Gender diversity and women in software teams

Citation for published version (APA):
Catolino, G., Palomba, F., Tamburri, D. A., Serebrenik, A., & Ferrucci, F. (2019). Gender diversity and women in software teams: how do they affect community smells? In *Proceedings - 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Society, ICSE-SEIS 2019* (pp. 11-20). [8797636] ACM/IEEE. https://doi.org/10.1109/ICSE-SEIS.2019.00010

DOI:
10.1109/ICSE-SEIS.2019.00010

Document status and date:
Published: 28/05/2019

Document Version:
Accepted manuscript including changes made at the peer-review stage

Gender Diversity and Women in Software Teams: How Do They Affect Community Smells?

Gemma Catolino[1], Fabio Palomba[2], Damian A. Tamburri[3,4], Alexander Serebrenik[4], Filomena Ferrucci[1]
[1]University of Salerno, Italy, [2]University of Zurich, Switzerland, [3]Jeronimus Academy of Data Science, The Netherlands,
[4]Eindhoven University of Technology, The Netherlands
gcatolino@unisa.it, palomba@ifi.uzh.ch, d.a.tamburri@tue.nl, a.serebrenik@tue.nl, fferrucci@unisa.it

***Abstract*—As social as software engineers are, there is a known and established gender imbalance in our community structures, regardless of their open- or closed-source nature. To shed light on the actual benefits of achieving such balance, this empirical study looks into the relations between such balance and the occurrence of *community smells*, that is, sub-optimal circumstances and patterns across the software organizational structure. Examples of community smells are Organizational Silo effects (overly disconnected sub-groups) or Lone Wolves (defiant community members). Results indicate that the presence of women generally reduces the amount of community smells. We conclude that women are instrumental to reducing community smells in software development teams.**

***Index Terms*—Gender Balance; Community Smells; Software Organizational Structures; Empirical Study**

I. Introduction

Software development is an inherently social activity, and as such, the way developers communicate and collaborate can be expected to, and has been found to affect software quality [1], [2]. Several recent studies have focused on the so-called "community smells", patterns indicating suboptimal organization and communication of software development teams that can lead to unforeseen project costs [3]. Community smells have been also linked to code smells, indications of poor design, coding, and implementation choices [4], [5].

While community smells have been studied and described both for open-source projects and closed-source projects, little is known about how the *composition* of a software team affects the presence of community smells, and in particular, about the impact of gender diversity on them. Indeed, women have been reported to show the socio-emotional behavior in an organizational structure [6] and discourage conflict [7], increasing team efficiency [7] and mediating organizational quality [8], [9]. Specifically in the software engineering context, communication is a crucial factor in project success as reported by architects of a feminine expertise [10], while more gender-diverse teams have been shown to be more productive than less gender-diverse ones [11].

Based on such findings, we conjecture that a more diverse composition of development teams impacts the number of emerging community smells. Should this be verified, it would have a direct implication for project managers with respect to how to allocate resources in a way to reduce community-related problems, as well as on how to indirectly improving also the quality of the product being developed [4], [5].

We seek therefore to verify our conjecture, and assess the role of gender diversity and women participation on community smells. More specifically, in this study we consider four of the community smells defined and operationalized by Tamburri et al. [3], [4], i.e., *Organizational Silo*, *Black Cloud*, *Lone Wolf*, and *Radio Silence*. All of them refer to suboptimal characteristics of a development community that might be mitigated by the presence of women.

As original and novel contribution of this study, we build a statistical model relating gender balance (as reflected by the well-known Blau-Index [12]) as well as the *number of women in a team* to the aforementioned community smells. Moreover, we control for a wide array of socio-technical and social-networks factors of influence (e.g., the well-known *socio-technical congruence* [13]), in order to verify whether the variables of interest remain statistically significant also in presence of such additional factors.

The results of our study show that the expected relations between gender balance and participation of women as mediators against the proliferation of community smells are valid in the cases of *Black Cloud* and *Radio Silence*, while we found only partial relations between the variables of interest and *Organizational Silo* and *Lone Wolf*. For example, as expected the role of women as mediators for increased organizational quality is not equal across all community smells. In particular, their role is increasingly important for those smells which affect the quantities and qualities of communication across the organizational structure (e.g., the *Radio Silence* effect) as opposed to the organizational arrangement of the structure (e.g., the *Organizational Silo* effect). These results indicate that the presence of women plays a beneficial role in *mitigating* strains faced by complex organizations. Finally, the results encourage further the study into the empirical relation between gender balance and the complex relations constituting software engineering organizational structures.

Reproducibility. In order to enable the *full* replication of our study, we provide a comprehensive package containing all data and scripts used as additional contribution [14].

Structure of the paper. In Section II we discuss the research methodology while in Section III we report the results of the study. Section IV examines the threats to the validity of the study and the way we mitigated them. Section V overviews the literature related to the role of women and gender balance in complex and, more specifically, in software engineering organizations. Finally, Section VI concludes the paper and provides insights on our future research agenda.

II. Research Methodology

The *goal* of the study is to assess to what extent gender diversity and participation of women in software communities influence the health of such communities, with the *purpose* of providing a deeper understanding on the way mixed software teams communicate and cooperate. The *perspective* is that of project managers, who are interested in how to effectively allocate resources or manage complex organizational structures.

A. Hypothesis and Research Question

This paper seeks to shed light over the role of gender balance within software teams and their socio-technical success. Thus, the working hypothesis behind this work is stated as follows:

> *The presence of women within the team improves communication, co-operation, and collaboration, thus reducing the number of community smells.*

We expect the number of community smells to be reduced based on evidence on the role of gender balance (and women, in general [8], [9]) as a mediator for organizational quality. The existing evidence does not, however, take into account software development or virtual teams, and as such providing empirical support to it or refuting it, requires a separate study.

More specifically, the work in this article is focused around establishing the relation between the constructs in the hypothesis above and community smells, as indicators of general organizational quality. Consequently, the article addresses the following research questions:

RQ$_1$. *How does the number of community smells differ in teams without women and in teams with women?*

RQ$_2$. *To what extent does the presence of women within teams influence the number of community smells?*

The first research question intents to be a preliminary analysis. Indeed, should the distribution of community smells be similar in the two sets of projects, this would mean that the presence of women does not impact the phenomenon of interest. On the contrary, an eventual difference would highlight that the projects in the two sets differ for some aspects and, therefore, it might make sense to further investigate which are the factors (if any) influencing the number of community smells arising within a community. The second research question aims to provide a closer look into the role of gender diversity and participation of women on the number of community smells.

B. Context of Study

The *context* of the study was represented by software communities and community smells. As for the context, we have to distinguish the communities considered when addressing **RQ$_1$** and **RQ$_2$**: in the former case we needed to compare non-gender-diverse versus gender-diverse projects, while in the latter we only needed to consider gender-diverse projects. Thus, starting from the dataset made available by Vasilescu et al. [15]—which reports information on open-source teams and corresponding members of 23,493 projects—we first randomly selected 20 projects whose development teams were composed by only men;[1] on the other hand, we randomly selected the same amount of projects (20) composed of mixed teams, i.e., number of women higher than 0. It is worth noting that, as team composition tends to change in time [16], [17], for each project the dataset records information about one or more quarters data on the composition, characteristics, and outcomes of their teams of contributors. We limited our analyses to 20 communities for each type due to the computation of community smells.

Table I: Projects used in our study

Projects	Progr. Languages	# Windows
Akretion	Python	6
Bigcheese	C++	1
Burke	Go	5
Chapuni	C++	9
Cloudfoundry	Shell	7
CTSRD-CHERI	C++	2
Django	Python	23
Emberjs	Python	7
Fangism	C++	1
Genome	Perl	5
Holman	C	7
Jedi4ever	Shell	8
Jrk	C++	1
Liferay	Java	12
Loganchien	C++	1
Moodle	PHP	14
Mozilla - gecko-dev	C++	1
Mozilla - OpenBadger	Javascript	2
Mxcube	Python	2
Puppetlabs	Ruby	14
RobbyRussell	Python	15
Rspec	Ruby	13
Symfony	Python	13
Torvalds	C	17
Travis-ci	Javascript	10

For the sake of space limits, we only report in Table I the communities whose teams are gender-diverse. Specifically, for each of them, we report information on (i) name, (ii) programming language of the system, and (iii) number of time windows analyzed in our study. More information about the non-gender-diverse teams are available in our online appendix [14]. As for the set of community smells, previous research defined more than 10 different patterns possibly leading to the emergence of social debt [18]. From this list, we decided to focus on four of them, *Organizational Silo*, *Black Cloud*, *Lone Wolf*, and *Radio Silence*, as they have characteristics for which the presence of women might consistently affect their emergence. More specifically:

Organizational Silo Effect. This smell appears in cases where a community presents siloed areas that essentially

[1] Note that we did not find projects characterized by only women.

do not communicate with each other, except through one or two of their respective members. The *Organization Silo* reflects rigid thinking, narrow tunnel-vision, as well as lack of community-wide communication. The presence of women (according to Razavian and Lago [10]) should be able to break down the barriers forming across community siloes, thus reducing (at least linearly) the amount of organizational silo effects manifesting in a certain community structure.

Black Cloud Effect. Software communities are affected by this smell in cases where there is an excessive information overload due to the lack of structured communication. As the *Black Cloud* effect corresponds to the manifested presence of overwhelming quantities of asynch and synch data exchanges across a community, we expect that the presence of more diversity mediates positively on the presence of black clouds; furthermore, it might have an effect with respect to the tenor of communications (i.e., more positive and focused as opposed to nastier and wider) [4].

Lone Wolf Effect. This smell appears when there are unsanctioned or defiant contributors who carry out their work irrespective or regardless of their peers. With their role of increasing connectedness, presence of women is expected to have a positive influence, i.e., reduce the amount of reported “lone wolves"; at the same time, however, we expect that this mitigating effect manifests only if women have had a direct involvement in the process of designing and developing the involved software. Indeed, lone wolves often manifest concomitant with components or software modules that were explicitly modularized for co-operation (contemporary operation over software artifacts) as opposed to collaboration between peers (joint work on shared artifacts) [19].

Radio Silence Effect. This is an instance of the “unique boundary spanner" [20] problem from social-networks analysis: one member interposes herself into every formal interaction across two or more sub-communities with little or no flexibility to introduce other parallel channels. Promotion-rates for women to cover boundary-spanning roles have previously been discovered and established, and thus we do expect their role to mediate for the presence of such a community smell.

It is worth noting that the considered smells can be grouped into two high-level categories, i.e., structural- and communication-based. While *Organizational Silo* and *Black-cloud* affect the overall community structure, *Lone Wolf* and *Radio Silence* are concerned with the way the community members communicate with each other. Thus, the selection of those smells allows us to understand what is the role of gender diversity and the presence of women on problems having two different levels of granularity.

C. Detecting Community Smells

The first step to address our research question consisted in the detection of the selected community smells. To do so, we exploited the CODEFACE4SMELLS tool, a fork of CODEFACE [21] designed to identify developers' communities. Starting from the developer networks built by CODEFACE, we detected instances of the considered smells according to the formalizations of Palomba et al. [4].

Organizational Silo. Let $G_m = (V_m, E_m)$ be the communication graph of a project and $G_c = (V_c, E_c)$ its collaboration graph. The set of *Organizational Silo* pairs S is defined as the set of developers that do not directly or indirectly communicate with each other, i.e., $\{(v_1, v_2) | v_1, v_2 \in V_c, (v_1, v_2) \notin E_m^*\}$, where E_m^* is the transitive closure of E_m.

Black-Cloud. Detection of the *Black cloud* smells starts with the identification of vertex clusters as already implemented in CODEFACE. Let $P = \{p_1, \ldots, p_k\}$ be a mutually exclusive and completely exhaustive partition of V_m induced by the clustering algorithm. Then, *Black Cloud* is the set of pairs of developers C that connect otherwise isolated sub-communities, i.e., $\{(v_1, v_2) | v_1, v_2 \in V_m, (v_1, v_2) \in E_m, \forall i, j(((v_1 \in p_i \wedge v_2 \in p_j) \Rightarrow i \neq j) \wedge \forall v_x, v_y((v_x \in p_i \wedge v_y \in p_j \wedge (v_x, v_y) \in E_m) \Rightarrow v_x = v_1 \wedge v_y = v_2))\}$.

Lone Wolf. Starting from the vertex clusters of the community network, the set of *Lone Wolf* pairs L is defined as the set of collaborators that do not directly or indirectly communicate with each other $\{(v_1, v_2) | v_1, v_2 \in V_c, (v_1, v_2) \in E_c, (v_1, v_2) \notin E_m^*\}$.

Radio Silence. Finally, the *Radio Silence* set B is the set of developers interposing themselves into every interaction across two or more sub communities. Developer can interpose themselves into interactions if either they are the only member of their cluster or they communicate with a member of the different cluster, and they are the only member of their cluster communicating with this different cluster: $\{v | v \in V_m, \exists i(v \in p_i \wedge \forall v_x(v_x \in p_i \Rightarrow v = v_x))\} \cup \{v | v \in V_m, \exists v_x, i, j(v \in p_i \wedge v_x \in p_j \wedge (v, v_x) \in E_m \wedge \forall v_y, v_z((v_y \in p_i \wedge v_z \in p_j \wedge (v_y, v_z) \in E_m) \Rightarrow v_y = v)\}$.

The detection techniques above were also evaluated in order to assess their actual ability to identify community smells. Specifically, we ran CODEFACE4SMELLS on 60 open-source projects and, through a survey study, we asked the original developers of such systems whether the results given by the tool actually reflect the presence of issues within the community. As a result, we discovered that the recommendations of the tool highlight real community-related problems. Furthermore, it should be noted that the effectiveness of the operationalisations above rely on the proven effectiveness of the approach by Joblin et al. [21], building upon the “Order Statistics Local Optimization Method" (OSLOM) [22] featured inside CODEFACE, which was never previously applied before on developer networks. Further operationalisation and evaluation details are discussed in the accompanying technical report [23].

*D. **RQ**$_1$ - Distribution of Community Smells in Teams without Women and in Teams with Women*

Once we had completed the detection of community smells for each of the time windows considered, we addressed **RQ**$_1$ by verifying the distribution of community smells in the two sets of projects considered, i.e., non-gender-diverse versus gender-diverse ones. In particular, we computed statistical metrics such as minimum, maximum, mean, median, first, and third quartile of the distribution of community smells in the two groups with the aim of investigating whether non-gender-diverse teams present a higher number of community smells with respect to gender-diverse ones. We also statistically assessed the differences observed with the Mann-Whitney test [24]. The results are intended as statistically significant at α = 0.05. We also measure the effect size with Cliff's Delta (d) [25]. To interpret the results, we followed the well-established guidelines: small for $|d| < 0.10$, small for $0.10 \leq |d| < 0.33$, medium for $0.33 \leq |d| < 0.474$ and large for d $\geq$ 0.474.

*E. **RQ**$_2$ - Building a Statistical Model*

To address **RQ**$_2$, we defined a statistical model relating the set of community metrics present in the dataset exploited [11] to the detected smells. The following subsections detail the definition of the modeling approach.

Independent Variables. Based on our hypothesis, we considered two factors as independent variables, namely:

- *Number of women in a team*. This information is computed as the difference between the total number of community members and the number of men belonging to the community. The exploited dataset [11] already contained this information;
- *Blau-Index [12]*. This index is a well-established diversity measure for categorical variables; It is defined as $1 - \sum_{i \in (m,f)} p_i^2$, where p_i is the fraction of male and female team members. Also in this case, the metric was already available in the considered dataset [11].

In the original dataset gender has been inferred based on personal names and, if available, countries, of the team members using the `genderComputer` tool [26]. The tool combines a number of transformations, such as diminutive and 1337 resolution, heuristics (e.g., users from Russia with surnames ending in -ova are female), and female/male frequency name lists collected for thirty different countries.

Response Variables. We aimed at understanding the impact of women and gender diversity on the presence of community smells. Thus, our response variable was represented by the number of community smells as identified by CODEFACE4SMELLS in each of the time windows of the considered projects. Note that we took into account the number of smells rather than a boolean value reporting their presence/absence because this response variable better fitted our hypothesis: indeed, we wanted to analyze whether our independent variables could mitigate the amount of problems appearing in the community rather than their presence.

Control Variables. While our hypothesis pertained to the relationship between women participation and gender diversity on the presence of community smells, it is important to note that many other community-related factors might have an influence on the dependent variable. To account for this, we consider the following control variables:

- *Number of Committers*: Theoretically speaking, the amount of committers in a community could represent a factor to consider when analyzing the number of community smells. Indeed, it might be possible that the higher the number of people in the community, the higher the likelihood that a smell appears. Thus, we considered the total committer count of a project as first control variable.
- *Number of Commits*: A high number of commits indicates a high activity in the community. Such an activity might influence the way the community communicates and evolves, possibly impacting the amount of community smells. This metric aimed at considering this aspect, and was computed as the total commit count in a project.
- *Team size*: The size of the community can clearly have an influence on the number of community smells it contains, i.e., it might be possible that the higher the size the higher the presence of smells. This metric is defined as a number of contributors per project team in a given quarter. It is worth noting that team size differs from the number of committers as it accounts for the number of members in a certain quarter, while number of committers represents the cumulative amount of people having at least one commit in the entire history of the considered projects.
- *Turnover*: This is defined as the fraction of the team in a given quarter that is different with respect to previous quarter (i.e., the turnover ratio). A high turnover indicates that community members are frequently changed: the constant introduction of new contributors might lead to the emergence of communication/coordination issues, possibly affecting the number of community smells.
- *Project age*: This is the difference between the maximum index and the index of the 90-day interval of the first commit. We controlled for changes in environment as GITHUB grows with time: later projects and their teams may have experienced a different culture, and this change might affect the presence of community smells.
- *Tenure diversity*: The experience of the team members possibly influence their ability to communicate and collaborate with other members of the community, thus influencing the emergence of problems. Thus, we took into account for this aspect by considering two metrics such as commit and project tenure. The first metric measures the coding experience of a contributor within GITHUB: this is computed as the number of days since the earliest ever recorded commit of contributor in a certain quarter (in any GITHUB repository) until the end of that quarter. The second metric computes the experience of a contributor in the context of the individual considered project: it is represented by the number of quarters since the earliest recorded event of a contributor in the current project until the end of that quarter. It is worth remarking that commit and project tenure are orthogonal measures: indeed, a development team might be diverse when considering commit tenure in cases where it has a mix of expert coders and newcomers, but homogeneous with respect to project tenure, if all contributors joined the team at the same time. As tenure measures are numerical variables,

the exploited dataset report them using the coefficient of variation, defined as the ratio between the adjusted standard deviation and the mean.

- *Tenure median:* To represent an average project or commit tenure as opposed to tenure diversity, we also included the project median tenure and the commit median tenure.

While the metrics described above were already available in the dataset provided by Vasilescu et al. [11], we also considered further community-related aspects such as:

- *Socio-Technical Congruence:* This metric represents "the state in which a software development organization harbors sufficient coordination capabilities to meet the coordination demands of the technical products under development" [13] and operationalized in this study as the number of development collaborations that do communicate over the total number of collaboration links present in the collaboration network.
- *Truck-Factor:* The truck factor measures the minimum number of members of a community that have to quit before the project will fail [27]–[29]. In our work, we operationalized truck factor based on core and peripheral community structures identified by CodeFace4smells, i.e., as the degree of ability of the community to remain connected without its core part. Further details on how core and periphery members are determined can be found in the work of Joblin et al. [21].
- *Centrality:* Centralization [30] is a social network analysis methodology that calculates the graph-level centrality score based on node-level centrality measures. Similarly, within this work we considered modularity measures [31] to measure the strength of a community structure: high modularity usually indicates that there exist a clear definition and distinction of sub-communities within the considered network and as the modularity tend to zero it indicates that there are no sub-communities within the considered network; social networks analysis literature suggests a threshold of 0.3 [31] over which a community is considered highly modular and thus with a clear distinction of the sub-communities present in its development network.

All the metrics above have been shown to influence community health [13], [32] and, therefore, might have all an influence on the number of community smells.

Statistical Models Contructions. Based on the independent, response, and control variables described above, we then built four statistical models, i.e., one for each community smell considered. Since the exploited dataset was composed of multiple time windows for each project, we had to build mixed-effects models [33], [34] in order to capture measurements from within the same group (i.e., within the same project) as a random effect. In our case, we used the time window as random effect. All other variables were modeled as fixed effects. We used multiple linear mixed-effects models [34], implementing the functions `lmer` and `lmer.test` available in the `R` package `lme4`.[2] With the aim of correctly interpreting the achieved results, we took into account the problem of multi-collinearity [35], which appears in cases where two or more variables are highly correlated and can be predicted one from the other, possibly biasing the way the model fits and the results interpreted. In this regard, we employed a stepwise variable removal procedure based on the Companion Applied Regression (`car`) `R` package,[3] and in particular based on the `vif` (variance inflation factors) function [35]. This method provides an index for each independent variable that measures how much the variance of an estimated regression coefficient is increased because of collinearity. The square root of the variance inflation factor indicates how much larger the standard error is, compared with what it would be if that variable were uncorrelated with the other predictor variables in the model. Based on this information, we could understand which metric produced the largest standard error, thus allowing the identification of the metric that was better to drop from the model. We also removed outliers, as they can seriously threat both results and assumptions behind statistical models [36]. To this aim, we employed the ROMR.FNC function, which is available in the `LMERConvenienceFunction` `R` package.[4] Finally, coefficients were considered important if they were statistically significant (ρ <0.05). Their effect sizes were obtained using the ANOVA statistical test [37].

III. Analysis of the Results

This section outlines results of our empirical study, showcasing the discussion of each research question individually.

Table II: Distribution of community smells over teams without women (NW) and teams with women (W) projects.

Group	Min	1st Q.	Mean	Median	3rd Q.	Max
NW	1	4	7	6	11	24
W	1	2	3	3	5	10

A. RQ_1 - Distribution of Community Smells in Teams without Women and in Teams with Women

Table II reports the results achieved when addressing $\mathbf{RQ}_1$. As it is possible to observe, development teams which do not include women (NW in the table) present a number of community smells generally higher than teams that include women (W). More in detail, the mean for community smells is 7 in the former case, while it is 3 in the latter one. The results are confirmed when looking at all the other indicators, and especially the maximum: this is 24 in the case of non-gender-diverse teams and 11 when considering gender-diverse ones. Our results are confirmed from a statistical perspective: the difference between the distributions are statistically significant (p-value < 0.05) and with a large effect size (d=0.68). It is important to note that we control our results for team size

[2] https://cran.r-project.org/web/packages/lme4/lme4.pdf

[3] https://cran.r-project.org/web/packages/car/index.html

[4] https://cran.r-project.org/web/packages/LMERConvenienceFunctions/index.html

(i.e., we verified whether the observed differences were just a reflection of the development team size): we observed that the W and NW teams are statistically indistinguishable with respect to this factor, meaning that the development size did not influence our observations. These findings seem to indicate that *there are factors within gender-diverse teams that influence the number of community smells*. This finding motivates the analysis conducted in **RQ**$_2$: indeed, on equal terms, there seem to be some factors influencing the distribution. Our study further analyzes whether gender diversity and women participation are among such factors.

> **Summary for RQ**$_1$**:** Non-gender diverse teams have a statistically higher number of community smells with respect to gender-diverse teams. This indicates the presence of factors within gender-diverse teams that possibly influence the presence of community smells.

B. *RQ*$_2$ *- Evaluating the Statistical Model*

In this section, we report the results that address our **RQ**$_2$. To better analyze and interpret our findings, we compared the model we built with two baseline statistical models: the first one containing all the control factors but not the independent variables; the second one only containing the random effect. On the one hand, the comparison with the former model allowed us to understand how the exploited control variables explain the number of community smells independently from gender diversity and women participation; on the other hand, the comparison with the latter model allowed us to understand whether the results were only a reflection of the random effect. In the following, we discuss the achieved results by presenting each community smell individually.

Radio Silence. Table III reports on the relations captured in a statistical model revolving around the *Radio Silence* effect. Unsurprisingly, software community members' turnover is a powerful mediator over the occurrence of the target community smell as implicitly revealed by Homscheid et al. [38]. More toward our study, the presence of women is indeed a powerful mediator with an estimate of 0.038 and a standard error of 0.012 with significance below 0.01. Both strong mediators are followed swiftly by socio-technical congruence and age, which assume mildly mediating roles. In the comparison with the baselines, we observed that the devised model has both a higher AIC (Akaike information criterion [39]) and BIC (Bayesian information criterion [39]), meaning that it can better explain the response variable: we then conclude that, in this case, the number of women represents a relevant factor to explain the number of *Radio Silence* instances affecting a community.

Organizational Silo. Table IV reports on the relations captured in a statistical model revolving around the *Organizational Silo* effect, for which the role of diversity and the presence of women is much less prominent than expected, playing a mild mediating role at best, with an estimate achieved at 0.039 and standard error at 0.021, with <0.1 significance. In other words, it seems that the reduction of the number of sub-communities that do not communicate with each other can be only partially explained by the existence of mixed teams. This was also confirmed when comparing the model with the baselines: indeed, our model achieved AIC and BIC values equal with respect to the one that only includes control factors, meaning that the addition of the independent variables does not necessarily help in better explaining the response variable.

Lone Wolf. Table V reports on the relations captured in a statistical model revolving around the *Lone Wolf* effect, for which the role of diversity is practically non-existent as mediator while the presence of women is much less prominent than expected, playing mediating role similar to what was previously reported in the scope of organizational silo effects, this time with an estimate achieved at 0.039 and standard error at 0.027, with <0.1 significance. Interestingly, team size and socio-technical congruence are the most relevant factors, both when considering the model that includes the independent variables and the one that does not include them. From a practical perspective, this result indicates that largest and poorly coordinated teams tend to exhibit a higher number of lone wolfs, and the presence of women can only partially mediate from their emergence. In this case, AIC and BIC of the devised model are slightly higher than the baselines, and this confirms our observations on the partial role of women on the number of *Lone Wolf* instances arising in software communities.

Black Cloud. Table VI reports on the relations captured in a statistical model revolving around the *Black Cloud* effect, for which our expectations and results are definitively confirmed: while the presence of women *per se* cannot explain the number of this smell type, our results show that the Blau-index, which denotes diversity, has *the strongest* estimate, achieved at -8.226 and standard error at 2.306, with <0.001 significance. Thus, we conclude that having a diverse team significantly help in reducing the number of *Black Cloud* instances. This is also confirmed when comparing the model with the baselines: AIC and BIC are higher for our model, and this indicates the strong ability of the Blau-index to explain the response variable.

> **Summary for RQ**$_2$**:** Gender diversity and women participation are relevant factors for *Black Cloud* and *Radio Silence*, while we found only partial relations between the variables of interest and *Organizational Silo* and *Lone Wolf.*

IV. Discussion and Threats to Validity

Below we discuss our main findings and provide an overview of the threats that might have influenced our conclusions.

A. *Discussion*

The role of gender diversity as a factor to consider for team-building and governance is still relatively unclear, especially with respect to the extent to which gender diversity changes or influences the software organizational structure. In this work, we strived to add to such clarity by means of quantitative empirical research.

Table III: Results achieved by the three models built for Radio Silence. If the values for a certain variable are not presented, it means that the *vif* function removed it from the model.

Factor	All Variables			Conf. Variables			Random		
	Estimate	S.E	Sig.	Estimate	S.E	Sig.	Estimate	S.E	Sig.
(Intercept)	-2.260	0.635		-2.743	0.624		2.335	0.128	
log(totalCommitters)	-0.049	0.085		-0.072	0.080				
log(totalcommits)	0.027	0.060		0.022	0.061				
projectAge	-0.025	0.015	.	-0.018	0.015				
turnover	10.048	0.455	***	10.119	0.458	***			
blauGender	0.194	0.804							
tenureMedian	0.044	0.035		0.034	0.035				
tenureDiversity	0.010	0.025		0.007	0.025				
stCongruence	-0.308	0.186	.	-0.265	0.187				
centrality	-0.038	0.146		-0.046	0.146				
log(teamSize)				0.195	0.068	**			
truckFactor	-0.014	0.046		-0.009	0.046				
numberOfWomen	0.038	0.012	**						
***p < 0.001 - **p < 0.01 - *p < 0.05 .p < 0.1									

Table IV: Results achieved by the three models built for Organization Silo. If the values for a certain variable are not presented, it means that the *vif* function removed it from the model.

Factor	All Variables			Conf. Variables			Random		
	Estimate	S.E	Sig.	Estimate	S.E	Sig.	Estimate	S.E	Sig.
(Intercept)	0.911	1.290		-1.626	1.226		2.342	0.122	
log(totalCommitters)	0.383	0.157	*	0.294	0.151	.			
log(totalcommits)	-0.154	0.117		-0.119	0.117				
projectAge	-0.022	0.027		-0.012	0.027				
turnover	-1.459	0.538	**	-1.183	0.518	*			
blauGender	2.740	1.769							
tenureMedian	-0.087	0.065		-0.094	0.066				
tenureDiversity	0.011	0.046		0.017	0.047				
stCongruence	0.379	0.343		0.417	0.348				
centrality	0.084	0.269		0.135	0.272				
log(teamSize)	0.202	0.146		0.124	0.132				
truckFactor	0.059	0.086		0.077	0.086				
numberOfWomen	0.039	0.021	.						
***p < 0.001 - **p < 0.01 - *p < 0.05 .p < 0.1									

Table V: Results achieved by the three models built for Lone Wolf. If the values for a certain variable are not presented, it means that the *vif* function removed it from the model.

Factor	All Variables			Conf. Variables			Random		
	Estimate	S.E	Sig.	Estimate	S.E	Sig.	Estimate	S.E	Sig.
(Intercept)	-8.462	4.112		-8.679	4.123		4.456	0.180	
log(totalCommitters)	0.267	0.289		0.414	0.262				
log(totalcommits)	0.121	0.219		0.086	0.217				
projectAge	0.089	0.051	.	0.068	0.049				
turnover	0.253	1.002		-0.193	0.929				
blauGender	-3.376	2.939							
tenureMedian	-0.031	0.121		-0.0313	0.122				
tenureDiversity	0.078	0.087		0.069	0.086				
stCongruence	-8.298	1.511	***	-8.465	1.510	***			
centrality	0.084	0.269		0.135	0.272				
log(teamSize)	4.315	0.892	***	4.250	0.893	***			
truckFactor	-0.168	0.159		-0.177	0.159				
numberOfWomen	0.013	0.038	.						
***p < 0.001 - **p < 0.01 - *p < 0.05 .p < 0.1									

A first key discussion point reflects our finding that the presence of women, rather than the Blau-index, exhibits a statistically significant relation with most community smells (with the exception of the Black Cloud smell); given the characteristics of our sample, this significance indicates that *even when outnumbered women can act as mediators against the proliferation of specific community smells*. Given the beneficial effects, we reported with respect to the considered smells, community leaders may want to encourage the participation of women including involvement in mediatory, and boundary-spanning roles across the organizational structure.

Secondly, focusing on the presence of women as mediators for community smells, we reported a considerably diverse effect across all community smells. The diversity of teams appears much more statistically significant for the *Black Cloud* community smell, which is connected to information overload; diversity in this instance could play a mediatory role forming more stable and well-coordinated organizational structures. Similarly, the more prominent role of women is beneficial to reduce a somewhat opposite condition, i.e., the *Radio Silence* smell, connected to miscommunicating sub-communities.

Third, our study confirms the role of established socio-technical quality metrics (e.g., socio-technical congruence, truck factor, etc.) as mediators for community smells. The relation between such factors and the presence/absence of women or, in general, gender diversity has not been studied,

Table VI: Results achieved by the three models built for Black Cloud. If the values for a certain variable are not presented, it means that the *vif* function removed it from the model.

Factor	All Variables			Conf. Variables			Random		
	Estimate	S.E	Sig.	Estimate	S.E	Sig.	Estimate	S.E	Sig.
(Intercept)	4.172	1.698		2.420	1.630		4.456	0.180	
log(totalCommitters)	-0.191	0.203		0.062	0.201				
log(totalcommits)	0.087	0.152		-0.0341	0.155				
projectAge	-0.017	0.035		-0.042	0.036				
turnover	0.794	0.702		-0.032	0.688				
blauGender	-8.226	2.306	***						
tenureMedian	0.060	0.084		0.057	0.088				
tenureDiversity	0.015	0.060		-0.009	0.062				
stCongruence	-0.179	0.444		-0.275	0.462				
centrality	-0.601	0.351	.	-0.572	0.361				
log(teamSize)	0.394	0.190	*	0.691	0.175	***			
truckFactor	0.009	0.111		0.003	0.115				
numberOfWomen	0.039	0.027							
****p < 0.001 - **p < 0.01 - *p < 0.05 .p < 0.1*									

yet: whether the collaboration (e.g., the co-commit networks [30]) and communication structures (e.g., the mailing-list networks [21]) are congruent could itself be connected to more gender balanced organizational structures and the connected beneficial organizational condition. In the scope of our results, we observed that these aspects could be closely connected, but further research is needed to look deeper into the relation.

B. Threats to Validity

We describe the threats to validity and limitations of our work, as posed by our methods and data.

Threats to Construct Validity. As for the dataset exploited, we relied on a publicly available source previously built [11]; most of the independent and control variables used in our models belong to it. Of course, we cannot exclude possible imprecisions in the computation of such variables. As for the detection of community smells, we exploited the CODEFACE4SMELLS tool. It is important to note that the tool was also evaluated in order to assess its actual ability to identify community smells [23]: the results of the evaluation make us confident in the precision with which the response variable has been computed. Finally, regarding the definition of teams without women (NW) and teams with women (W) within the project, we are aware that we simplified gender as a binary variable (for instance, there are developers who do not identify themselves as women or men). Moreover, the GENDERCOMPUTER tool, in some cases, may return the value of "unknown", indicating its inability in detecting the gender of the contributor either due to limitations of the tool or due to the conscious decision of the contributor not to reveal their gender. Importance of not misindentifying women was highlighted by Lin and Serebrenik [40]. In cases we analyzed a team consisting of men and unknowns, this was considered as NW, i.e., if the tool does not explicitly say that a woman is involved, then it is classified as NW.

Threats to Conclusion Validity. A major threat to the conclusions drawn is related to the statistical methods employed. To ensure that the selected linear mixed model was appropriate for the available data, we first investigated how similar studies performed their analyses [15], [26], [41], [42]. Moreover, since the dataset was composed of multiple time windows for each project, we built mixed-effects models [33], [34] in order to capture measurements from within the same group as a random effect, as recommended by literature [33], [34]. Afterwards, to ensure that the experimented model did not suffer from multi-collinearity, we adopted the *variance inflation factors* function [35] to discard non-relevant variables from the considered features, setting the threshold to 5 as suggested by O'Brien [35]. In addition, we discarded outliers to not bias our interpretations [36]. Finally, to statistically verify the significance of the model variables, we employed ANOVA [37], a widely recognized efficient technique to interpret the results of statistical models.

Threats to External Validity. Threats in this category mainly concern the generalization of results. We analyzed a total of 40 software systems coming from different application domains and having different characteristics (size, programming languages, number of classes, etc.). Of course, we cannot claim the generalizability, however part of our future research agenda is to extend the study with more different set of systems.

V. RELATED WORK

Team formation represents the creation of the social networks of individuals partaking the team, the properties of such an invisible structure as well as the mediators that influence both structure and behavior [43]. From this perspective, in the context of software engineering, little is known about what are the key performance indicators behind team formation, let alone organizational structure quality [44]. At the same time, however, organizations and social-networks research literature present early works towards establishing the role of gender diversity as a mediating factor against work-conflicts or sub-optimal organizational behavior. For example, Randel [45] established that numerical distinctiveness of gender in group composition triggers the *salience* [46] of group members' gender identities for men in the group and, in turn, identity salience affects work group conflict. The aforementioned gender-related problem was relatively latent in the software engineering literature until Huff [47] made explicit reference to it in terms of occupational equity for software designers. More recently, the latest edition of the flagship software engineering conference organized an explicit workshop to account for more discussion around gender imbalance and equity; in that venue, Reeves [48] made an attempt at circumscribing and elaborating more on the phenomenon; a most related effort is found in Mahmod et al. [49] who discuss gender inequality in open-source and link it qualitatively to the potential lack of organizational innovation, a phenomenon known as organizational inertia [50].

The role of team formation [51] as a success driver for software engineering efforts has been studied from several perspectives [3], [11], [18], [43]–[50], [52]–[56]. Farhangian et al. [52] investigated the phenomenon in self-assembling teams typically occurring in open-source commons. Similarly, team formation dynamics and composition has been studied extensively from the perspective of software engineering education [53], where the formation of teams needs to be factored into the educational structure and an equal balance of collaboration. In the context of open source systems, other study tried to analyzed the diversity on a team [11], [54]–[56]. Daniel et al. [54] investigated the effects of diversity on community engagement and market success for 357 SourceForge projects. As a result, they found that developers with different reputation and role influence the market success and community engagement, while diversity of spoken language and nationality influence in negative manner the community engagement, but in positive the market success. Chen et al. [55] analyzed how the diversity of experience and interests, within WIKIPEDIA Projects, influence the productivity and turnover; results, indeed, showed how different interest increases productivity and decreases member withdrawal. The same study was conducted by Wang et al. [56] that showed how different tenure of developers than the over-all group tenure influence the productivity. Finally, Vasilescu et al. [11] analyzed the diversity of team from a gender diversity perspective: the authors have studied more than 2 million GITHUB projects, built a linear mixed model and showed how both gender and tenure diversity influence the productivity in open source systems.

In a different work, Vasilescu et al. [26] have observed higher participation and activity levels of women in mailing lists of open source projects as opposed to more competitive environment of STACK OVERFLOW, and Qiu et al. [57] have studied the impact of participation in open teams wrt diversity of ties and information on the chance of prolonged engagement of women in open source projects. Terrell et al. [42] have observed gender-related bias in pull request acceptance. Ford et al. [58] have identified barriers to Stack Overflow use for females, while Mendez et al. [59] have identified aspects of open source tools and infrastructure such as GITHUB inducing gender bias in open source barriers to entry previously identified by Steinmacher et al. [60]. Finally, Robles et al. [61] have conducted a large-scale study of women in Open Source Software, and Izquierdo-Cortazar et al. [62] have recently reported on gender participation in Open Stack, a large open source ecosystem. Going beyond the specifics of software engineering, the broader topic of gender and information technology has been extensively studied by feminist scholarship [63], [64].

Similarly to the work of Vasilescu [11], in this article, we wanted to understand the extent to which gender (im-)balance in software project teams may influence, in our case, the emergence of circumstances within the organizational structure known as *community smells* [3], [18].

VI. CONCLUSIONS

This article reports on empirical evidence to shed light over the connections between the presence of women and the manifestation of community smells. Results indicate a considerable and consistent role as mediators for the presence of woman professionals with respect to the occurrence of community smells, however, such presence seems to mediate more strongly with respect to those smells which affect the quality of communication (e.g., *Radio Silence* effect) as opposed to how the organizational structure is arranged or the quantities involved (e.g., *Organizational Silo* effect).

The impact of our results is connected to the effective composition of teams as a mitigation against the proliferation of nasty community smells across complex software organizations. Furthermore, our results confirm the generally-established positive effect of gender balance in software teams, but not equally for all organizational circumstances.

In the future, we plan to increase the types of smells with which we operated our experimentation to understand further the role of women for organizational quality. In addition, we aim to look deeper into the technical benefits of gender balance along with riddance of community smells.

REFERENCES

[1] M. Pinzger, N. Nagappan, and B. Murphy, "Can developer-module networks predict failures?" in *FSE*. ACM, 2008, pp. 2–12.

[2] N. Bettenburg and A. E. Hassan, "Studying the impact of social structures on software quality," in *ICPC*. IEEE, 2010, pp. 124–133.

[3] D. A. Tamburri, P. Kruchten, P. Lago, and H. van Vliet, "Social debt in software engineering: insights from industry." *J. Internet Services and Applications*, vol. 6, no. 1, pp. 10:1–10:17, 2015.

[4] F. Palomba, D. 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.

[5] D. A. Tamburri and E. Di Nitto, "When software architecture leads to social debt." in *WICSA*. IEEE, 2015, pp. 61–64.

[6] R. Croson and U. Gneezy, "Gender differences in preferences," *Journal of Economic Literature*, vol. 47, no. 2, pp. 448–474, 2009.

[7] S. G. Rogelberg and S. M. Rumery, "Gender diversity, team decision quality, time on task, and interpersonal cohesion," *Small Group Research*, vol. 27, no. 1, pp. 79–90, 1996.

[8] P. Araújo-Pinzón, C. Álvarez Dardet, J. M. Ramón-Jerónimo, and R. Flórez-López, "Women and inter-organizational boundary spanning: A way into upper management?" *European Research on Management and Business Economics*, vol. 23, no. 2, pp. 70–81, 2017.

[9] H. Ibarra, "Network centrality, power, and innovation involvement: Determinants of technical and administrative roles," *The Academy of Management Journal*, vol. 36, no. 3, pp. 471–501, Jun. 1993.

[10] M. Razavian and P. Lago, "Feminine expertise in architecting teams." *IEEE Software*, vol. 33, no. 4, pp. 64–71, 2016.

[11] B. Vasilescu, D. Posnett, B. Ray, M. G. J. van den Brand, A. Serebrenik, P. T. Devanbu, and V. Filkov, "Gender and tenure diversity in github teams," in *CHI*. ACM, 2015, pp. 3789–3798.

[12] P. M. Blau, *Inequality and heterogeneity: A primitive theory of social structure*. Free Press New York, 1977, vol. 7.

[13] G. Valetto, M. Helander, K. Ehrlich, S. Chulani, M. Wegman, and C. Williams, "Using software repositories to investigate socio-technical congruence in development projects," in *MSR*, 2007, pp. 25:1–25:4.

[14] G. Catolino, F. Palomba, D. A. Tamburri, A. Serebrenik, and F. Ferrucci. (2018) Gender diversity and women in software teams: How do they affect community smells? - replication package - https://figshare.com/s/a144c4bcf3839952477b.

[15] B. Vasilescu, A. Serebrenik, and V. Filkov, "A data set for social diversity studies of github teams," in *MSR*, 2015, pp. 514–517.

[16] S. Augustine, B. Payne, F. Sencindiver, and S. Woodcock, "Agile project management: steering from the edges," *Communications of the ACM*, vol. 48, no. 12, pp. 85–89, 2005.

[17] A. Fuggetta and E. Di Nitto, "Software process," in *Proceedings of the on Future of Software Engineering*. ACM, 2014, pp. 1–12.

[18] D. A. Tamburri, R. Kazman, and H. Fahimi, "The architect's role in community shepherding." *IEEE Software*, vol. 33, no. 6, pp. 70–79, 2016.

[19] O. E. Danzell and L. M. M. Montañez, "Understanding the lone wolf terror phenomena: assessing current profiles," *Behavioral Sciences of Terrorism and Political Aggression*, vol. 8, no. 2, pp. 135–159, 2016.

[20] S. Wasserman and K. Faust, *Social Network Analysis. Methods and Applications.* Cambridge University Press, 1994.

[21] M. Joblin, W. Mauerer, S. Apel, J. Siegmund, and D. Riehle, "From developer networks to verified communities: A fine-grained approach." in *ICSE*, A. Bertolino, G. Canfora, and S. G. Elbaum, Eds., 2015, pp. 563–573.

[22] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," *PLoS ONE*, vol. 6, no. 4, p. e18961, 04 2011.

[23] S. Magnoni, "An approach to measure community smells in software development communities," 2016, politecnico di Milano, Italy.

[24] W. J. Conover and W. J. Conover, *Practical nonparametric statistics*. Wiley New York, 1980.

[25] R. J. Grissom and J. J. Kim, *Effect sizes for research: A broad practical approach.* Lawrence Erlbaum Associates Publishers, 2005.

[26] B. Vasilescu, A. Capiluppi, and A. Serebrenik, "Gender, representation and online participation: A quantitative study," *Interacting with Computers*, vol. 26, no. 5, pp. 488–511, 2014.

[27] L. Williams and R. R. Kessler, *Pair Programming Illuminated.* Addison Wesley, 2003.

[28] G. Avelino, L. T. Passos, A. C. Hora, and M. T. Valente, "A novel approach for estimating truck factors," in *ICPC*, 2016, pp. 1–10.

[29] M. M. Ferreira, M. T. Valente, and K. A. M. Ferreira, "A comparison of three algorithms for computing truck factors," in *ICPC*, 2017, pp. 207–217.

[30] A. Meneely, L. Williams, W. Snipes, and J. A. Osborne, "Predicting failures with developer networks and social network analysis." in *FSE*. ACM, 2008, pp. 13–23.

[31] J.-P. Hatala and J. G. Lutta, "Managing information sharing within an organizational setting: A social network perspective," *Performance Improvement Quarterly*, vol. 21, no. 4, pp. 5–33, 2009.

[32] D. A. Tamburri, F. Palomba, A. Serebrenik, and A. Zaidman, "Discovering community patterns in open-source: A systematic approach and its evaluation," *Empirical Software Engineering*, 2018.

[33] D. Bates, Mächler, B. Bolker, S. Walker *et al.*, "lme4: Linear mixed-effects models using eigen and s4," *R package version*, vol. 1, no. 7, pp. 1–23, 2014.

[34] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[35] R. M. O'Brien, "A caution regarding rules of thumb for variance inflation factors," *Quality & quantity*, vol. 41, no. 5, pp. 673–690, 2007.

[36] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, vol. 22, no. 2, pp. 85–126, 2004.

[37] A. Cuevas, M. Febrero, and R. Fraiman, "An anova test for functional data," *Computational statistics & data analysis*, vol. 47, no. 1, pp. 111–122, 2004.

[38] D. Homscheid and M. Schaarschmidt, "Between organization and community: Investigating turnover intention factors of firm-sponsored open source software developers," in *Proceedings of the 8th International ACM Web Science Conference*, 2016.

[39] K. P. Burnham and D. R. Anderson, "Multimodel inference: understanding aic and bic in model selection," *Sociological methods & research*, vol. 33, no. 2, pp. 261–304, 2004.

[40] B. Lin and A. Serebrenik, "Recognizing gender of stack overflow users," in *MSR*, M. Kim, R. Robbes, and C. Bird, Eds. ACM, 2016, pp. 425–429.

[41] B. Winter, "A very basic tutorial for performing linear mixed effects analyses," *arXiv preprint arXiv:1308.5499*, 2013.

[42] 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 Comp Sci*, vol. 3, p. e111, 2017.

[43] P. Rathod and M. des Jardins, "Stable team formation among self-interested agents," in *AAAI Workshop on Forming and Maintaing Coalitions in Adaptive Multiagent Systems*, 2004.

[44] D. A. Tamburri, P. Lago, and H. van Vliet, "Organizational social structures for software engineering." *ACM Comput. Surv.*, vol. 46, no. 1, p. 3, 2013.

[45] E. R. Amy, "Identity salience: a moderator of the relationship between group gender composition and work group conflict," *Journal of Organizational Behavior*, vol. 23, no. 6, pp. 749–766, 2002.

[46] N. Röper, "Confirmatory information processing in group decision making: the impact of group composition, identity salience, time pressure, and accountability," Ph.D. dissertation, University of Regensburg, Germany, 2014.

[47] C. Huff, "Gender, software design, and occupational equity," *ACM SIGCSE Bulletin*, vol. 34, no. 2, pp. 112–115, 2002.

[48] J. Reeves, "Gender equality in software engineering." in *GE@ICSE*. ACM, 2018, pp. 33–36.

[49] M. Mahmod, S. Affendi Mohd Yusof, and Z. M. Dahalin, "Where are the female developers? Exploring the gender issues in open source software innovation process," in *Knowldege Management International Conference*, 2010, pp. 555–560.

[50] M. T. Hannan and J. H. Freeman, "Structural inertia and organizational change," *American Sociological Review*, vol. 49, no. 2, pp. 149–164, Apr. 1984.

[51] C. M. Karapicak and O. Demirors, "A case study on the need to consider personality types for software team formation." in *SPICE*, ser. Communications in Computer and Information Science, vol. 349. Springer, 2013, pp. 120–129.

[52] M. Farhangian, M. K. Purvis, M. Purvis, and B. T. R. Savarimuthu, "Modeling team formation in self-assembling software development teams: (extended abstract)." in *AAMAS*. ACM, 2016, pp. 1319–1320.

[53] V. Yannibelli and A. Amandi, "A memetic algorithm for collaborative learning team formation in the context of software engineering courses." in *ADNTIIC*, ser. LNCS, vol. 7547. Springer, 2011, pp. 92–103.

[54] S. Daniel, R. Agarwal, and K. J. Stewart, "The effects of diversity in global, distributed collectives: A study of open source project success," *Information Systems Research*, vol. 24, no. 2, pp. 312–333, 2013.

[55] J. Chen, Y. Ren, and J. Riedl, "The effects of diversity on group productivity and member withdrawal in online volunteer groups," in *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, 2010, pp. 821–830.

[56] L. S. Wang, J. Chen, Y. Ren, and J. Riedl, "Searching for the goldilocks zone: trade-offs in managing online volunteer groups," in *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 2012, pp. 989–998.

[57] H. S. Qiu, A. Nolte, A. Brown, A. Serebrenik, and B. Vasilescu, "Going farther together: The impact of social capital on sustained participation in open source," in *ICSE*. IEEE, 2019, pp. xx–xx.

[58] D. Ford, J. Smith, P. J. Guo, and C. Parnin, "Paradise unplugged: identifying barriers for female participation on stack overflow," in *FSE*. ACM, 2016, pp. 846–857.

[59] C. J. Mendez, H. S. Padala, Z. Steine-Hanson, C. Hilderbrand, A. Horvath, C. Hill, L. Simpson, N. Patil, A. Sarma, and M. M. Burnett, "Open source barriers to entry, revisited: a sociotechnical perspective," in *ICSE*. ACM, 2018, pp. 1004–1015.

[60] I. Steinmacher, T. Conte, M. A. Gerosa, and D. F. Redmiles, "Social barriers faced by newcomers placing their first contribution in open source software projects," in *CSCW*, 2015, pp. 1379–1392.

[61] G. Robles, L. Arjona Reina, J. M. González-Barahona, and S. D. Domínguez, "Women in free/libre/open source software: The situation in the 2010s," in *OSS*, 2016, pp. 163–173.

[62] D. Izquierdo-Cortazar, N. Huesman, A. Serebrenik, and G. Robles, "Openstack gender diversity report," *IEEE Software*, 2019.

[63] L. Van Zoonen, "Feminist theory and information technology," *Media, Culture and Society*, vol. 14, no. 1, pp. 12–35, 1992.

[64] S. Turkle, *The Second Self: Computers and the Human Spirit*. The MIT Press, 2005.