. Authors with high citation counts were found to publish lesser number of papers. Thus a highly cited researcher seems more likely to focus on specialised topic(s) of interest, striving to write relatively few but higher impact papers, and seeking to establish him or herself as an authority in the chosen topics.The investigation also found that reciprocity (mutual citations) is not a significant phenomenon in software engineering research.

The second research question (RQ-02) showed that highly published authors interact among one another quite differently to a similar group of highly cited authors. The latter group co-authors papers or cites their peers to a much lesser extent than the former. When the observation from RQ-01 that highly cited authors tend to publish lesser number of papers is combined with the findings from RQ-02, it seems plausible that researchers with high citation counts – each viewing oneself as an expert in one's field – would have little inclination or necessity to co-author a paper with, or cite one another.

Our results offer unique insights into the ecosystem of software engineering research, and we hope these insights will assist research students and their supervisors in discerning the

key patterns of behaviour that influence research impact in terms of number of publications and citations.

REFERENCES

[1] J. Favela and F. Peña-Mora, "An experience in collaborative software engineering education," *Software, IEEE*, vol. 18, no. 2, pp. 47–53, 2001.

[2] L. Ohlsson and C. Johansson, "A practice driven approach to software engineering education," *Education, IEEE Transactions on*, vol. 38, no. 3, pp. 291–295, 1995.

[3] O. Hazzan, "The reflective practitioner perspective in software engineering education," *Journal of Systems and Software*, vol. 63, no. 3, pp. 161–171, 2002.

[4] P. B. Henderson, "Mathematical reasoning in software engineering education," *Communications of the ACM*, vol. 46, no. 9, pp. 45–50, 2003.

[5] L. Williams and R. Upchurch, "Extreme programming for software engineering education?" in *Frontiers in Education Conference, 2001. 31st Annual*, vol. 1. IEEE, 2001, pp. T2D–12.

[6] E. Ye, C. Liu, and J. A. Polack-Wahl, "Enhancing software engineering education using teaching aids in 3-d online virtual worlds," in *Frontiers In Education Conference-Global Engineering: Knowledge Without Borders, Opportunities Without Passports, 2007. FIE'07. 37th Annual*. IEEE, 2007, pp. T1E–8.

[7] K. Toth, "Experiences with open source software engineering tools," *Software, IEEE*, vol. 23, no. 6, pp. 44–52, 2006.

[8] M. Shaw, "Writing good software engineering research papers: minitutorial," in *Proceedings of the 25th international conference on software engineering*. IEEE Computer Society, 2003, pp. 726–736.

[9] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, "Selecting empirical methods for software engineering research," in *Guide to advanced empirical software engineering*. Springer, 2008, pp. 285–311.

[10] M. Newman, "Scientific collaboration networks. i. network construction and fundamental results," *Physical Review E*, vol. 64, no. 1, p. 016131, 2001. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.64.016131

[11] M. Shaw, "Continuing prospects for an engineering discipline of software," *IEEE Software*, vol. 26, pp. 64–67, 2009.

[12] A. L. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A 311, (3-4) (2002), pp. 590-614*, 2001.

[13] M. E. J. Newman, "The structure of scientific collaboration networks," *Proc. Natl. Acad. Sci. USA 98*, pp. 404–409, 2000.

[14] M. Newman, "Scientific collaboration networks. II. shortest paths, weighted networks, and centrality," *Physical Review E*, vol. 64, no. 1, p. 016132, 2001. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.64.016132

[15] K. Börner, L. Dall'Asta, W. Ke, and A. Vespignani, "Studying the emerging global brain: Analyzing and visualizing the impact of co-authorship teams," *Complexity*, vol. 10, p. 57–67, 2005.

[16] L. M. A. Bettencourt, D. I. Kaiser, J. Kaur, C. Castillo-Chvez, and D. E. Wojick, "Population modeling of the emergence and development of scientific fields," *Scientometrics*, vol. 75, no. 3, pp. 495–518, 2008.

[17] M. Herrera, D. C. Roberts, and N. Gulbahce, "Mapping the evolution of scientific fields," *PLoS ONE*, vol. 5, no. 5, p. e10355, May 2010. [Online]. Available: http://dx.doi.org/10.1371/journal.pone.0010355

[18] S. Gerrish and D. M. Blei, "A language-based approach to measuring scholarly impact." in *ICML*, J. Frnkranz and T. Joachims, Eds. Omnipress, 2010, pp. 375–382. [Online]. Available: http://dblp.uni-trier.de/db/conf/icml/icml2010.html#GerrishB10

[19] J. Huang, Z. Zhuang, J. Li, and C. L. Giles, "Collaboration over time: characterizing and modeling network evolution," in *Proceedings of the international conference on Web search and web data mining*, 2008, pp. 107–116.

[20] C. Bird, E. Barr, A. Nash, P. Devanbu, V. Filkov, and Z. Su, "Structure and dynamics of research collaboration in computer science," in *SDM*, 2009, pp. 826–837.

[21] A. Hassan and R. Holt, "The small world of software reverse engineering," in *Proceedings of 11th Working Conference on Reverse Engineering*, 2004, pp. 278–283.

[22] R. L. Glass, I. Vessey, and V. Ramesh, "Research in software engineering: an analysis of the literature," *Information and Software Technology*, vol. 44, no. 8, pp. 491–506, 2002.

[23] R. L. Glass, V. Ramesh, and I. Vessey, "An analysis of research in computing disciplines," *Commun. ACM*, vol. 47, pp. 89–94, June 2004. [Online]. Available: http://doi.acm.org/10.1145/990680.990686

[24] S. Datta, N. Kumar, and S. Sarkar, "The social network of software engineering research," in *Proceedings of the 5th India Software Engineering Conference*, ser. ISEC '12. New York, NY, USA: ACM, 2012, pp. 61–70. [Online]. Available: http://doi.acm.org/10.1145/2134254.2134265

[25] D. L. Parnas, "Stop the numbers game," *Commun. ACM*, vol. 50, no. 11, pp. 19–21, Nov. 2007. [Online]. Available: http://doi.acm.org/10.1145/1297797.1297815

[26] P. Dugard, J. Todman, and H. Staines, *Approaching multivariate analysis: a practical introduction*. Routledge, 2010.

[27] B. Tabachnick and L. Fidell, *Using Multivariate Statistics*. Boston: Pearson Education, 2007.

[28] L. Stanca, "Measuring indirect reciprocity: Whose back do we scratch?" *Journal of Economic Psychology*, vol. 30, no. 2, pp. 190–202, 2009.

[29] E. Fehr and S. Gächter, "Fairness and retaliation: The economics of reciprocity," *The Journal of Economic Perspectives*, vol. 14, no. 3, pp. 159–181, 2000.

[30] J. Cohen, *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum, 1988.

[31] M. Ley and P. Reuther, ": The problem of data quality," *EGC*, vol. RNTI-E-6, pp. 5–10, 2006.