Going farther together

Going Farther Together: The Impact of Social Capital on Sustained Participation in Open Source

Huilian Sophie Qiu
Carnegie Mellon Univ.
hsqq@cmu.edu

Alexander Nolte
University of Tartu
alexander.nolte@ut.ee

Anita Brown
Bryn Mawr College
anitab1046@gmail.com

Alexander Serebrenik
Eindhoven Univ. of Tech.
a.serebrenik@tue.nl

Bogdan Vasilescu
Carnegie Mellon Univ.
vasilescu@cmu.edu

***Abstract*—Sustained participation by contributors in open-source software is critical to the survival of open-source projects and can provide career advancement benefits to individual contributors. However, not all contributors reap the benefits of open-source participation fully, with prior work showing that women are particularly underrepresented and at higher risk of disengagement. While many barriers to participation in open-source have been documented in the literature, relatively little is known about how the social networks that open-source contributors form impact their chances of long-term engagement. In this paper we report on a mixed-methods empirical study of the role of social capital (i.e., the resources people can gain from their social connections) for sustained participation by women and men in open-source GitHub projects. After combining survival analysis on a large, longitudinal data set with insights derived from a user survey, we confirm that while social capital is beneficial for prolonged engagement for both genders, women are at disadvantage in teams lacking diversity in expertise.**

I. Introduction

Sustained participation by contributors in open source software (OSS) is critical to the survival of OSS projects [1], [2], and it can provide many benefits to individual contributors [3]. For example, a recent survey [4] found that OSS work helped more than half of the respondents obtain their current positions, and that OSS work in general helps people build their professional reputation. Given the advantage that open source experience can bring to an individual and the benefit that sustained participation can provide to OSS projects, it is essential to study what retains or repels contributors.

Not surprisingly, sustained participation in OSS has attracted considerable attention among researchers, with prior work focusing on developers' motivation [1], [5], [6], the kind of tasks they perform [7], [8] , and rejection experiences [9]–[13]. However, the benefits that contributors can gain from their OSS social relations and structures have not been studied. Such benefits are known in the social sciences as *social capital* [14], [15]. Social capital can be built through individuals' social networks and has been shown to affect various kinds of human endeavors, from knowledge sharing [16] to labor force participation [17] and from philanthropy [18] to financial development [19]. In OSS, studies have shown that prior social ties can influence forming or joining a new team [20], [21]. However, they did not explore whether and how social ties can prolong contributors' participation.

While social capital can be built and leveraged by everyone, it can impact women differently in male-dominated environments. For example, prior work in the film industry [22] found that while men benefit from strongly connected networks, women do not; moreover, women benefit from diversity in teams and tasks. In OSS, women are severely underrepresented and, as we show, likely to disengage from GitHub participation earlier than men (Figure 1).

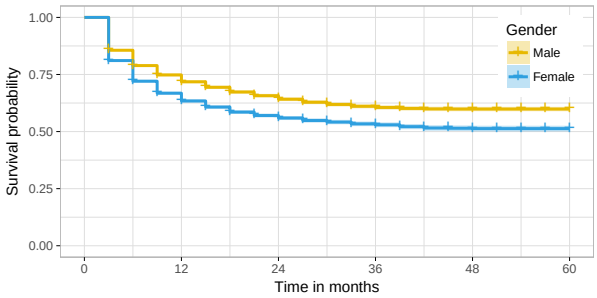

Fig. 1: Kaplan-Meier estimators: women disengage significantly earlier. (chi-sq= 645, p$< 2e^{-16}$ per a log-rank test)

To better understand contributors' disengagement, we perform a longitudinal, quantitative analysis of the structure of OSS contributors' social networks on GitHub and the impact of this structure on prolonged engagement, through the lens of social capital theory. Moreover we report on a user survey to better understand what constitutes social capital for GitHub open source contributors and how it is associated with their participation sustainability. Our findings highlight that:

- Contributing to projects where team members are more familiar with each other (from prior collaborations) is in general associated with decreased risk of disengagement;
- Women are at higher risk of disengagement than men.
- Higher team diversity along dimensions of programming language expertise is associated with a decreased risk of both short and long term disengagement. *Moreover*, gender and language diversity interact: when team members have more diverse programming language backgrounds, women are less likely than men to disengage early.

Our results have implications for project choosing, team formation, and project management in OSS. Based on our results, we especially recommend that women take project social capital and expertise diversity into consideration when choosing a project to join, and that project managers consider these aspects when allocating developers to tasks, in more centrally managed contexts. We also argue that social coding

platforms like GITHUB could benefit from recommendation engines for newcomers looking for projects to join; these should take social capital into account when making a recommendation (cf. [23]); furthermore, GITHUB could facilitate project maintainers tracking trends in factors negatively associated with the development of social capital, particularly among women.

II. DEVELOPMENT OF HYPOTHESES

We build on *social capital theory*, a popular social sciences theory used to explain individual and group success and performance (for an overview see Adler and Kwon [24]). *Social capital* is the set of benefits individuals can gain from their social connections and social structures, such as access to information and emotional support [24]; it is a complement to human capital, which refers to an individual's ability [15].

OSS is a social environment that can be modeled as collaborative social networks [25], where social capital can form: projects are community-based in nature; contributors have ample opportunities to connect with each other by interacting and collaborating over time; they agree on common norms; and they share collective goals—the development and maintenance of OSS. Once present, social capital can "make individuals' experiences of working on open source projects both satisfying and rewarding" [26]. In this paper we argue that *social capital also impacts the overall open source tenure of contributors*, and that female and male contributors benefit from social capital differently, on average.

There are two main network structures conducive of social capital: strong, dense, and cohesive ties generate *bonding social capital* [27], while weakly connected ties, acting as brokers between subgroups, generate *bridging social capital* [15].

The first, *bonding social capital*, emerges from *network closure*, *i.e.*, strongly connected ties [27]. Tie strength increases with the amount of interaction between individuals, emotional density, intimacy, or reciprocal service [28]. In a closed network, information is passed more accurately through direct communication [29], and trust develops more easily since it is more expensive for people to break norms when actions are more easily noticed [27]. At the same time, network closure increases group cohesiveness and solidarity among group members, who become more likely to remain engaged.

In OSS, contributors are motivated by both intrinsic and extrinsic factors, among which aspects related to bonding social capital, such as identifying with the community and feeling obligated to contribute back, are highly important [6]. Prior work showed how identification, obligation, emotional attachment, trust relationships, and shared goals and norms (all of which are more likely to develop in cohesive teams [30]) positively impact individual and team outcomes. It follows that bonding social capital should positively impact the contributors' willingness to sustain their OSS activity. In OSS participants are often free to disengage at any time, therefore the extent to which they have a sense of social identity, or perceive themselves to be part of the community, may substantially increase their intention to continue [31], [32].

In contrast to bonding social capital, *bridging social capital* focuses on how network individuals who maintain weak ties can benefit from a brokerage position [15]. In closed networks people who are strongly connected may have the same information or the same source of information. Bridging otherwise disconnected groups, what Burt calls structural holes [15], can enable access to broader sources of information and improve the information's quality, relevance, and timeliness [24]. While bridging social capital is especially beneficial in competitive scenarios, when timely and non-redundant information about job opportunities can be an advantage, it can also be an asset in OSS. Weak ties can expose contributors to, *e.g.*, new technologies and new projects, providing opportunities to continue their engagement. Already, evidence suggests that past collaborative ties impact contributors' choice of OSS projects to participate in [21]. Network brokers can also decrease the centralization of OSS communities and increase communication between experts and peripheral users [33].

To summarize, network closure and structural holes, representing both types of social capital, seem important for sustained participation in open source. We expect that:

H$_1$. *During their open source tenure, the more often people participate in projects with high potential for building social capital, the higher their chance of prolonged engagement.*

However, network closure may not always be beneficial. As Lutter [22] notes "cohesive networks might foster discrimination and exclusion, as network closure is likely to divide [individuals] into insiders and outsiders". Outsiders, *i.e.*, those who are not part of the "core" group, can have a harder time accessing information, leading them to miss out on some chances [14], [34], [35]. Furthermore, people within a social group tend to develop their own habitus, often unconsciously. Such habitus embodies membership but also restricts outsiders from accessing and identifying with the group [36]–[38].

In OSS in general and GITHUB in particular, sociodemographic diversity is lower than anywhere else in tech [39]. Women are particularly underrepresented, with recent surveys placing them at less than 5% [40]; women are also more likely than men to encounter stereotyping or unwelcoming language [41]–[43]. However, as prior results from the film industry, a similarly male-dominated field, show, women can overcome the negative effects of network closure: being more often attached to open teams with regard to diversity of ties, information flow, and genre background increases chances of career survival [22]. That is, since women tend to be outsiders to the strongly connected groups of (mostly male) decision-makers, diversifying their ties makes them less dependent on the in-group for acceptance [44]. Therefore, given women's minority (and likely outsider) status in OSS in aggregate, we expect:

H$_2$. *During their open source tenure, the more often women participate in open teams wrt diversity of ties and information, the higher their chance of prolonged engagement.*

III. Related work

Discrimination exists in online software engineering communities and women are known to face greater barriers than men [45]. Terrell *et al.* show that women whose gender identities are revealed have lower pull request acceptance rate [43]. Mendez *et al.* have observed biases against women in GitHub tools and infrastructure [23], while Ford *et al.* identified barriers for female participation on Stack Overflow [46]. Social network analysis has also been applied to OSS [20], [21], [25], [47]–[51], although these studies did not consider gender.

Sustained participation, turnover and disengagement have attracted significant attention as well, *e.g.*, using qualitative methods, Fang *et al.* reveal that situated learning and identity construction are associated with sustained participation [1], while Lin *et al.* show that contributors who join the project earlier, write code instead of documents, or are responsible for modifying code have higher chances of remaining in the team [7]. The relation between turnover and project quality has been studied by Foucault *et al.* [52]. A complementary perspective has been taken by Zhou and Mockus that identified metrics such as number of comments and the size of the peers' groups as characteristics of new contributors that will become long-term contributors [53]. These conclusions, however, focused on individual behaviors and project qualities. In this paper, we analyze sustained participation from the perspective of contributors' social connections on GitHub.

IV. Methods

We designed a mixed-methods study characterized by a concurrent triangulation strategy [54] to help triangulate our findings. Quantitatively, we collected a multivariate longitudinal data of 58,091 GitHub contributors, and performed survival analysis to model the effects of social capital on disengagement. Qualitatively, we surveyed a sample of 88 contributors to gain additional insights into the role of social capital on GitHub.

A. Data

Our main data source is the February 2017 version of GHTorrent [55], a publicly available historical database of GitHub public activity traces, containing data for approximately 16M users. Gender is not recorded in GitHub profiles and, consequently, is also not available in GHTorrent. Therefore, we inferred it from people's names, as described in Section IV-B, and augmented the GHTorrent data. However, since social network analysis on a data set of GitHub's size would be computationally unfeasible, we first compiled a smaller sample of 58,091 users, as follows.

Preprocessing and Filtering. Starting from the $\sim$ 16M users in GHTorrent, we filtered out organizational users (*i.e.*, metausers, not usually corresponding to a single person), users with deleted accounts, users who never authored any commits and users with names not containing any space (gender inference techniques rely on a person's first and last names; *e.g.*, Alice would be excluded, but Alice Smith and Alice Marie Smith would not). We acknowledge that some cultures do not split names into parts, or some people are known mononymously. We chose this conservative heuristic, which excludes some valid names, since we noticed during manual exploration of the data that many single-part names are English words or nicknames from which we cannot extract gender information. Approximately 1.8M GitHub users in our data had non-organizational, non-deleted accounts, authored at least one commit, and had names consisting of at least two parts.

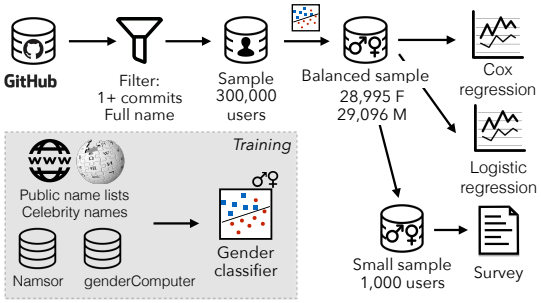

Fig. 2: Overview of our methodology.

Identity Merging. Since git version control settings are set locally by each client, there are some cases where git commits are not attributed to the correct GitHub account, which introduces noise in the data. Moreover, the same contributor may have used different git "aliases" (*i.e.*, names and emails) in different projects or over time [56]. To have a more accurate representation of one's activity and contributions, we performed identity merging on the different (name, email) tuples in our data using a series of heuristics (cf. [56]–[58]).

Sampling. After initial filtering and identity merging, we randomly sampled 300,000 users and applied our gender inference technique (Section IV-B) to label each account as Female (9.7%), Male (84.85%), or Unknown (5.45%). Some of our social network analysis measures (Section IV-C) require, for every person, to collect all the repositories they contributed to, and for every repository, to collect all other contributors and all *their* repositories. To reduce computational effort and to address the Female–Male imbalance in our sample, we randomly down-sampled the group of male contributors to the same size as the female group. After removing users who have only contributed to educational projects, our final dataset contains 28,995 users labeled Female and 29,096 users labeled Male. Figure 2 gives an overview of our data collection process.

B. Gender Inference

Various approaches and tools for name-based gender inference have been proposed [59], [60]. All operate with the simplifying assumption that gender is binary; we also assume binary gender here to simplify data collection and analysis. We tried many of these tools and found that each has strengths and blind spots. In particular, most tools are based on databases of English names and as such fail, *e.g.*, on Asian names.

We have considered approaches that use social network data, specifically *Google+* [43], but the gender API has been deprecated; tools that can infer gender from photos, *e.g.*, *Face++*, but discarded these since GitHub profile photos

TABLE I: Accuracy of the different gender inference methods (bolded are the highest accuracy for that language).

Language	genderComputer (%)	NamSor (%)	Our classifier (%)
Chinese	17.58	6.70	**60.00**
Japanese	76.76	26.88	**79.71**
Korean	18.82	13.51	**68.07**
All	79.41	74.07	**83.62**

are scarcely available; and tools that can infer gender from text [61], but discarded these since we have a very limited amount of text for each user – mostly commit messages, which are usually too short to provide enough information.

Instead, we identified two main contenders among tools that rely on broader datasets of names in different languages, and integrate them in a classifier (*i.e.*, a voting system). Our first contender is *genderComputer*[1] [62]. As opposed to other tools it uses location information to disambiguate; *e.g.*, it is able to distinguish between Italian Andrea (predominantly male) and German Andrea (predominantly female). Our second contender is *NamSor*[2] which classifies personal names by gender, country of origin, and ethnicity, with good coverage of different languages, countries, and regions. We trained and tested a Naive Bayes classifier that takes as input the gender predictions output by *genderComputer* and *NamSor* for a given name as well as features of the name itself, and produces a gender label as output, *i.e.*, one of Female, Male, or Unknown.

As training (80%) and test (20%) data, we compiled a list of 11,706 names from two sources. First, we randomly sampled 8,706 names from *genderComputer*'s open source dataset, which covers 28 countries. Second, since both input gender tools often have difficulty with East Asian names, we further collected a total of 3,000 romanized Chinese, Japanese, and Korean names from celebrity name lists on Wikipedia, websites for baby names, or name lists found in online public datasets, *e.g.*, lists of recent school graduates or of enrolment.

For each name, we obtained the gender inferences from *NamSor* and *genderComputer*. We also extracted features from the name itself, including the last character (*e.g.*, in Spanish, names ending in 'a' tend to be female), the last two characters (*e.g.*, in Japan, names ending in 'ko' tend to be female), and tri-grams and 4-grams to capture romanized Chinese, Japanese, and Korean names. We also included *NamSor*'s inference on the contributors' countries of origin from their last names as a feature. Using the country of national origin inferred from last names, instead of the country of residence declared on the GitHub profile, is an improvement on prior work, because it can increase the gender inference accuracy for people residing outside their (or their ancestors') country of origin, *e.g.*, Italian Andrea's living in the US. We note, however, that this approach can still fail in some cases, *e.g.*, for a person with a Chinese last name and a non-Chinese first name such as Andrea Zhang.

Table I reports the accuracy of the gender inference tools and our classifier overall as well as on names in East Asian languages, which are typically the hardest to make inferences

[1]https://github.com/tue-mdse/genderComputer

[2]http://www.namsor.com

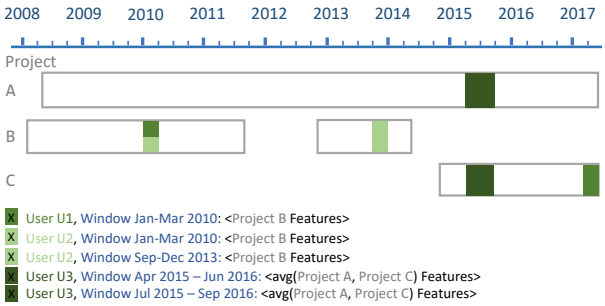

Fig. 3: Illustration of data points we collect.

on [60]. Overall, our combination classifier has higher accuracy on all categories of names than either *genderComputer* or *NamSor*. Our classifier fails mostly on gender neutral names, such as Robin and a Chinese name Yan that can be both male and female, depending on what Chinese character it is associated with. We also do not have enough training samples to make accurate inference from languages such as Burmese.

C. Operationalizations of Concepts

To model the effects of different dimensions of social capital on sustained participation on GitHub, our statistical modeling technique (survival analysis, Section IV-D) involves operationalizations of the different theoretical concepts discussed in Section II. We introduce the following operationalizations.

Panel Data. An implicit assumption for social capital effects to manifest is that project members had a chance to interact with each other. Since GitHub projects can be long-lived and since open-source projects in general face high turnover [7], [52], we assemble a longitudinal panel data set with measures computed over shorter time intervals; specifically, we aggregate all data from 2008 to 2016 into consecutive three-month windows, *i.e.*, we compute quarterly values for all measures.

Note that this involves two levels of aggregation. First, for every person and every project they contributed to, we compute quarterly values for different project-level measures (details below). Second, whenever someone contributed to more than one project in the same three-month window, thus having different sets of values for different projects in that window, we average out their project-level measures across their different projects that window; our results are qualitatively similar (significance and directionality of regression coefficients) if we compute the maximum instead of the average across projects. Figure 3 illustrates the structure of our data.

GitHub Disengagement—Outcome Variable. The dependent variable in our model is the occurrence of the disengagement event: *i.e.*, if every commit a person authors is an indication of repeated engagement, we consider a person's last recorded commit as an indication of disengagement if "long enough" time has elapsed for potential subsequent commits to be observable. Naturally, programmers may take a break from GitHub and return later for more contributions. Moreover, one's last recorded commit may be very close to the end of the observation period, so it is not clear whether they will return to contribute more; this common phenomenon in longitudinal

data is known as right censorship (the disengagement event did not happen during the course of study) [63].

We considered 12 months of inactivity as "long enough" to confidently detect disengagement, and used this operationalization in our survival models. Specifically, we consider that a GITHUB contributor has disengaged at time t if they have not committed anything to any open-source project for 12 months after t; *i.e.*, the *has_disengaged* value is 1 in the three-month window containing t, and 0 in all previous windows. Consequently, we also consider that people whose last recorded commit is less than 12 months prior to the end of our data are still active. Our models are robust to this operationalization and the results are qualitatively similar (significance and directionality of regression coefficients) with 6 months instead of 12. Note that we excluded 9,269 people with 12 months or more of inactivity that returned to make new contributions. Among them, 4,932 were male, 4,337 female.

Team Cohesion Measures. $\mathbf{H}_1$ assumes that during their open-source tenure, the more often GITHUB contributors participate in projects with high potential for building social capital, the higher their chance of prolonged engagement, *i.e.*, strongly connected networks and presence of ties between subgroups increase the likelihood of sustained participation in open-source. While subgroup or community detection has been extensively studied in the social network analysis literature [51], as argued by de Vaan *et al.* [64] these techniques are not suited for the operationalization of social capital constructs. Indeed, community detection techniques interpret ties as a static construct, while interpersonal relations, trust, and the implied social capital develop in time. Hence, to argue presence of a tie between two developers, the relationship between them should be durable, and this durability should be reflected in the operationalization. Therefore, as operationalizations for ties in team structures, we follow Lutter [22] and de Vaan *et al.* [64] and compute two distinct but related measures of social capital: interpersonal team familiarity and team recurring cohesion.

Team Familiarity. We adapt Newman's [65] measure of average interpersonal familiarity within a team, which captures the intensity of prior collaborations between each pair of current team members; the measure of strength of a developer's social connection to a project by Casalnuovo *et al.* [21] is conceptually similar. Team familiarity is aggregated over pairs of contributors (dyads), and as such it is capable of capturing both ties within subgroups and between subgroups, corresponding to bonding and bridging social capital.

To calculate dyadic interpersonal familiarity for project p in time window t, we iterate over all time windows prior to t. Let i and j be two contributors to project p and let and r_{is} and r_{js} be the sets of projects they worked on in time window s, respectively. The familiarity between i and j at time t is defined as the number of projects they worked on together in past windows $s < t$, adjusted by the team size of each project at that time, assuming that people who work in a smaller team are more familiar with each other. Only collaborative projects ($|r_s| > 1$) are considered. Then, the values of each window s are summed to result in the interpersonal familiarity measure w_{ijt} defined as $\sum_{s=1}^{t-1} \sum_{r_s \in (r_{is} \cap r_{js}), |r_s|>1} \frac{1}{|r_s|-1}$.

To measure team familiarity for project p in time window t, we define Team familiarity$_{pt}$ as the sum of w_{ijt} for all pairs of contributors i and j normalized by the number of pairs of contributors to p in time window t: $\frac{1}{\binom{|p_t|}{2}} \sum_{i>j \wedge i,j \in p_t} w_{ijt}$. The values range from 0 to 299.0.

Recurring Cohesion. To capture tendencies for possible network closure from team cohesion, we again follow Lutter [22] and de Vaan et al [64] in calculating a measure of recurring cohesion, which captures cliques of at least three people who have previously worked together. If three programmers have worked on some project before, and they later worked together again, the network containing this three-person clique can be considered more cohesive than that where any three people only share dyadic ties. A clique is defined as a group of people who at some time prior to current window t worked on a common project within a three-month window; to reduce the complexity of enumerating and checking all possible cliques of large teams, we only consider cliques of up to five members.

After identifying all q_p cliques for a project p at time t, we construct a $q_p \times q_p$ matrix $\mathbf{M}^p$, where each entry (v, w) contains the number of people shared by cliques v and w. Then we use all the off-diagonal, lower triangular values of $\mathbf{M}^p_{v,w}$ to calculate the recurring cohesion as:

$$\text{Recurring cohesion}_{pt} = \frac{1}{2(q_{pt}-1)} \sum_{v<w \wedge v,w \in p_t} \frac{|v| + |v \cap w|}{|p_t|}$$

If there are no cliques, this measure is assigned 0; if there is exactly one clique, say v, the measure is calculated as $\frac{|v|}{|p_t|}$. The values range between 0 and 1547.5.

Team Diversity Measures. $\mathbf{H}_2$ tests whether attachment of women to open teams with regard to diversity of ties and information increases their chance of prolonged engagement relative to men's. To operationalize diversity of information we compute the share of newcomers and heterogeneity of programming language expertise. Indeed, the more newcomers are in a team, and the more diverse expertise team members have, the more diverse is information exchanged in the team.

Share of Newcomers. Following Lutter [22] and Perretti and Negro [66], we calculated each team's share of newcomers, *i.e.*, the fraction of newcomers in a project in time window t relative to the size of the project team at time t. The more newcomers there are in a team, the more new ideas can be brought in, and the more new combinations of relationships can be formed. We operationalize newcomers at project level, *i.e.*, people who never contributed to a given project prior to time t.

Heterogeneity of Programming Language Expertise. Prior work has shown that diverse knowledge is important to innovation and sustainable competitive advantage in many domains [67]. A similar effect may be visible in OSS teams, where assembling a diverse team with expertise in different programming languages or technologies may provide a competitive

advantage, and may help create social connections between members that bridge communities and create opportunities.

Following Lutter's measure of genre diversity in the film industry, based on the distance measure of de Vaan *et al.* [64], we calculate a measure of programming language background heterogeneity at project team level, that considers each team member's prior experience with different programming languages from prior open-source GITHUB projects. We begin with a list of the most popular 33 languages on GITHUB [68]; all other languages in our data are labeled 'Other', generating a set of $K = 34$ languages. On GITHUB each project is labeled with the predominant programming language used therein. Given a project p labeled with the predominant language k, we consider that all developers who contributed to p have experience with k: while individuals may vary in their experience with k, given the size of the dataset we expect a reduction to the mean in terms of individual knowledge; *i.e.*, we expect that, on average, project contributors would have had experience in the predominant language.

For each contributor i in project p in the current time window t, we calculate the vector $f_i = (f_{i1}, ..., f_{iK})$ for each language k, where f_{ik} is 1 if i has worked in projects labeled with the predominant language k. Then, the programming language background distance d_{ijt} between two contributors i and j in the time window t is defined as the cosine of their respective experience vectors. Possible values for this measure range from 1, indicating complete similarity in the language histories of i and j, to 0, indicating complete dissimilarity. Future refinements to this measure, beyond the scope of the current paper, could also consider how similar different programming languages are with each other [69]. We then aggregate these similarity measures at project level, over all pairs of contributors i and j, $i > j$, adjusted for team size, and subtract the result from 1 to obtain a degree of dissimilarity:

$$\text{Language heterogeneity}_{pt} = 1 - \frac{1}{\binom{|p_t|}{2}} \sum_{i>j \wedge i,j \in p_t} d_{ijt},$$

Control Variables. As control variables we consider:

Is Project Owner and *Is Project Major Contributor* both control for the contributor's position in the project. We define major contributors as those authored at least 5% of the project commits during a given window [70]. Being a repository owner or major contributor indicates higher levels of commitment, hence, we expect differences in disengagement rates.

Number of Followers and *Number of Repository Stars* both control for visibility of the contributors and projects, respectively [71]. Popular developers, or developers contributing to popular projects, tend to have a different experience on GITHUB and may be less likely to disengage [72], [73].

Niche width, *i.e.*, the number of programming languages of the developer's past GITHUB commits are spread across. We expect individuals knowing multiple languages to be more versatile and less likely to disengage.

D. Survival Analysis (Quantitative)

To test our hypotheses quantitatively, we use survival analysis, a statistical modeling technique that specializes in time to event data [63]. Survival analysis is particularly suitable for modeling right-censored data like ours.

Estimation. We model jointly the effects of the different social capital factors in Section IV-C on the time to the GITHUB disengagement event, while controling for covariates. For each GITHUB developer in our sample, we have a *survival time* T on record (number of quarters until *has_disengaged* becomes 1). The probability of reaching a given survival time t is given by the *survival function* $S(t) = P(T > t)$, and the probability of leaving the state at time t is given by the *hazard rate* $h(t) = \frac{P(T<t+\Delta t|T \geq t)}{\Delta t}$. The Cox model is a non-parametric regression which can estimate, using partial likelihood, the effect of some independent variables X on the hazard rate, $h(t, X) = \theta(t)f(X)$; *i.e.*, it can estimate the coefficients β of the regression $h(t, X) = \theta(t)\exp(\beta' X)$, where β' denotes the vector transpose of β [63]. The coefficients β can be directly interpreted, *e.g.*, if $\beta_i = 2$, then a unit increase in X_i decreases the probability of survival by $\exp(2) = 7.4$ times.

Many developers disengage early, in their first quarter. In open-source, occasional contributions [74] are common. To model how the different factors contribute to explaining the variability in disengagement rates differently early compared to later on, we split the data set into two parts: developers who disengage in the first quarter and the rest. Since the former only contribute one observation each (one quarter), we model this group using logistic regression (`glm` in R). For the remaining developers, the data set contains repeated quarterly observations. To model these, we estimate a Cox proportional-hazards model.

Diagnostics. Whenever variables had highly skewed distributions, we removed the top 1% of values as potential high-leverage outliers, to increase model robustness [75]; we also log-transformed variables, as needed, to reduce heteroscedasticity [76]. We then tested for multicollinearity (and removed predictors, as needed) using the variance inflation factor (VIF), comparing to the recommended maximum of 5 [77]. Next we inspected the Schoenfeld residual plots [78] (graphical diagnostics) to test the assumption of constant hazard ratios over time. Finally, we report p-values for model coefficients as well as estimates of their effect sizes (fraction of variance explained) from ANOVA analyses.

E. Developer Survey (Qualitative)

To better understand how social capital might impact women and men on GITHUB differently, we conducted a user survey.

Survey design. The aim of the survey was to gain additional context information about how open source contributors perceive their respective projects and the way they collaborate in those project. The survey instrument thus focuses on contributors to collaborative open source GITHUB projects (with at least three contributors, to exclude "toy" projects [79]). Respondents were instructed to choose such a project and base their answers on their experience therein.

We asked open ended questions focusing on their perceived responsibilities and (if applicable) reasons for them to stop contributing. Furthermore, we asked Likert scale questions covering individual satisfaction of contributors being part of this particular project [80], perceived work engagement [81], perceived social capital [82] (the principal construct of our study) and the frequency of communication using different means of communication. We opt to measure individual satisfaction since it has been repeatedly related to loyalty [83], and therefore more satisfied developers can be expected to be less likely to disengage; while work engagement has been shown to be related to turnover intentions [84]. We also aim to assess communication as additional context information about how open source contributors collaborate. For the first three scales we rely on existing instruments that we adapted for our context. In order to assess the frequency of communication we developed a scale that covers different potential means of communication such as *reading each other's code*, *text messaging*, *email* and others. This scale is divided into four levels ranging from *"never or hardly ever"* to *"every day or almost every day"*. The provided means of communication cover typical technologies, *e.g.*, text, audio/video messaging, and typical means of communication in OSS projects, *e.g.*, reading each other's code, commenting on existing code. We also included in person communication for co-located teams.

We also included multiple questions that focus on individual programming skills. The purpose of these questions is not only to assess the potential bandwidth of different skill levels. It can also be expected that differences related to skill level can have an impact on the social structure within a project. Similarly to the niche width in the repository data analysis, we asked participants to identify programming languages that they feel comfortable using. The list we used was based on the most commonly used programming languages in GITHUB. We also asked contributors for how many years they have been active in OSS projects in general and how they rate their skills in comparison to their fellow project contributors. This question has been found to be mostly related to actual programming experience by Siegmund *et al.* [85]. The latter question is related to the tenure diversity shown to be a predictor for turnover in GITHUB teams [42]. Finally we included typical demographic questions: the age and gender of the participants and their education level. Wang and Fesenmaier have shown that when keeping age and educational level constant, men have been members of an online community for a longer period of time [86]. The educational level was based on the Educational Attainment scale by the United States Census Bureau.

Procedure. The population of interest for our study includes female and male contributors to open source GITHUB projects with at least 3 members. We piloted the survey internally with 3 individuals and externally by contacting a total of 800 individuals (400 identified as female and 400 as male by the gender prediction algorithm). Based on the 43 responses we received (5.38% response rate), we revised the survey instrument. For the final survey, we sent 500 invitations to contributors identified by the gender prediction algorithm as women and 500 invitations to those identified as men. The delivery of 6 invitations failed. The survey was available for 2 weeks. We received 107 responses, for a response rate of 10.7%. Responses were anonymous and participation was voluntary. Out of the 107 survey responses received, 93 were complete. Out of the complete responses, 32 respondents identified as female, 56 as male, and 5 did not disclose their gender, which leaves 88 usable responses for the following analysis.

The average reported GITHUB tenure of our survey respondents was 2.50 years, slightly less than what other studies found (*e.g.*, [45] found an average of 3.07 years). This difference could be explained by the larger share of female participants in our survey (36% as opposed to 25% in the survey by Vasilescu et al. [45]) and the fact that female participants in general report shorter tenures than male participants. The tenure of our survey participants is thus generally comparable to that of others in a similar setting. For open ended questions, we conducted an open coding procedure (one author, expert qualitative researcher). For perceived responsibilities we referred to the contributor types that can be found in the GITHUB open source survey [40]. For potential reasons to discontinue contributing to an OSS project we reversed the motivations to contribute to open source [87]. The categories were iteratively refined.

Accuracy of gender prediction. We found a strong correlation between the computed and reported gender. Out of the 107 responses we received, a total of 53 were responses to the survey that we sent to contributors that were identified by the algorithm as female and 54 were responses that were identified by the algorithm as male. Out of the 54 participants our algorithm identified as male, 52 identified themselves as male in the survey and 2 elected not to disclose their gender. Out of the 53 participants our algorithm identified as female, 37 identified themselves as female, 13 identified as male and 3 elected not to disclose their gender.

The algorithm was thus nearly perfect in terms of predicting whether or not a contributor indeed is of male gender (96.30%), as expected given that males are the majority group. The accuracy for predicting whether or not a contributor is of female gender was lower (69.81%) but still above chance. Our algorithm also did not classify female as male contributors: indeed, all participants that were classified as male either reported to be male or did not disclose their gender. This also suggests that the probability of the algorithm missing the contributions of women should be low, since it is capable of detecting male contributors with high accuracy (cf. [59] for discussion of the importance of not misclassifying women).

F. Replication Package

Our data collection and data analysis scripts, the survey instrument, and the input data for the regression models in Table III, are part of a replication package.[3]

[3]https://doi.org/10.5281/zenodo.2550931

V. Results

A. Survey results

What responsibilities do survey respondents have? We asked participants about what they perceive to be their overall responsibilities in the project they selected. To analyze the answers we conducted an open coding procedure based on on the different contributor types in the GitHub open source survey.[4] While applying the contributor types to the survey responses we discovered additional codes ending up with nine distinct but not mutually exclusive responsibility categories.

While participants reported anything between no responsibilities at all and five different responsibilities, most participants reported either one or two. For both genders contributing code is by far the most common perceived responsibility (76.14%), with project management (30.68%) and project lead (22.73%) following at a distance. Male contributors mainly perceive themselves as leaders or managers (37.50% of males report those as their perceived responsibilities) while females appear to take over more non-code related activities such as documentation and proposing ideas (62.50% of females report those as their perceived responsibilities). While this observation concurs with the higher participation of males in the mailing lists related to designing technology [62], the difference is not statistically significant (p = .869 for non-code related activities).

How do survey respondents communicate? We analyzed whether and how respondents interact with each other based on different means of communication. We found that 10 out of 88 respondents never communicated with their fellow project members; Eight of those identified as male (9.09%) and two as female (2.27%). Most of our survey participants thus communicated via any of the provided means of communication.

Participants most commonly communicated via text messages, comments on code and reading each others code in general (almost half of respondents communicate in this way at least once or twice a week). Mail and in person communication are less popular (35.23% and 28.41%, respectively) followed by social networks (11.36%), video messaging (15.91%) and audio messaging (20.45%). Although there are no statistically significant differences between female and male contributors in terms of their communication behavior (p = .979), a closer look into the respective frequencies reveals that female contributors are slightly more active communicating with their fellow project members. This observation concurs with the results of Razavian and Lago: their study has shown that communication is seen by software architects as feminine expertise [88]. In particular, women use text and audio messages as well as social networks more frequently. Males on the other hand appear to use comments on code more frequently than females.

How experienced are the survey respondents? We also asked survey participants about their age, educational background and experience related to both programming in general and contributing to open source projects in particular.

The respondents were mostly between 18 and 34 years old (56.8%) and have a bachelor's or master's degree (67.0%). They reported feeling comfortable using between two and six of the proposed programming languages (77.3%; niche width). Comparing female and male contributors we found that male contributors reported a significantly higher number of programming languages they feel comfortable using ($F = 6.646$, $p < .05$, $\eta^2 = 0.072$). We also found males to report a significantly higher level of expertise ($F = 5.643$, $p < .05$, $\eta^2 = 0.062$). Both are medium effects as demonstrated by η^2 values [89]. There were however no significant differences between female and male contributors in terms of reported age, level of education and years of experience in open source projects. One explanation could be that female contributors are less confident about their programming expertise than male contributors, while neither their education level nor their experience in contributing to open source suggest a valid reason for this perceived difference. This would concur with Wang *et al.*'s finding on women's confidence-competence gap [90].

Why do people stop contributing to GitHub projects? Most of our survey participants are still active in open source (73.9%). Out of the 32 respondents who identified as female, 6 reported that they stopped contributing to open source, while 26 reported that they are still active. Among males, out of the 56 respondents, 17 reported that they stopped contributing while 39 reported that they are still active.

We then conducted a logistic regression analysis on the survey data, using data from the different scales, to model the factors that explain and predict disengagement (binary variable). The multi-item scales we used (individual satisfaction, perceived work engagement, and perceived social capital) are all reliable (Cronbach's α between 0.84 and 0.92). We built an explanatory model, including data from the three scales above, as well as programming experience and reported gender as independent variables. Results from this regression analysis (Table II) showed that **perceived bridging social capital** and **years of programming experience** are significant predictors of individual disengagement. Both bridging social capital and years of experience are comparably strong predictors for individual disengagement (cf. deviance explained in Table II). Gender had no significant direct influence on disengagement.

When looking into self-reported reasons for discontinuing to participate in a GitHub open source project, we found two main reasons: (1) not having enough time to contribute anymore; and (2) no immediate personal need for the respective project. Lack of time was reported to be caused by work related (*"changes in job"*, *"work became over bearing"*) as well as personal reasons (*"diversifying hobbies"*, *"personal life"*). Lack of time was also identified by Lee *et al.* as the most common barrier to participation faced by one-time-contributors to FLOSS projects [91]. Other reported reasons were *"the end of funding of our project"*, frustration (*"failure of our team of backend and front-end"*) or the perception that *"the project [...]*

[4] https://github.com/github/opensource-survey/blob/master/survey-instrument.md

TABLE II: Regression model for the user survey data ($N = 88$).

	GitHub disengagement response: *has_disengaged = 1*	
	exp(Coeffs) (Err.)	LR Chisq
(Intercept)	14.41 (2.55)	
Individual satisfaction (Avg)	2.23 (0.52)	2.95
Work engagement (Avg)	2.00 (0.38)	3.97*
Bridging social capital (Avg)	0.22 (0.60)*	8.37**
Bonding social capital (Avg)	0.61 (0.34)	2.18
Experience relative to team	0.74 (0.31)	0.91
Years of experience	0.72 (0.14)*	6.87**
Education	0.77 (0.24)	1.27
Self-reported gender	2.83 (0.69)	2.44
Niche width	0.96 (0.17)	0.06

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

is finished". When comparing reasons to disengage we found female contributors to report personal reasons significantly more often (F = 4.87, p < .05, η^2 = 0.188). This is a large effect, concurring with the higher likelihood of women leaving and reentering the labor force for personal reasons [92].

B. Survival analysis results

Who are the GITHUB **data developers?** Out of 58,091 programmers, 39,643 have taken a break longer than half a year, and 25,196 programmers have taken a break longer than 1 year. The average age of an account (number of months since the first commit) is 15.01 months; women are statistically younger than men ($p < 2.2^{-16}$, Cliff's $\delta = 0.23$) these results concur with our survey and earlier observations [45], [93]. On average, a programmer contributes to 9.55 projects (median = 4); statistically, women contribute to fewer projects than men ($p < 2.2^{-16}$, Cliff's $\delta = 0.16$). The effect size is in both cases are small (< 0.33) [94].

How does social capital associate with disengagement? Figure 1 plots the Kaplan-Meier estimates revealing that contributors are most likely to drop out in the first two years, and women are more likely to drop out than men in general. Table III presents summaries of our regression models: a logistic regression for contributors who disengage within their first three months of activity (left), and a Cox regression for contributors who disengage later (right).

In both models the control variables behave as expected. More popular (*i.e.*, followers), active (*i.e.*, commits to date) and versatile (*i.e.*, niche width) developers are less likely to disengage. Similarly, project owners, major contributors and contributors to highly starred projects are less likely to disengage. Moreover, as expected, female contributors are at higher risk of disengagement than males: in the short term, being female increases the odds of disengagement from GITHUB by 27%; in the long term, by 32%.

The two variables related to team cohesion have statistically significant effects, and these effects are consistent between the two models. Contributing to projects where team members are more familiar pairwise with each other from prior collaborations (Team familiarity), or projects where cliques of three or more developers recur from prior projects (Recurring cohesion), is associated with decreased risk of disengagement.

TABLE III: Regression models for early-stage disengagement ($N = 29,235$ users; $140,441$ data rows) and later-stage disengagement ($N = 26,299$ users; $143,984$ data rows).

	Early-stage (GLM) response: *Disengaged = 1*		**Later-stage (Cox)** response: *Disengaged = 1*	
	Coeffs (Err.)	LR Chisq	Coeffs (Err.)	LR Chisq
(Intercept)	1.61 (0.07)***			
Followers	0.61 (0.02)***	990.53***	0.70 (0.02)***	394.39***
Stars	0.89 (0.02)***	45.18***	0.86 (0.02)***	103.26***
Commits to date	0.63 (0.01)***	1635.38***	0.64 (0.02)***	718.15***
Is major contrib.	0.77 (0.05)***	29.05***	0.63 (0.06)***	62.96***
Is repo owner	0.56 (0.03)***	363.80***	0.51 (0.04)***	310.35***
Niche width	0.47 (0.05)***	244.20***	0.54 (0.05)***	132.70***
Is female	1.27 (0.03)***	68.79***	1.32 (0.04)***	59.96***
Team familiarity	0.84 (0.08)*	4.83*	0.79 (0.09)**	13.22***
Rec. cohesion	0.85 (0.04)***	30.77***	0.86 (0.04)***	28.46***
Share newcomers	1.07 (0.04)	3.37	0.78 (0.04)***	35.70***
Lang. heterogen.	0.70 (0.11)**	44.44***	0.63 (0.14)***	44.43***
Lang. heter.:Female	0.73 (0.15)*	4.36*	0.69 (0.18)*	4.30*
Female:Team fam.	1.09 (0.11)		1.05 (0.17)	
Female:Cohesion	1.02 (0.05)		1.01 (0.04)	

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

The variables related to team diversity also have statistically significant effects. Heterogeneity in the programming language backgrounds of project team members is associated with decreased risk of disengagement both short and long term. Moreover, language heterogeneity has a statistically significant interaction with gender: women are more likely to disengage when language heterogeneity is low. Contributing to projects with high turnover (Share of newcomers) is associated with higher risk of disengagement after the first three months.

VI. DISCUSSION

A. Hypotheses

$\mathbf{H}_1$ linked social capital to the duration of engagement of OSS developers. Both aspects related to bonding social capital, such as the need to reciprocate, and those related to bridging social capital, such as exposure to new technologies and ideas can be related to developers' motivation. Therefore, $\mathbf{H}_1$ stated that the more often people participate in projects with high potential for building social capital, the higher their chance of prolonged engagement. Our study **strongly supports** this hypothesis. Both regression models (Tables II and III) indicate that social capital, measured by an established survey measurement instrument [82] and by team familiarity and recurring cohesion metrics respectively, is a statistically significant predictor for disengagement. The regression coefficients are lower than one, meaning that *the increase in social capital decreases the chance of disengagement*, other variables held fixed.

$\mathbf{H}_2$ stated that attachment of women to open teams with regard to diversity of ties and information increases their chance of prolonged engagement relative to men's. Table III shows that $\mathbf{H}_2$ is **partially supported**. On the one hand, we found evidence that attachment of women to open teams with regard to diversity of information (language heterogeneity) increases their chance of prolonged engagement: language heterogeneity

interacts with gender. On the other hand, no such interaction could be found for diversity of ties (recurring cohesion and team familiarity), therefore we conclude the support is only partial.

B. Implications

Our results provide empirical evidence that social capital impacts the prolonged engagement of contributors to open-source. Hence, researchers can consider social capital as a lens to investigate social phenomena in OSS.

Given the importance of and concerns about the sustainability of OSS [95], [96], our results suggest that social coding environments like GITHUB should be redesigned to support women in developing social capital, on the one hand, and project maintainers in tracking and being able to react to factors that negatively impact the formation of social capital, on the other hand. We envision: 1) better search functionality and recommendation engines for newcomers looking for projects to join, that take the target project team cohesion and expertise diversity explicitly into account when making a recommendation, to facilitate the formation of social capital, in particular for women (cf. [23]); 2) stemming from the previous point, better mentorship support for newcomers in general and women in particular, whereby mentors can be automatically recommended to potential mentees to facilitate the formation of social capital (cf. [97]); and 3) UI elements besides the ones currently available on GITHUB repository pages, such as badges [98], that allow project maintainers to track worrisome trends in factors negatively associated with the development of social capital (*e.g.*, team expertise diversity and turnover).

C. Threats to Validity

Like any empirical study, our work is subject to threats to validity. First, our results depend on the data collected by GHTORRENT, which may not be a full replica of GITHUB data [79]. We carefully cleaned and filtered our data to avoid the GITHUB mining "perils" [79]. The project-level metrics are calculated based both on the contributors' own forks and their base repositories (the repository to which they make pull requests). We also focus on commits instead of pull requests because only a fraction of projects use pull requests [79]. We repeatedly manually checked data outliers *e.g.*, large repositories that are not software projects, but tutorials. We excluded projects with large number of zero-commit forks and repositories with huge numbers of forks and commits (top 1%).

A second threat to validity may come from our gender classifier. The accuracy of the classifier is limited by the information users display on GITHUB. Many users do not use their real names so we cannot extract their gender information reliably [40]. Some users display names in a language for which our gender classifier does not have data. Moreover, there are many top female developers from East Asia [90]. It is difficult to verify their gender identity because their names are gender neutral and their profile pictures are not necessarily their own photos. Furthermore, our gender classifier, as any automatic classifier we are aware of, is based on the assumption of binary gender, and as such our work cannot explicitly take into account contributions by non-binary software developers.

Third, we used a single coder for the open ended survey questions which might result in a subjective interpretation of the responses. We attempted to mitigate this threat by building on established categories.

Finally, statistical modeling required many operational decisions (*e.g.*, time windows, length of inactivity): ours follow best practices and prior work. Again following best practices, we tested sensitivity of our operational decisions. Given space restrictions, we prioritized replicability and validity, reporting all decisions made, but in cases of insensitive parameters did not always discuss the rationale for a specific value.

VII. CONCLUSIONS

In this paper we have studied the impact of social capital on sustained participation of open source contributors and, in particular, on gender differences in this impact. We have performed a mixed-methods empirical study combining survival analysis on a longitudinal data set of 58,091 open source contributors and their GITHUB contributions, with a survey of 98 developers. Our studies show that in general social capital positively affects sustained participation in open source on GITHUB. For women, diversity of the project members' expertise becomes crucial to sustain their participation: we found that higher team diversity along dimensions of programming language expertise is associated with decreased risk of disengagement both short and long term.

Our secondary contribution is the very first gender inference tool explicitly targeting Chinese, Japanese, and Korean names, achieving 83.62% accuracy overall, and at least 60.00% on (South) East Asian names. This opens multiple directions of further research from replication of earlier gender studies [42], [43], [62], [99] for East Asian contributors to exploration of new datasets such as STACK OVERFLOW in Japanese.[5]

In the same way as we have studied the impact of language heterogeneity on the disengagement of women, future work should also consider the impact of gender diversity and gender homophily, i.e., preference of people to interact more with people of the same gender, of the teams on the disengagement of women [42], [100]. Furthermore, our study can be replicated to investigate the relation between social capital and sustained participation on other platforms, *e.g.*, STACK OVERFLOW, and the impact of different demographic aspects.Finally, understanding the relation between social capital and sustained participation on GITHUB is the key to designing appropriate interventions aiming at ensuring engagement of women in open source software projects more broadly.

ACKNOWLEDGEMENTS

Qiu and Vasilescu gratefully acknowledge support from the Alfred P. Sloan Foundation. Many thanks to Mark Lutter for useful discussion on earlier versions of this work, Elian Carsenat for access to NamSor, our survey respondents, and the anonymous reviewers.

[5]https://ja.stackoverflow.com/

REFERENCES

[1] Y. Fang and D. Neufeld, "Understanding sustained participation in open source software projects," *J Manage Inform Syst*, vol. 25, no. 4, pp. 9–50, 2009.

[2] J. Coelho and M. T. Valente, "Why modern open source projects fail," in *ESEC/FSE*. ACM, 2017, pp. 186–196.

[3] J. Marlow and L. Dabbish, "Activity traces and signals in software developer recruitment and hiring," in *CSCW*, 2013, pp. 145–156.

[4] GitHub, "Open source survey," http://opensourcesurvey.org/2017/.

[5] 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*, vol. 52, no. 7, pp. 984–999, 2006.

[6] G. Hertel, S. Niedner, and S. Herrmann, "Motivation of software developers in open source projects: an internet-based survey of contributors to the Linux kernel," *Research Policy*, vol. 32, no. 7, pp. 1159–1177, 2003.

[7] B. Lin, G. Robles, and A. Serebrenik, "Developer turnover in global, industrial open source projects: Insights from applying survival analysis," in *ICGSE*, 2017, pp. 66–75.

[8] A. Schilling, S. Laumer, and T. Weitzel, "Who will remain? an evaluation of actual person-job and person-team fit to predict developer retention in FLOSS projects," in *HICSS*, 2012, pp. 3446–3455.

[9] Y. Jiang, B. Adams, and D. M. German, "Will my patch make it? and how fast?: Case study on the Linux kernel," in *MSR*, 2013, pp. 101–110.

[10] V. J. Hellendoorn, P. T. Devanbu, and A. Bacchelli, "Will they like this?: Evaluating code contributions with language models," in *MSR*, 2015, pp. 157–167.

[11] R. Padhye, S. Mani, and V. S. Sinha, "A study of external community contribution to open-source projects on GitHub," in *MSR*, 2014, pp. 332–335.

[12] Y. Tao, D. Han, and S. Kim, "Writing acceptable patches: An empirical study of open source project patches," in *ICSME*, 2014, pp. 271–280.

[13] G. Gousios, M. Pinzger, and A. v. Deursen, "An exploratory study of the pull-based software development model," in *ICSE*, 2014, pp. 345–355.

[14] R. S. Burt, "The gender of social capital," *Rationality and Society*, vol. 10, no. 1, pp. 5–46, 1998.

[15] ——, "Structural holes versus network closure as social capital," in *Social Capital: Theory and Research*. De Gruyter, 2001, pp. 31–56.

[16] C.-M. Chiu, M.-H. Hsu, and E. T. Wang, "Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories," *Decision Support Systems*, vol. 42, no. 3, pp. 1872–1888, 2006.

[17] M. B. Aguilera, "The impact of social capital on labor force participation: Evidence from the 2000 social capital benchmark survey," *Social Science Quarterly*, vol. 83, no. 3, pp. 853–874, 2002.

[18] E. Brown and J. M. Ferris, "Social capital and philanthropy: An analysis of the impact of social capital on individual giving and volunteering," *Nonprof Volunt Sec Q*, vol. 36, no. 1, pp. 85–99, 2007.

[19] L. Guiso, P. Sapienza, and L. Zingales, "The role of social capital in financial development," *American Economic Review*, vol. 94, no. 3, pp. 526–556, June 2004.

[20] J. Hahn, J. Y. Moon, and C. Zhang, "Emergence of new project teams from open source software developer networks: Impact of prior collaboration ties," *Information Systems Research*, vol. 19, no. 3, pp. 369–391, 2008.

[21] C. Casalnuovo, B. Vasilescu, P. Devanbu, and V. Filkov, "Developer onboarding in GitHub: The role of prior social links and language experience," in *ESEC/FSE*, 2015, pp. 817–828.

[22] M. Lutter, "Do women suffer from network closure? the moderating effect of social capital on gender inequality in a project-based labor market, 1929 to 2010," *American Sociological Review*, vol. 80, no. 2, pp. 329–358, 2015.

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

[24] P. S. Adler and S.-W. Kwon, "Social capital: Prospects for a new concept," *Acad Manage Rev*, vol. 27, no. 1, pp. 17–40, 2002.

[25] G. Madey, V. Freeh, and R. Tynan, "The open source software development phenomenon: An analysis based on social network theory," in *AMCIS*, 2002, pp. 1806–1813.

[26] J. Wang, "The role of social capital in open source software communities," *AMCIS*, p. 427, 2005.

[27] J. S. Coleman, *Foundations of social theory*. Belknap, 1990.

[28] M. S. Granovetter, "The strength of weak ties," *Am J Sociol*, vol. 78, no. 6, pp. 1360–1380, 1973.

[29] W. E. Baker and A. V. Iyer, "Information networks and market behavior," *Journal of Mathematical Sociology*, vol. 16, no. 4, pp. 305–332, 1992.

[30] B. Xu and D. R. Jones, "Volunteers' participation in open source software development: a study from the social-relational perspective," *ACM SIGMIS Database*, vol. 41, no. 3, pp. 69–84, 2010.

[31] W. Oh and S. Jeon, "Membership herding and network stability in the open source community: The Ising perspective," *Management Science*, vol. 53, no. 7, pp. 1086–1101, 2007.

[32] P. Song and C. W. Phang, "Promoting continuance through shaping members' social identity in knowledge-based versus support/advocacy virtual communities," *IEEE T Eng Manage*, vol. 63, no. 1, pp. 16–26, 2016.

[33] S. Toral, M. Martínez-Torres, and F. Barrero, "Analysis of virtual communities supporting OSS projects using social network analysis," *Inform Software Tech*, vol. 52, no. 3, pp. 296–303, 2010.

[34] S. Christopherson, "Working in the creative economy: Risk, adaptation and the persistence of exclusionary networks," *Creative labour: Working in the creative industries*, pp. 72–90, 2009.

[35] I. Grugulis and D. Stoyanova, "Social capital and networks in film and tv: Jobs for the boys?" *Organization Studies*, vol. 33, no. 10, pp. 1311–1331, 2012.

[36] H. Blair, "Active networking: action, social structure and the process of networking," *Creative Labour: Working in the Creative Industries*, pp. 116–134, 2009.

[37] A. Portes, "Social capital: Its origins and applications in modern sociology," *Annual Review of Sociology*, vol. 24, no. 1, pp. 1–24, 1998.

[38] B. Groysberg, *Chasing Stars: The Myth of Talent and the Portability of Performance*. Princeton University Press, 2010.

[39] K. Finley, "Diversity in open source is even worse than in tech overall," https://www.wired.com/2017/06/diversity-open-source-even-worse-tech-overall/.

[40] R. S. Geiger, "Summary analysis of the 2017 GitHub open source survey," *arXiv preprint 1706.02777*, 2017.

[41] D. Nafus, "'patches don't have gender': What is not open in open source software," *New Media & Society*, vol. 14, no. 4, pp. 669–683, 2012.

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

[43] 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.

[44] N. Lin, *Social Capital: A Theory of Social Structure and Action*. Cambridge University Press, 2001.

[45] B. Vasilescu, V. Filkov, and A. Serebrenik, "Perceptions of diversity on GitHub: A user survey," in *CHASE*, 2015, pp. 50–56.

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

[47] L. López-Fernández, G. Robles, and J. González-Barahona, "Applying social network analysis to the information in CVS repositories," in *MSR*, 2004, pp. 101–105.

[48] H.-L. Yang and J.-H. Tang, "Team structure and team performance in is development: A social network perspective," *Inform Manage*, vol. 41, no. 3, pp. 335–349, 2004.

[49] K. Ehrlich and K. Chang, "Leveraging expertise in global software teams: Going outside boundaries," in *ICGSE*, 2006, pp. 149–158.

[50] S. Toral, M. Martínez-Torres, and F. Barrero, "Analysis of virtual communities supporting oss projects using social network analysis," *Inf Sw Tech*, vol. 52, no. 3, pp. 296–303, 2010.

[51] D. A. Tamburri, P. Lago, and H. van Vliet, "Uncovering latent social communities in software development," *IEEE Software*, vol. 30, no. 1, pp. 29–36, Jan 2013.

[52] M. Foucault, M. Palyart, X. Blanc, G. C. Murphy, and J.-R. Falleri, "Impact of developer turnover on quality in open-source software," in *ESEC/FSE*, 2015, pp. 829–841.

[53] M. Zhou and A. Mockus, "What make long term contributors: Willingness and opportunity in OSS community," in *ICSE*. IEEE, 2012, pp. 518–528.

[54] S. Easterbrook, J. Singer, M.-A. Storey, and D. Damian, "Selecting empirical methods for software engineering research," *Guide to Advanced Empirical Software Engineering*, pp. 285–311, 2008.

[55] G. Gousios, "The GHTorrent dataset and tool suite," in *MSR*, 2013, pp. 233–236.

[56] I. S. Wiese, J. T. da Silva, I. Steinmacher, C. Treude, and M. A. Gerosa, "Who is who in the mailing list? comparing six disambiguation heuristics to identify multiple addresses of a participant," in *ICSME*, 2016, pp. 345–355.

[57] C. Bird, A. Gourley, P. Devanbu, M. Gertz, and A. Swaminathan, "Mining email social networks," in *MSR*, 2006, pp. 137–143.

[58] B. Vasilescu, A. Serebrenik, M. Goeminne, and T. Mens, "On the variation and specialisation of workload—a case study of the Gnome ecosystem community," *Empir Softw Eng*, vol. 19, no. 4, pp. 955–1008, 2014.

[59] B. Lin and A. Serebrenik, "Recognizing gender of Stack Overflow users," in *MSR*, 2016, pp. 425–429.

[60] F. Karimi, C. Wagner, F. Lemmerich, M. Jadidi, and M. Strohmaier, "Inferring gender from names on the web: A comparative evaluation of gender detection methods," in *WWW Companion*, 2016, pp. 53–54.

[61] F. Rangel, P. Rosso, M. Potthast, and B. Stein, "Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter," *Working Notes Papers of the CLEF*, 2017.

[62] B. Vasilescu, A. Capiluppi, and A. Serebrenik, "Gender, representation and online participation: A quantitative study," *Interact Comput*, vol. 26, no. 5, pp. 488–511, 2013.

[63] R. G. Miller Jr, *Survival analysis*. Wiley, 2011, vol. 66.

[64] M. De Vaan, B. Vedres, and D. Stark, "Disruptive diversity and recurring cohesion: Assembling creative teams in the video game industry, 1979–2009," Institute for Social and Economic Research and Policy, Tech. Rep. 3, 2011.

[65] M. E. J. Newman, "Scientific collaboration networks II. Shortest paths, weighted networks, and centrality," *Physical Review E*, vol. 64, no. 1, pp. 016 132:1–7, 2001.

[66] F. Perretti and G. Negro, "Mixing genres and matching people: a study in innovation and team composition in Hollywood," *J Organ Behav*, vol. 28, no. 5, pp. 563–586, 2007.

[67] S. Rodan and C. Galunic, "More than network structure: How knowledge heterogeneity influences managerial performance and innovativeness," *Strategic management journal*, vol. 25, no. 6, pp. 541–562, 2004.

[68] C. Zapponi, "Programming languages and github," http://githut.info/, 2017, visited 29 June 2017.

[69] B. Vasilescu, A. Serebrenik, and M. G. J. van den Brand, "The Babel of software development: Linguistic diversity in open source," in *SocInfo*, 2013, pp. 391–404.

[70] C. Bird, N. Nagappan, B. Murphy, H. Gall, and P. Devanbu, "Don't touch my code!: examining the effects of ownership on software quality," in *FSE*, 2011, pp. 4–14.

[71] J. Sheoran, K. Blincoe, E. Kalliamvakou, D. Damian, and J. Ell, "Understanding watchers on GitHub," in *MSR*. ACM, 2014, pp. 336–339.

[72] L. Dabbish, C. Stuart, J. Tsay, and J. Herbsleb, "Social coding in GitHub: transparency and collaboration in an open software repository," in *CSCW*. ACM, 2012, pp. 1277–1286.

[73] J. Marlow, L. Dabbish, and J. Herbsleb, "Impression formation in online peer production: activity traces and personal profiles in GitHub," in *CSCW*, 2013, pp. 117–128.

[74] G. Pinto, I. Steinmacher, and M. A. Gerosa, "More common than you think: An in-depth study of casual contributors," in *SANER*, 2016, pp. 112–123.

[75] J. K. Patel, C. Kapadia, and D. B. Owen, *Handbook of statistical distributions*. M. Dekker, 1976.

[76] A. Gelman and J. Hill, *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[77] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.

[78] P. M. Grambsch and T. M. Therneau, "Proportional hazards tests and diagnostics based on weighted residuals," *Biometrika*, vol. 81, no. 3, pp. 515–526, 1994.

[79] E. Kalliamvakou, G. Gousios, K. Blincoe, L. Singer, D. M. German, and D. Damian, "The promises and perils of mining GitHub," in *MSR*, 2014, pp. 92–101.

[80] A. Filippova, E. Trainer, and J. D. Herbsleb, "From diversity by numbers to diversity as process: supporting inclusiveness in software development teams with brainstorming," in *ICSE*, 2017, pp. 152–163.

[81] W. B. Schaufeli, A. B. Bakker, and M. Salanova, "The measurement of work engagement with a short questionnaire: A cross-national study," *Educ Psychol Meas*, vol. 66, no. 4, pp. 701–716, 2006.

[82] N. B. Ellison, C. Steinfield, and C. Lampe, "The benefits of facebook "friends:" social capital and college students' use of online social network sites," *J Comput-Mediat Comm*, vol. 12, no. 4, pp. 1143–1168, 2007.

[83] J. Drengner, S. Jahn, and H. Gaus, "Events and loyalty formation: The role of satisfaction, felt community, emotional experience, and frequency of use," in *Stand und Perspektiven der Eventforschung*. Wiesbaden: Gabler, 2010, pp. 151–165.

[84] W. B. Schaufeli and A. B. Bakker, "Job demands, job resources, and their relationship with burnout and engagement: a multi-sample study," *J Organ Behav*, vol. 25, no. 3, pp. 293–315, 2004.

[85] J. Siegmund, C. Kästner, J. Liebig, S. Apel, and S. Hanenberg, "Measuring and modeling programming experience," *Empir Softw Eng*, vol. 19, no. 5, pp. 1299–1334, 2014.

[86] Y. Wang and D. R. Fesenmaier, "Modeling participation in an online travel community," *J Travel Res*, vol. 42, no. 3, pp. 261–270, 2004.

[87] K. R. Lakhani and R. G. Wolf, "Why hackers do what they do: Understanding motivation and effort in free/open source software projects," MIT, Tech. Rep. 4425-03, 2003.

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

[89] J. Cohen, "Statistical power analysis for the behavioral sciences," 1988.

[90] Z. Wang, Y. Wang, and D. Redmiles, "Competence-confidence gap: A threat to female developers' contribution on GitHub," in *ICSE*, 2018, pp. 81–90.

[91] A. Lee, J. C. Carver, and A. Bosu, "One-time contributors to FLOSS: surveys and data analysis," in *ICSE*, 2017, pp. 187–197.

[92] S. Elder and L. J. Johnson, "Sex-specific labour market indicators: What they show," *Int'l Labour Review*, vol. 138, no. 4, pp. 447–464, 2008.

[93] G. Robles, L. A. Reina, J. M. González-Barahona, and S. D. Domínguez, "Women in free/libre/open source software: The situation in the 2010s," in *Open Source Systems: Integrating Communities*, 2016, pp. 163–173.

[94] J. Romano, J. D. Kromrey, J. Skowronek, and L. Devine, "Exploring methods for evaluating group differences on the NSSE and other surveys: Are the t-test and Cohen's d indices the most appropriate choices?" in *Ann. meeting, South Assoc Institutional Research*, 2006, pp. 1–51.

[95] N. Eghbal, "Roads and bridges: The unseen labor behind our digital infrastructure," Ford Foundation, Tech. Rep., 2016.

[96] M. Valiev, B. Vasilescu, and J. Herbsleb, "Ecosystem-level determinants of sustained activity in open-source projects: A case study of the PyPI ecosystem," in *ESEC/FSE*. ACM, 2018, pp. 644–655.

[97] S. Balali, I. Steinmacher, U. Annamalai, A. Sarma, and M. A. Gerosa, "Newcomers' barriers. . . is that all? an analysis of mentors' and newcomers' barriers in OSS projects," *Comp Support Coop W*, vol. 27, no. 3-6, pp. 679–714, 2018.

[98] A. Trockman, S. Zhou, C. Kästner, and B. Vasilescu, "Adding sparkle to social coding: An empirical study of repository badges in the npm ecosystem," in *ICSE*, 2018.

[99] V. Kuechler, C. Gilbertson, and C. Jensen, "Gender differences in early free and open source software joining process," in *IFIP International Conference on Open Source Systems*. Springer, 2012, pp. 78–93.

[100] D. Ford, A. Harkins, and C. Parnin, "Someone like me: How does peer parity influence participation of women on Stack Overflow?" in *VL/HCC*, 2017, pp. 239–243.