Characterizing Women (Not) Contributing To Open-Source

Blinded Authors' Names
Blinded Authors' Affiliations

Abstract—**Women are under-represented not only in software development, but also in the Open-Source Software (OSS) community. Based on previous research, there are observed differences between developers who contribute to OSS and those who do not. In this study we examine the existence of the same differences as present in a sample of women. Characterizing women who participate in OSS may help to attract other women to contribute to OSS. Furthermore, it might uncover potential biases in data about female developers that are gathered through the mining of software repositories.**

Using the data from the Stack Overflow Developer Survey 2018, counting 100,000+ respondents (6.9% female), we compare the characteristics of women who report to contribute to OSS and those who report to not contribute. Surprisingly, we did not found the differences that we expected based on previous literature, thus suggesting that open-source software data seem to represent well the closed-source population, in the context of female developers. However, the correlates of female under-representation in OSS remain unexplained.

Index Terms—**Women in Software Community; Human Aspects in Software Engineering; Open-Source Software.**

I. Introduction

Empirical software engineering research relies greatly on open-source software (OSS) data being available to use [16]. Based on this data, various characteristics of software systems, e.g., code quality or community and social aspects of software development, are investigated [9], [5], [10].

Collecting demographic data (such as age, education, professional experience, and gender [30]) pertaining to OSS contributors is far from trivial, in particularly when analyzing solely a project's code base, defects, and its historical changes [9]. Nevertheless, research showed that demographic characteristics play a key role in software development, in the development process itself but also, for example, in the way developers interact with each other [27], [29], [1]. Supporting gender diversity in OSS teams can bring several benefits, such as higher productivity [30].

When studying the developers community, an appropriate female representation in the investigated samples is of a great importance due to existing gender differences. Often times, research does not account for the representation bias and draws conclusions about gender differences on samples having as little as 2% of women [3].

In an OSS setting, gender is more anonymous than in closed-source setting; indeed, as the famous New Yorker cartoon stated: "On the Internet, nobody knows you're a dog" [13]. Nonetheless, there is even greater under-representation of women in OSS than in closed-source software development. In OSS, women account for 4% of developers as opposed to closed-source environment, where there is 8% of women [4].

Due to to the lower representation of women in OSS environment, it is harder to achieve team gender diversity. Thus, gender diversity research has better chances of obtaining representative samples and of identifying more cases to study in the closed-source environment. Nevertheless, there are studies focusing on the issue of gender diversity that use OSS data exclusively [7], [30].

Even though research results from OSS settings might be transferable to the closed-source development, little is known about the differences between the population of developers in both contexts. The first study of this kind pointed out there might be considerable differences between OSS contributors and other developers [3]. With our previous work [4] we have addressed the representativeness of investigating developers in OSS; except for the greater gender representation gap, we found that OSS developers are more learning-oriented, more experienced, and with higher perception of their own competence than their closed-source counter-part.

In this paper, we extend on this line of research by further examining the differences between women contributing to OSS and those who do not. By investigating these differences, we attempt to provide insights on the correlates of the representation mismatch. Further, this investigation can help (1) guiding future efforts in attracting women to contribute to OSS, (2) collecting data about female developers with empirical software engineering research, and (3) pointing at potential differences in women's behaviour in software development.

II. Considered Dimensions

In this section we outline the rationale for the dimensions that we use to compare women software practitioners contributing to OSS vs. those who do not. These dimensions were selected based on their relationship with code quality as well as with interpersonal interaction during the development process. As such, differences in these dimensions across different populations could potentially be related to differences in code, networking and collaboration of (female) developers, as well as their participation in OSS.

Experience and Age. The developers coding experience is related to the code quality and to the amount of introduced bugs, as well as to the quality of social bonds within the team [6], [12]. It is, however, not connected with voluntary OSS contribution, even though younger developers are more likely to contribute [3]. Further, there is a bias in OSS population towards more experienced developers [4]. We examine whether this bias holds for females as well, as women compose a greater portion of developers with short experience [19].

Likewise, we expect distinct developer roles may be represented differently between OSS and closed-source. While system admins or DevOps specialists are much more likely to be men than women, academics, QA developers, data scientists or designers have relatively higher female representation [19]. In our analysis, we focus on differences between Developers, Data Scientists, DevOps-related positions, and Students.

Education. Education does not seem to be related with the probability of OSS contribution nor with the extent of activity within an OSS project, even though the OSS contributors are typically well-educated [1], [23] [4]. We examine whether this assumption holds true for female.

Perceived Competence. Women are generally more affected by Competence-Confidence Gap, an unjustified low belief in own competence, that prevents them from contributing to OSS projects [22], [31]. The belief in own competence is also important for a successful progression within OSS projects [11]. Therefore, we examine how pressing this issue is for women and their further development in the OSS community.

Kinship and Competition. Empirical software engineering research has provided a rich source of information on how social factors may influence code quality and community health [20], [26]. Previous studies show that developers values and motivations have an effect on communication quality and on task completion in a project, namely the collaborative values [25]. Competitive values are on the other hand related to a higher effort put in the projects [1]. OSS contributors are specific with their motivations, as they hold collaborative values over individual ones, but in comparison with other open-source contributors – e.g., Wikipedia's contributors – they are more focused on self-enhancement than on altruism [18], [25]. We investigate whether there is a difference between females contributing to OSS and those who do not with respect to their feeling of competition and feeling of kinship with other developers. In our previous study [4] there was no apparent difference in these two feelings between the overall OSS and the overall closed-source developers populations.

Self-Education Activity. Lastly, we include a self-education activity dimension. We expect female developers who are more proactive in self-education and more learning-oriented to be more likely to contribute to OSS. This finding is supported by our previous research for the overall developers population on the Stack Overflow Survey data [4]. The proactive personality and learning orientation is related to multiple positive outcomes, such as job performance and career success [15], [21]. In the development context, Software Engineers with Proactive Personality perform better in innovative tasks [24].

III. Research Methodology

In this section, we describe the goal and methodological steps we followed in our study.

A. Research Question

As shown in the Section II, source code quality, developer's interaction, productivity, and research data quality are dependent on the people developing the software [17], [20], [26]. Earlier work suggests that there exist differences between developers who are present in OSS environment and those outside of it. One of these differences is the even greater under-representation of women in OSS, when compared to a closed-source setting [3], [19], [4]. Therefore, we examine if this group of developers is well represented or not, as current research sometimes draws conclusions from surveys and research that has a very low women representation in the sample [3]. Hence, we ask:

RQ$_1$: *Are there differences between women in OSS and women in closed-source software development?*

Based on literature, we assess whether differences in personal characteristics influencing source code quality and software development exist between women contributing to OSS and those not contributing. Identifying these differences may be important in multiple contexts, such as:

- To understand whether women are represented equally in OSS and non-OSS setting.
- To identify what are the differences between women in OSS and closed-source environment.
- To confirm the representativeness of the data collected about women in software development through mining of software repositories.

Describing this data about women may indicate (i) how the code and the interaction of developers is different between OSS and non-OSS setting, (ii) what kind of women are better represented in OSS and potentially more attracted to contribute, and (iii) whether the data about women collected in MSR research could be affected by these differences.

We test the null hypotheses of no difference between the two groups in the proposed dimensions. The methodology of the study is described in further detail in the following section.

B. Subjects

We used openly available data from the Stack Overflow Developer Survey 2018 that counts more than 100,000 respondents and is the most widely spread survey of demographics and other characteristics of developers and their work [19]. Stack Overflow users are a recognized and commonly used population to be surveyed and investigated as to draw conclusions about the OSS environment (e.g., [2], [14]). Gender was a multiple choice question and there was an option allowing developers not to share the information. For the clarity of our analysis we included only the respondents who identify themselves solely as women. The final amount of analyzed responses was 3,436, as we removed responses for which there are missing values in the considered dimensions.

TABLE I
COMPARISON OF WOMEN CONTRIBUTING TO OSS AGAINST WOMEN NOT CONTRIBUTING TO OSS

	OSS		**non-OSS**		**Total**		**Chi Square Test**		
	N	% of OSS	N	% of non-OSS	N	% of total	Sig.	Effect Size	Meaning
All	1,025	100.00	2,411	100.00	3,436	100.00			
Developer Type									
- Developer	923	90.05	2,183	90.54	3,106	90.40	0.653		
- Data Science	218	21.27	402	16.67	620	18.04	<0.01	-0.046	negligible
- DevOps	281	27.41	466	19.33	747	21.74	<0.001	-0.081	negligible
- Student	124	12.10	319	13.23	443	12.89	0.336		
Education							<0.05		
- Primary	5	0.49	6	0.25	11	0.32	0.257		
- Secondary	43	4.20	112	4.65	155	4.51	0.561		
- College Dropout	85	8.29	182	7.55	267	7.77	0.456		
- Associate Degree	31	3.02	82	3.40	113	3.29	0.571		
- Bachelor Degree	523	51.02	1,327	55.04	1,850	53.84	<0.05	0.240	small
- Master Degree	283	27.60	624	25.88	907	26.40	0.293		
- Professional Degree	12	1.17	28	1.16	40	1.16	0.981		
- Ph.D.	43	4.20	50	2.07	93	2.70	<0.001	-0.271	small
	OSS		**non-OSS**		**Total**		**Independent T Test**		
	Mean	SD	Mean	SD	Mean	SD	Sig.	Effect Size	Meaning
Age	3.05	0.85	2.98	0.81	3.00	0.82	<0.01	-0.055	negligible
Experience	3.59	2.26	3.19	2.09	3.31	2.15	<0.001	-0.109	negligible
Competition	2.8	1.17	2.83	1.14	2.82	1.15	0.512		
Kinship	3.84	0.89	3.75	0.86	3.77	0.87	<0.01	-0.062	negligible
Competence	2.56	1.13	2.86	1.14	2.77	1.14	<0.001	0.144	negligible
Self-Education	2.44	1.34	2.24	1.2	2.3	1.25	<0.001	-0.082	negligible

C. Data Analysis

We aimed to identify the differences between women who contribute to OSS either voluntarily or as part of their job, and women who do not. For that we compared the two groups using Chi-Square Test in case of Formal Education and Developer Type and T-test for independent groups in case of Age, Experience, Feeling of Kinship and Competition, Perceived Competence and Self-Education Activity. We present Cliff's Delta effect sizes where relevant.

D. Threats to Validity

Our study analyzed openly available data from Stack Overflow Developer Survey 2018 [19]. This data set might not be representative of the developers community. Though, it is the most representative survey so far and multiple sources confirm its relevance [8], [28].

We have addressed similar dimensions of differences as in our previous research, where we compared the differences between OSS and non-OSS developers of all genders. In this study we used a subsample of the same dataset and the comparison of results is not independent. Even though we were analyzing the same dataset, we did not observe the same differences, thus giving us more confidence to assume that the results are not affected by one another.

IV. ANALYSIS OF THE RESULTS

Our final sample consisted of 3,436 women distributed in OSS (i.e., the respondents who reported to contribute to Open Source, N = 1,025) and non-OSS (i.e., the respondents who reported to not contribute to Open Source, N = 2,411) samples, showing twice as much women are present in non-OSS setting. As depicted in Table I, the only characteristics yielding a significant difference with at least small effect size are Bachelor and PhD level of education. Women who reported to have achieved a Ph.D. are more represented among OSS contributors (4% of OSS women vs. 2% in non-OSS) and women with Bachelor degree are represented more in closed-source environment (51% vs. 55%). The other types of education did not show a significant difference.

The Developer Type Student and Developer are not significantly different between OSS and non-OSS setting. Even though type DevOps and Data Scientist are significantly different, the effect size is negligible. It is interesting to note that Developer Type is a variable where the developers could choose multiple options while filling out the survey and women in OSS fall more frequently in multiple categories (amount of responses 50% higher than amount of respondents) than other women (40% higher).

Age, experience, perceived competence, feeling of kinship and self-education activity were significantly different, but the effect sizes were negligible as well. Feeling of competition is not significantly different between the OSS and non-OSS population of women.

V. DISCUSSION

The results of this study are surprising when compared to previous research. Firstly, in our previous study on the same dataset, we have found differences between OSS and non-OSS developers [4]. These effects were not replicated for the female sample.

The absence of previously observed differences between developers is hard to explain. As previously stated, the main difference between OSS and non-OSS groups is the greater under-representation of women [19], [4]. Nevertheless, we observed almost no differences in their characteristics, meaning that these factors do not seem to be related to female OSS contribution and the low OSS participation of women remains unexplained.

This lack of differences is particularly interesting for perceived competence where an even bigger difference in the female group were to be expected, according to previous research [31]. However, we did not find a difference in perceived competence between OSS and non-OSS women. Furthermore, women have in general less coding experience than men [19] and OSS contributors have more experience than non-OSS ones [4]. We analyzed whether there is difference between women in OSS and in non-OSS and we found that the populations seem to be comparable.

The positive conclusion of this initial exploration for the research community is that the women investigated through OSS may be highly representative of female developers in general.

VI. Conclusion

There is a research interest in investigating developers community and gender differences within. However, women are under-represented in the tech industry and in research samples as well. This might have an effect on the conclusions researchers draw from available data.

OSS data is a commonly used data source. For that reason, we have investigated for potential differences in the female population contributing and not contributing to OSS, aiming to identify how the research data and conclusions drawn from the OSS contact might be biased, and to identify which women are more attracted to contribute to OSS. Based on previous literature we expected some differences to appear, however the examined samples seem to be well representative of one another. The only identified differences lie in higher portion of women with Ph.D. and lower portion of those with Bachelor degree in the OSS community. The main difference remains in the even higher under-representation of women in OSS. Future research is needed to determine this difference and to identify the opportunities to attract women to contribute to OSS.

References

[1] S. O. Alexander Hars. Working for free? motivations for participating in open-source projects. *International Journal of Electronic Commerce*, 6(3):25–39, 2002.

[2] S. Baltes and S. Diehl. Usage and attribution of stack overflow code snippets in github projects. *arXiv preprint arXiv:1802.02938*, 2018.

[3] J. Bitzer and I. Geishecker. Who contributes voluntarily to oss? an investigation among german it employees. *Research policy*, 39(1):165–172, 2010.

[4] BLINDED. How is oss representative? characterizing open-source software contributors. In *Under submission*, 2019.

[5] E. Capra, C. Francalanci, and F. Merlo. An empirical study on the relationship between software design quality, development effort and governance in open source projects. *IEEE Transactions on Software Engineering*, 34(6):765–782, 2008.

[6] C. Casalnuovo, B. Vasilescu, P. Devanbu, and V. Filkov. Developer onboarding in github: the role of prior social links and language experience. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*, pages 817–828. ACM, 2015.

[7] G. Catolino, F. Palomba, D. A. Tamburri, A. Serebrenik, and F. Ferrucci. Gender diversity and women in software teams: How do they affect community smells? In *41st International Conference on Software Engineering, Software Engineering in Society*, 2018.

[8] C. Chen, Z. Xing, and Y. Liu. What's spain's paris? mining analogical libraries from q&a discussions. *Empirical Software Engineering*, pages 1–40, 2018.

[9] M. D'Ambros, M. Lanza, and R. Robbes. An extensive comparison of bug prediction approaches. In *Mining Software Repositories (MSR), 2010 7th IEEE Working Conference on*, pages 31–41. IEEE, 2010.

[10] D. Di Nucci, F. Palomba, G. De Rosa, G. Bavota, R. Oliveto, and A. De Lucia. A developer centered bug prediction model. *IEEE Transactions on Software Engineering*, 2017.

[11] N. Ducheneaut. Socialization in an open source software community: A socio-technical analysis. *Computer Supported Cooperative Work (CSCW)*, 14(4):323–368, 2005.

[12] J. Eyolfson, L. Tan, and P. Lam. Do time of day and developer experience affect commit bugginess? In *Proceedings of the 8th Working Conference on Mining Software Repositories*, pages 153–162. ACM, 2011.

[13] G. Fleishman. Cartoon captures spirit of the internet. *The New York Times*, 14:2000, 2000.

[14] D. Ford, K. Lustig, J. Banks, and C. Parnin. We don't do that here: How collaborative editing with mentors improves engagement in social q&a communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 608. ACM, 2018.

[15] B. Fuller Jr and L. E. Marler. Change driven by nature: A meta-analytic review of the proactive personality literature. *Journal of Vocational Behavior*, 75(3):329–345, 2009.

[16] A. E. Hassan. The road ahead for mining software repositories. In *Frontiers of Software Maintenance, 2008. FoSM 2008.*, pages 48–57. IEEE, 2008.

[17] H. Muccini, D. A. Tamburri, and V. S. Rekha. On the social dimensions of architectural decisions. In *European Conference on Software Architecture*, pages 137–145. Springer, 2015.

[18] S. Oreg and O. Nov. Exploring motivations for contributing to open source initiatives: The roles of contribution context and personal values. *Computers in human behavior*, 24(5):2055–2073, 2008.

[19] S. Overflow. Stack overflow annual developer survey, 2018.

[20] F. Palomba, D. A. A. Tamburri, F. A. Fontana, R. Oliveto, A. Zaidman, and A. Serebrenik. Beyond technical aspects: How do community smells influence the intensity of code smells? *IEEE Transactions on Software Engineering*, 2018.

[21] P. R. Pintrich. The role of goal orientation in self-regulated learning. In *Handbook of self-regulation*, pages 451–502. Elsevier, 2000.

[22] E. M. Pomerantz, E. R. Altermatt, and J. L. Saxon. Making the grade but feeling distressed: Gender differences in academic performance and internal distress. *Journal of Educational Psychology*, 94(2):396, 2002.

[23] J. A. Roberts, I.-H. Hann, and S. A. Slaughter. Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the apache projects. *Management science*, 52(7):984–999, 2006.

[24] N. Rodrigues and T. Rebelo. Incremental validity of proactive personality over the big five for predicting job performance of software engineers in an innovative context. *Revista de Psicología del Trabajo y de las Organizaciones*, 29:21–27, 2013.

[25] K. J. Stewart and S. Gosain. The impact of ideology on effectiveness in open source software development teams. *Mis Quarterly*, pages 291–314, 2006.

[26] D. A. Tamburri, P. Kruchten, P. Lago, and H. Van Vliet. Social debt in software engineering: insights from industry. *Journal of Internet Services and Applications*, 6(1):10, 2015.

[27] J. Terrell, A. Kofink, J. Middleton, C. Rainear, E. Murphy-Hill, C. Parnin, and J. Stallings. Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science*, 3:e111, 2017.

[28] B. Vasilescu, V. Filkov, and A. Serebrenik. Stackoverflow and github: Associations between software development and crowdsourced knowledge. In *Social computing (SocialCom), 2013 international conference on*, pages 188–195. IEEE, 2013.

[29] B. Vasilescu, V. Filkov, and A. Serebrenik. Perceptions of diversity on github: A user survey. In *Proceedings of the Eighth International Workshop on Cooperative and Human Aspects of Software Engineering*, pages 50–56. IEEE Press, 2015.

[30] B. Vasilescu, D. Posnett, B. Ray, M. G. van den Brand, A. Serebrenik, P. Devanbu, and V. Filkov. Gender and tenure diversity in github teams. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3789–3798. ACM, 2015.

[31] Z. Wang, Y. Wang, and D. Redmiles. Competence-confidence gap: A threat to female developers' contribution on github. In *ICSE-SEIS'18. Proceedings of 40th International Conference on Software Engineering: Software Track*, page 10, 2018.