Life in Software Engineering Conferences: A Text Mining Study

George Mathew
Computer Science, NCSU
Raleigh, North Carolina
george.meg91@gmail.com

Tim Menzies
Computer Science, NCSU
Raleigh, North Carolina
tim.menzies@gmail.com

Amritanshu Agrawal
Computer Science, NCSU
Raleigh, North Carolina
aagrawa8@ncsu.edu

ABSTRACT

Abstract here

Keywords

ACM proceedings; LaTeX; text tagging

1. INTRODUCTION

From 1993 to 2013 there have been more than 9000 papers published in top 11 Software Engineering(SE) conferences and the number of conferences have been increasing gradually since the inception of SE in 1968. Since the research literature has grown tremendously over the last two decades, there is a need of identifying the pattern of papers published in SE conferences.

Bibliometrics studies in SE have focused in the following areas; (a) generating ranking lists of top performing institutions and scholars [7], (b) citation analysis to identify the most popular articles [13], and (c) content analysis of SE research [8, 1, 2]. Papers in area (a) can mainly be used internally within the SE research community. Papers on areas (b) and (c) can be used to explain our science to outsiders, e.g. to funding authorities or to scientists representing other disciplines. Additionally, such works can be helpful in teaching students about software engineering research or to highlight the top areas under study to industry, and help outsider to get acquainted with the latest research trends. Thus, bibliometrics papers can be important aid in distributing knowledge beyond the software engineering community. But, there is a need for identifying the patterns in the papers and their in software engineering and how it has changed over the last two decades which can aid future researchers in selecting the SE sections that call for new research and the approach to be taken to successfully publish in SE.

To carry out this research we propose the following research questions:

- **RQ1: What research career is more rewarded?**
  Do authors who publish in multiple conferences of diverse topics or do they publish in select conferences with specific conferences?

Result 1
We find that most authors publish in around half of all the topics and only very few authors venture into all the topics in software engineering.

- **RQ2: Are conferences diverse?**
  Do most conferences in SE cover similar sets of topics? or Are different conferences specific to different topics?

Result 2
80% papers from all conferences are from 60% of the conferences. i.e Papers from a limited set of topics are mostly published in all the conferences.

- **RQ3: Do conferences evolve?**
  Does the focus of a conference change over time? or Does a conference tend to focus on similar topics?

Result 3
From 4 of the top conferences in SE, we find that in 3 of them the primary topic of the papers published change over time.

- **RQ4: Is there a program committee bias?**
  What percentage of papers published in a conference have authors in the program committee(PC) of the conference?

Result 4
In 3 of the 4 conferences in this study, there is less than 30% bias from the PC; and in the last conference there exists around 40% bias from the PC.

- **RQ5: Do topics evolve?**
  What percentage of papers published in a conference have authors in the program committee(PC) of the conference?

Result 5
From papers in top 11 conferences of SE, there is a change in the topic of focus almost every 3-4 years.

WOODSTOCK '97 El Paso, Texas USA

DOI: 10.475/123_4

2. RELATED WORK

Add stuff from these papers [3, 5, 11, 12, 10].

Numerous bibliometric studies analyzing SE conferences have been published in SE, most of which have been in the last 2 decades. This section discusses few of the prominent works with their notable findings.

Cai & Card [1] analyzed 691 papers from 7 leading journals SE and 7 leading conferences SE in the year 2006. Among their findings was that 89% of conference papers focus on 20% of subjects in SE, including software/program verification, testing and debugging, and design tools and techniques. Moreover, the average number of 7 top international conferences in SE references cited by a conference paper is 24.

Garousi & Vahid in 2014 presents a bibliometric analysis of the Turkish SE community (researchers and institutions) based on the Scopus academic search engine. Based on their study they find that there was a lack of diversity in the general SE spectrum(for eg. limited focus on requirements engineering, software maintenance and evolution, and architecture). They also identify a low involvement from the industry in SE. Although this research was conducted in only in Turkish SE community, this study can be extended to identify similar patterns in the entire SE community. Similar study was conducted early in 2010 on the Canadian SE community [6].

More recently, a citation based study was conducted by Garousi and Fernandes to identify the top cited paper in SE community [4]. The authors use two metrics: total number of citations and average annual number of citations to identify the top paper. The authors also try characterizing the overall citation landscape in SE motive and hope that this method encourages further discussions in the SE community towards further analysis and formal characterization of the highly-cited SE papers. Although the authors encourage an important need in SE of characterization, the approach using citations has other alternatives.

In Computer Science(CS), a 2013 paper by Hoonlor et al.highlighted the prominent trends in CS [9].This paper identified trends, bursty topics, and interesting inter-relationships between the American National Science Foundation (NSF) awards and CS publications, finding, for example, that if an uncommonly high frequency of a specific topic is observed in publications, the funding for this topic is usually increased. The authors adopted a Term Frequency Inverse Document Frequency(TFIDF) based approach to identify trends and topics. A similar approach can be performed in SE considering how closely CS is related to SE.

Garousi and Mantyla recently have adopted a Topic Modeling and Word Clustering based approach to identify topics and trends in SE [5] similar to the research by Hoonloor et al. [9]. Although their method is very novel and in line with the current state of the art, they use only the titles of the papers for modelling topics. This might lead to inaccurate topics as titles are generally not very descriptive of the field the paper is trying to explore. This issue is addressed in the current work where we use the abstracts of the paper to build topic models.

XXX Research to discussing inbreeding in SE

3. DATASET & RESEARCH METHOD

3.1 Gathering Data

XXX George fills this

3.2 Latent Dirichlet Allocation

XXX Amrit fills this

4. STUDIES

Studies answering the Research Questions

4.1 RQ1: What research career is more rewarding?

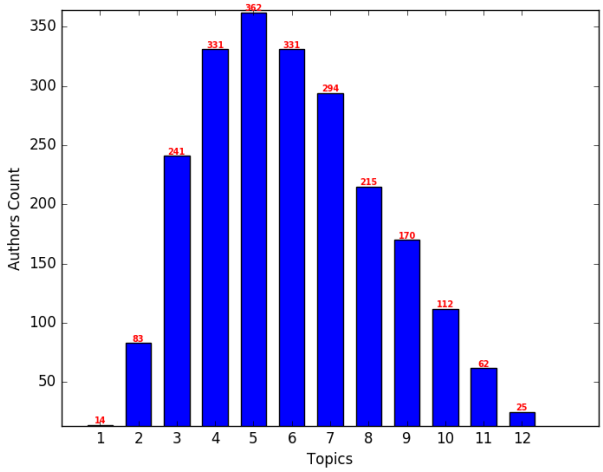

Figure 1: Number of authors vs Number of Topics Used.

In this section we study the number of authors who publish papers in different topics and identifying the optimal success strategy for publishing in software engineering conferences. Figure 1 shows the number of authors publishing in "x" number of distinct topics where "x" varies from 1 to 12 as shown on the x-axis of the graph.It is clear that from over 2240 thousand authors publishing in various software engineering conferences in the past 20 years close to 46% of the authors publish in 4-6 topics and close to 80% of the authors publish in 3-8 topics. But the question arises what percentage of the top authors fall in the 80% of authors publishing between 3-8 topics.

Figure 2 shows the results of the study which was repeated considering only the top 1% cited authors in SE. From the figure we can see that all of the top 1% authors publish between 10-12 topics. Thus, none of the top 1% authors fall into the category of authors who publish in the majority crop of authors who publish in 3-8 topics.

Result 6
Successful authors publish in different topics.

4.2 RQ2: Are conferences diverse?

In this section we check if most conferences are same or different. A heatmap as shown in Figure 3 shows which conferences are similar to each other. The y axis of the heatmap shows the index of the topic ranging between 0-11 and the x axis shows the conference names used in the study.The color on the heatmap indicates how strongly or

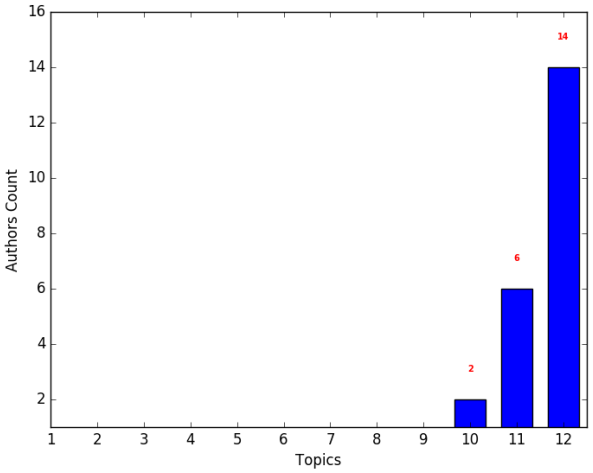

Figure 2: Number of authors vs Number of Topics Used for top 1% of the authors

weakly a topic on the y-axis is covered in a conference on x-axis.

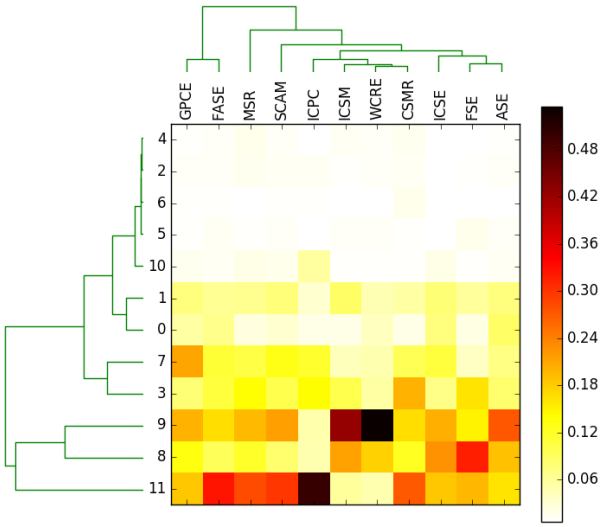

Figure 3: Heatmap showing correlation between topics and conferences.

From the hierarchical tree on the x-axis we can see that there is a strong correlation between conferences with respect to the topics covered in them. For example, ICPC, ICSM, WCRE and CSMR cover similar topics and top conferences like ICSE, FSE and ASE cover similar topics. On the other hand, conferences like GPCE and FASE are completely different from other conferences. Thus we can conclude that, there are conferences staggeringly similar to each other but there are also conferences which significantly differ in the topics published in them.

Result 7
Top conferences publish papers on similar topics but there exists smaller conferences that focus on different topics.

4.3 RQ3: Do conferences evolve?

In this section we examine if the topics concentrated in a conference changes over time. In Figure 4 to Figure 7 4 top conferences ICSE, ASE, FSE and MSR are used to study topic coverage % of papers published in them from 1993-2013. The y-axis in each figure represents a stacked bar chart in descending order indicating the topic coverage percentage in a conference(title) for a certain year(x-axis). The x-axis represents the year of the conference from 1993-2013.

Consider ICSE, in years 1998-2004 topic 7 is the most published topic and the percentage of papers published in topic 7 decreases almost consistently. Similarly topic 11 slowly gains popularity from 2007-2010 and eventually becomes the most popular topic from 2011-2013. Similar trends can be observed in ASE and FSE.But when it comes to MSR, we can see that the most popular topic has always been topic 11(expect for 2010).

Result 8
Most conferences evolve over time but there are conferences that prefer concentrating on a niche set of topics.

4.4 RQ4: Is there a program committee bias?

In this section we check what percentage of papers published in a conference are (co)authored by members of the program committee(PC) and examine if a bias exist in paper acceptance. Figure 8 shows the % of papers (co)authored my PC members per year(on y-axis) in four prominent conferences from 1993 - 2013(on x-axis). We can observe that in ICSE, FSE and ASE less than 30% of papers published have atleast one of the authors in the PC(average of 15%). Whereas in MSR over 40% of papers in average each year have atleast one of the authors in the PC.

Result 9
Conferences have a low Program Committee bias with few exceptions of high bias.

4.5 RQ5: Do topics evolve?

In this section we study the trend in topics of papers published from 1993-2013 and check if there is a change in affinity towards a topic every year. Figure 9 shows the stacked coverage % of top 5 topics in descending order(y-axis) in a year(x-axis) for 9281 papers from top 11 conferences over 20 years.

From the Figure 9, we can see that topic 7 was highly popular from 1998-2004 and since then the percentage of accepted papers have gradually reduced. On the contrary, topic 13 has gained popularity gradually from 2008-2013. Topic 8 is slowly gaining popularity from 2005 and based on the trend, it might become the most popular topic in which papers are published in the next 5 years.

5. DISCUSSION

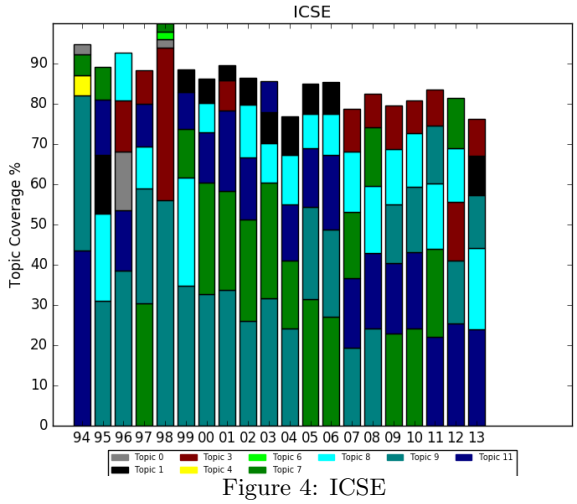

Figure 4: ICSE

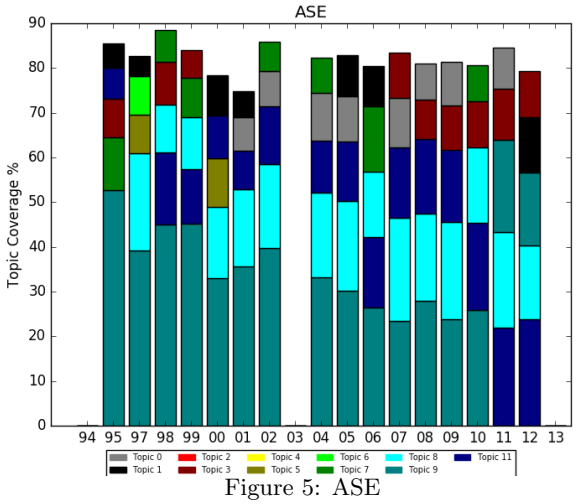

Figure 5: ASE

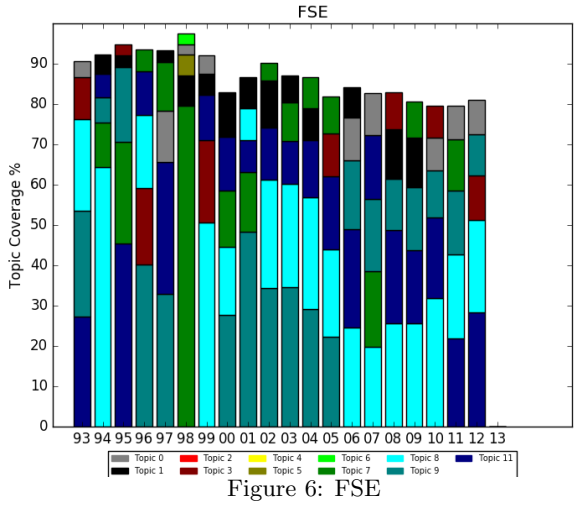

Figure 6: FSE

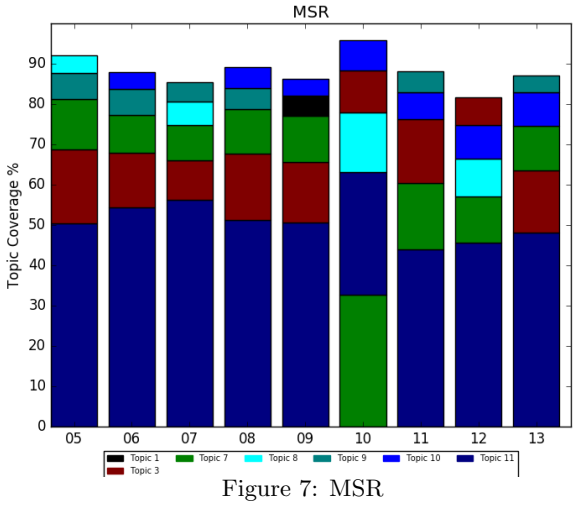

Figure 7: MSR

Topic Coverage % vs Year for 4 top conferences from 1993-2013

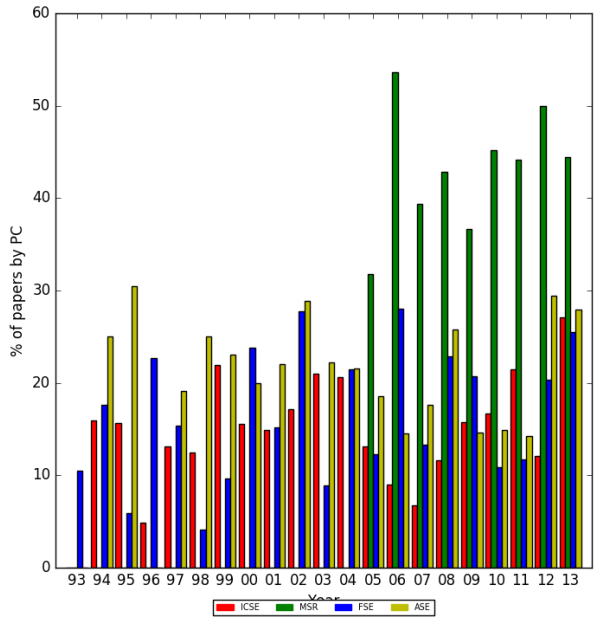

Figure 8: % of papers by PC vs year

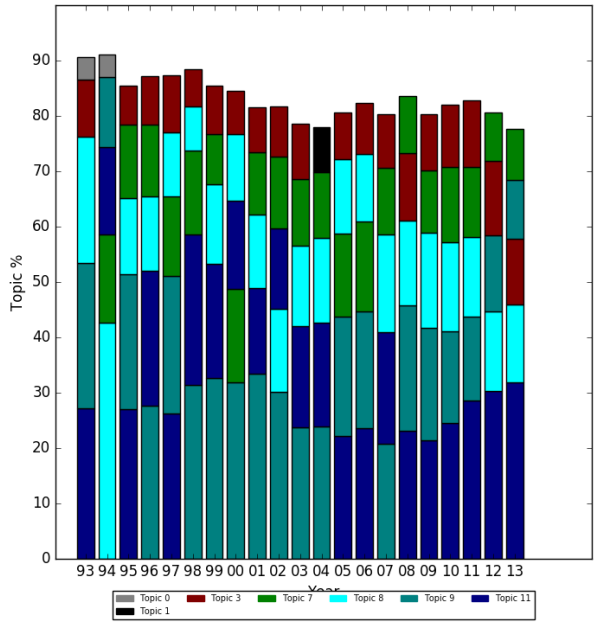

Figure 9: Topic Coverage % vs Year

6. THREATS TO VALIDITY

7. ACKNOWLEDGMENTS

8. REFERENCES

[1] K.-Y. Cai and D. Card. An analysis of research topics in software engineering–2006. *Journal of Systems and Software*, 81(6):1051–1058, 2008.

[2] N. Coulter, I. Monarch, and S. Konda. Software engineering as seen through its research literature: A study in co-word analysis. *Journal of the American Society for Information Science*, 49(13):1206–1223, 1998.

[3] J. M. Fernandes. Authorship trends in software engineering. *Scientometrics*, 101(1):257–271, 2014.

[4] V. Garousi and J. M. Fernandes. Highly-cited papers in software engineering: The top-100. *Information and Software Technology*, 71:108–128, 2016.

[5] V. Garousi and M. V. Mäntylä. Citations, research topics and active countries in software engineering: A bibliometrics study. *Computer Science Review*, 19:56–77, 2016.

[6] V. Garousi and T. Varma. A bibliometric assessment of canadian software engineering scholars and institutions (1996-2006). *Computer and Information Science*, 3(2):19, 2010.

[7] R. L. Glass and T. Y. Chen. An assessment of systems and software engineering scholars and institutions (1999–2003). *Journal of Systems and Software*, 76(1):91–97, 2005.

[8] R. L. Glass, I. Vessey, and V. Ramesh. Research in software engineering: an analysis of the literature. *Information and Software technology*, 44(8):491–506, 2002.

[9] A. Hoonlor, B. K. Szymanski, and M. J. Zaki. Trends in computer science research. *Communications of the ACM*, 56(10):74–83, 2013.

[10] O. Hummel, A. Gerhart, and B. Schäfer. Analyzing citation frequencies of leading software engineering scholars. *Computer and Information Science*, 6(1):1, 2013.

[11] T. Systä, M. Harsu, and K. Koskimies. Inbreeding in software engineering conferences, 2012.

[12] B. Vasilescu, A. Serebrenik, and T. Mens. A historical dataset of software engineering conferences. In *Proceedings of the 10th Working Conference on Mining Software Repositories*, pages 373–376. IEEE Press, 2013.

[13] C. Wohlin. An analysis of the most cited articles in software engineering journals-2000. *Information and Software Technology*, 49(1):2–11, 2007.