

# Towards Mining OSS Skills from GitHub Activity

Jenny T. Liang  
University of Washington  
Seattle, Washington, USA  
jliang9@cs.washington.edu

Thomas Zimmermann  
Microsoft Research  
Redmond, Washington, USA  
tzimmer@microsoft.com

Denae Ford  
Microsoft Research  
Redmond, Washington, USA  
denae@microsoft.com

## ABSTRACT

Open source software (OSS) development relies on diverse skill sets. However, to our knowledge, there are no tools which detect OSS-related skills. In this paper, we present a novel method to detect OSS skills and prototype it in a tool called `DISKO`. Our approach relies on identifying relevant *signals*, which are measurable activities or cues associated with a skill. Our tool detects how contributors 1) teach others to be involved in OSS projects, 2) show commitment towards an OSS project, 3) have knowledge in specific programming languages, and 4) are familiar with OSS practices. We then evaluate the tool by administering a survey to 455 OSS contributors. We demonstrate that `DISKO` yields promising results: it detects the presence of these skills with precision scores between 77% to 97%. We also find that over 54% of participants would display their high-proficiency skills. Our approach can be used to transform existing OSS experiences, such as identifying collaborators, matching mentors to mentees, and assigning project roles. Given the positive results and potential impact of our approach, we outline future research opportunities in interpreting and sharing OSS skills.

## CCS CONCEPTS

• **Software and its engineering** → **Open source model.**

## KEYWORDS

open source software, skills detection, mining software repositories

### ACM Reference Format:

Jenny T. Liang, Thomas Zimmermann, and Denae Ford. 2022. Towards Mining OSS Skills from GitHub Activity. In *New Ideas and Emerging Results (ICSE-NIER'22)*, May 21–29, 2022, Pittsburgh, PA, USA. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3510455.3512772>

## 1 INTRODUCTION

Constructing open source software (OSS) depends on contributors with a diverse set of skills. Each skill is vital to OSS: problem-solving skills allow software developers to build new features to address issues; organizational skills help software maintainers manage the moving parts of an OSS project; and communication skills enable writers to generate clear, concise documentation and facilitate collaboration. Unlike in software engineering, where programmers

primarily write code, many OSS contributors provide equally valuable non-code contributions [8, 37]. While OSS-related skills include a subset of software engineering skills, contributors also work in contexts unique to OSS (e.g., wrangling contributors, identifying funding, consistently collaborating in a distributed form).

Despite the importance of skills related to OSS development, to our knowledge, there are no tools that currently exist which detect such skills. Related work in software engineering has developed techniques to detect specific software engineering skills, for example, Java programming skills [12] and general programming experience [35]. Meanwhile, other work detects programming-related skills by pulling data from version control systems such as GitHub [19, 21, 29–32]. Montandon et al. showed that this data could help identify experts in OSS communities [30] and predict technical roles of GitHub users [31]. Other work has demonstrated that GitHub data can be used to extract skills for job recommendations [19, 21].

These studies have significantly advanced the field, but focus largely on a single topic: mining technical programming skills. Thus, a gap still remains in mining OSS expertise, which encapsulates both soft and hard skills [39] across many roles aside from software engineering. Rather than simply focusing on technical software development skills, our work focuses on mining GitHub data to detect both soft skills and hard skills, which are vital to OSS development.

In this paper, we introduce a method to detect OSS skills and implement it in a tool called `DISKO` (**DETECTING SKILLS in OSS**), with promising results. Our approach relies on identifying accurate *signals*, which are measurable activities or cues associated with a skill used to identify the presence of having that skill. For example, having proficiency in a programming language could be measured using a certain number of lines of code written in that language [12]. The notion of signals follows prior work, such as Marlow et al., who identified cues that GitHub users utilized to judge a contributor's coding ability, project-relevant skills, and personality [28].

We discuss how we identify relevant signals (Section 2) and outline `DISKO`'s features (Section 3). We then present an evaluation of `DISKO` and its results (Section 4). Finally, we discuss the implications (Section 5) and next steps for this research (Section 6).

## 2 IDENTIFYING SIGNALS FOR SKILLS

To develop `DISKO`, OSS skills and signals need to be identified. The first author extracted relevant **OSS skills** by reading prior literature on software engineering expertise [9, 10, 25, 33] and social factors of OSS [15, 18, 28, 36, 38]. These papers were selected to include a broad set of roles and contribution types. The first author read each manuscript and identified key skills.

Next, we performed additional literature review to identify **signals** for each skill. In this work, *signals are expressed as true or false statements* (i.e., the signal is either present or not). Potential signals were elicited through the nine papers to identify OSS skills

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
*ICSE-NIER'22*, May 21–29, 2022, Pittsburgh, PA, USA

© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 978-1-4503-9224-2/22/05...\$15.00  
<https://doi.org/10.1145/3510455.3512772>

**Table 1: Skills-signals model based on literature review. [A] denotes an author-defined signal. [CE] denotes a signal from one of three OSS community experts.**

Signal	Literature Source
<b>Teaches others to be involved in the OSS project</b>	
Contributes to a pull request with a newcomer at least 3 times	[36]
Comments on code in others' pull requests at least once	[15]
Comments overall on others' pull requests at least once	[15]
Contributes at least 5 changes to Markdown files at least once	[13] [36]
Contributes at least 5 changes to community health files at least once	[2] [13] [36]
Is the only person (aside from the owner) who commented on an issue or pull request at least 3 times	[15] [28]
Has not responded to someone's open issue in less than an hour in the past three months	[40]
There is no more than 1 comment on pull requests or issues with toxicity scores over 0.5 in the past year	[20] [34]
<b>Shows commitment towards the OSS project</b>	
At least 36 months where there is at least one contribution per month across all projects	[40] [CE]
At least 12 months where there is at least one contribution per month across all projects	[40] [CE]
Is involved in the discussion of their own PR around 70% of the time in a particular project in the past year	[24]
Number of GitHub members who follow the new contributor is at the 75th percentile across users	[11] [16]
Has write rights to a repository they don't own	[16]
Is at the 75th percentile by number of commits to a repository across users	[16]
<b>Has knowledge in specific programming languages</b>	
Has made a commit in the language at least once	[A]
The lines changed in the language is at the 20th percentile across users	[12]
The lines changed in the language is at the 40th percentile across users	[12]
The lines changed in the language is at the 60th percentile across users	[12]
The lines changed in the language is at the 80th percentile across users	[12]
<b>Is familiar with OSS practices</b>	
Has made a commit	[A]
Has opened a pull request	[A]
Has opened an issue	[A]
Has made a comment on another's pull requests	[A]
Has made a comment on another's issue	[A]
Has been assigned to an issue, closed an issue, or has merged a pull request	[A]

[9, 10, 15, 18, 25, 28, 33, 36, 38] and relevant literature on mining OSS activity [11–14, 16, 20, 24, 34, 40]. We read through each manuscript, identified potential signals for as many OSS skills as possible, and recorded the source of each signal. We finalized the signals from literature by converting each one as a true or false statement by defining thresholds (e.g., an activity happens at least  $N$  times, an activity occurs with a frequency at least at the  $N$ th percentile). We then generated signals which seemed viable to compute and were representative of the skill. We designed our own signals for *Is familiar with OSS practices* because it was cited often in literature [9, 18, 36], but prior work did not define clear signals for this skill.

During this process, authors consulted three OSS community experts who made significant contributions to OSS and regularly work with OSS stakeholders for their profession. They provided insight on important skills and signals. Based on the experts' feedback and the frequency of the skills cited in literature, we reduced the final model to four skills, which is shown in Table 1.

### 3 AUTOMATICALLY DETECTING SKILLS

After developing a model of OSS skills and their associated signals (see Section 2), the first author implemented `DISKO` in Python to detect a contributor's OSS skills from the model based on GitHub user data. We use the GitHub GraphQL API [4], GitHub REST API [5], and GHTorrent [17] as sources of GitHub data.

Our tool rates skills on a zero to five scale, where zero represents no proficiency and five represents a high level of proficiency. We use the function  $rate(N, M)$  to determine skill level, where  $N$  represents the number of signals present and  $M$  represents the total number of signals a skill has. This function weights each signal equally.

$$rate(N, M) = \begin{cases} 0 & \frac{N}{M} = 0 \\ 1 & 0 < \frac{N}{M} \leq 0.2 \\ 2 & 0.2 < \frac{N}{M} \leq 0.4 \\ 3 & 0.4 < \frac{N}{M} \leq 0.6 \\ 4 & 0.6 < \frac{N}{M} \leq 0.8 \\ 5 & 0.8 < \frac{N}{M} \leq 1.0 \end{cases}$$

*Selecting programming languages.* Programming languages that the tool detected were selected based on if they were present in both the 2020 State of the Octoverse [1] and the 2020 Stack Overflow Developer Survey [7]. This resulted in 9 programming languages: C, C#, Java, JavaScript, PHP, Python, Ruby, Shell, and TypeScript.

*Excluding third-party libraries as user contributions.* Contributors often included code from third-party libraries in their commits, which has also been shown in prior work [27]. Code from third-party libraries thus should be excluded from contributors' mined contributions. To that end, we identify popular package managers

for each programming language from a Wikipedia article on popular package managers [6]. We compile a list of installation folder names across all the package managers. Next, we analyze the top folder names for each programming language. We use a list of OSS for social good (OSS4SG) projects (i.e., OSS projects which address a societal issue and target a specific community) from Huang et al. [22]. We download the contents of each repository in this list and record its file tree. From this, we identify files written in the language based on file extensions and extract the path from the repository folder. We then retrieve the top 50 folder names associated with each programming language. For each language, we exclude the entire file tree under a folder name from the analysis when: 1) the folder name corresponds to an installation location of the programming language’s package manager(s), or 2) the folder name is in the top 50 folder names for the programming language and the list of installation folder names across all package managers.

**3.0.1 Computing distributions.** Some signals rely on the user activity being at a certain percentile and thus depends on an underlying distribution to compare to. To compute this, we generate a list of OSS contributors by recording the top 30 contributors per project from Huang et al.’s list of 437 OSS4SG projects [22]. We randomly select 500 GitHub users from this list. Each distribution is computed based on these 500 users’ relevant activity for the distribution.

## 4 EVALUATION

### 4.1 Design

We designed a two-part survey to validate DISKO (see Section 3):

- (1) a Qualtrics survey where participants submitted anonymized responses and
- (2) a Microsoft Forms survey where participants submitted personal identifying information (PII). This survey was displayed after the completion of the Qualtrics survey.

The survey was implemented in two parts so PII was linked only to a small number of questions. Participants could choose to only take the Qualtrics survey and not provide any PII. Topics in the Qualtrics survey included the importance of our tool’s OSS skills and the participants’ willingness to display their skills and ratings on GitHub. Topics in the Microsoft Forms survey included GitHub usernames and self-assessments of the skills from DISKO. Some survey questions are shown in Figure 1 and the full survey instrument is available as supplemental material [26]. We sent the survey to a subset of contributors who authored commits, opened issues and pull requests, or commented on others’ issues and pull requests from Huang et al.’s list of 1,079 OSS and OSS4SG projects [22]. Our survey was sent to a total of 9,095 OSS contributors with a response rate of 5%. The Qualtrics survey received 455 responses. The Microsoft Forms survey received 386 responses, resulting in 316 valid usernames with public code contributions from merged pull requests. After completing the survey, participants could join a sweepstakes to win one of four \$100 Amazon.com gift cards.

The GitHub usernames allowed us to compare the skill self-assessments with the skills scores computed by DISKO. To evaluate DISKO, we focus on its precision—that is, when the participant is confident they have a skill, does DISKO agree? We used two measurements of precision: 1) the precision of users who had self-evaluated

#### PART 1: QUALTRICS SURVEY QUESTIONS (ANONYMOUS)

- Rate the following skills based on their importance for OSS projects.
- For each of the skills below, rate your agreement with the following statement: if I were proficient at this skill, I would publicly display that I have proficiency in this skill on my GitHub profile.
- For each of the programming languages below, rate your agreement with the following statement: if I were proficient at this programming language, I would publicly display that I have proficiency in this programming language on my GitHub profile.

#### PART 2: MICROSOFT FORMS SURVEY QUESTIONS

- Please enter your GitHub username.
- Rate your proficiency in each skill below.
- Rate your proficiency in each programming language below.

**Figure 1: A subset of the survey questions. The complete survey instrument is in the supplemental materials [26].**

**Table 2: Participants who found each skill important.**

Skill	Importance
Teaches others to be involved in the OSS project	64%
Shows commitment towards the OSS project	67%
Has knowledge in specific programming languages	45%
Is familiar with OSS practices	56%

skill levels greater than 0 (i.e., detecting the presence of a skill) and 2) the precision of users who had self-evaluated skill levels greater than 3 (i.e., detecting moderate to high skill proficiency).

For the analysis of the survey, we only look at close-ended questions and use standard statistical analysis techniques. We report percentages on how frequently participants agreed or strongly agreed with a statement and how frequently participants said a skill was important or very important. This follows Kitchenham’s and Pflieger’s best practices to analyze survey data [23].

### 4.2 Preliminary Results

**4.2.1 Skill importance.** All the detected skills were important to participants; the results are shown in Table 2. A majority of participants found the soft skills (*Teaches others to be involved in the OSS project*, *Shows commitment towards the OSS project*) to be important. Notably, participants found soft skills more important than hard skills. *Is familiar with OSS practices* was also rated as important by a majority of participants, but was less important than the soft skills.

**4.2.2 Displaying skills.** All detected skills would be displayed by a majority of participants (see Table 3). Participants were most willing to share *Is familiar with OSS practices*, *JavaScript*, and *Python*. *PHP* was the least popular skill to share. Overall, there was variation in how willing participants were to share certain skills. In the free response, some participants were excited by sharing skills and

**Table 3: Results of the survey and evaluation of DISKO. Mean values in the survey are from people who reported to have the skill. The tool’s precision scores are from those who had self-evaluated their skill level to be greater than 0 and 3.**

Skill	Survey			Tool		
	Distribution 0 1 2 3 4 5	Mean	Would display	Distribution 0 1 2 3 4 5	Prec. (>0)	Prec. (>3)
OSSFamiliarity		4.06	79%		97%	74%
Commitment		3.92	67%		96%	63%
TeamPlayer		3.94	69%		97%	73%
Javascript		3.38	80%		95%	65%
Python		3.45	80%		96%	88%
Java		2.80	67%		96%	67%
Php		2.52	54%		89%	56%
C#		2.54	64%		83%	100%
Typescript		2.98	73%		92%	79%
Shell		3.13	69%		91%	61%
C		2.89	73%		98%	92%
Ruby		2.46	66%		77%	58%

suggested new languages to detect (e.g., Golang, Rust) or expressed encouragement. Others expressed concerns about the impact of DISKO, questioning whether it should be deployed.

**4.2.3 Tool accuracy.** DISKO’s performance is displayed in Table 3. We find that DISKO identifies the presence of OSS skills with impressive performance, with precision scores ranging from 77% (*Ruby*) to 97% (*Is familiar with OSS practices*, *Teaches others to be involved in the OSS project*). However, the overall performance of DISKO drops while detecting moderately skilled to expert users. It is also notable that the rating distributions for *Teaches others to be involved in the OSS project*, *Shows commitment towards the OSS project*, and *Is familiar with OSS practices* are skewed positively in the survey.

## 5 DISCUSSION

Our evaluation shows encouraging results for DISKO. Participants are most excited to display programming language-related skills, which our tool detects with reasonable confidence. However, the high importance placed on soft skills by our participants should not be overlooked. This is especially relevant with emerging neural models that generate code with high quality, such as GitHub Copilot [3]. Given its importance, future versions of DISKO could be enhanced to more accurately detect soft skills.

**Potential applications.** Our results indicate a potential future for skill detection approaches within OSS development and software engineering. While skills underlie these activities, they are not widely supported in tooling. Skills detection methods such as ours could be applied in practice to transform existing experiences. For example, DISKO could assist project maintainers with OSS team formation by supporting a search engine for potential collaborators based on skills or automatically recommending contributors for particular OSS roles. Furthermore, DISKO’s skill ratings could be used to identify a contributor’s potential areas of growth or recommend mentors with the expertise for the contributor’s professional goals. Skills could also be displayed publicly to GitHub profiles—users

could self-select skills they were proficient in, while the platform could display a verified badge for detected skills.

**Limitations.** One limitation of our evaluation is that self-rated skills is a biased measure, as participants may systematically overestimate or underestimate their skills. Additionally, limitations in describing skills in a survey may cause mismatched expectations of a skill. These may contribute to lower evaluation scores. For example, the signals for *Is familiar with OSS practices* was designed to be simple to compute and beginner friendly, but participants rated themselves more harshly on the skill. Thus, when designing experiences with automatic skills detection, transparency in how the skill rating is computed is paramount and should be communicated clearly.

## 6 FUTURE PLANS

Our preliminary results indicate that there are some promising directions for DISKO. However, additional steps are required to improve upon our current approach, which we outline below.

**Identifying additional signals from practitioners.** We hope to interview OSS contributors to understand how they evaluate their peers’ expertise for the skills our tool detects. This could generate new signals to improve the accuracy of our tool. 258 participants have agreed to be interviewed for this work.

**Defining weights of the signals.** In practice, each signal is not equally weighted in predicting a contributor’s skill. Since our current approach does not support this, one next step could use a linear regression model to examine the relationship between signals and participants’ self-evaluated skill level and then use the resulting model weights to weigh each signal in our tool.

**Evaluating the tool with other measures of skill.** We hope to run an additional evaluation using data sources less impacted by personal bias. For example, soft skills may be evaluated using peer evaluations, while hard skills may be evaluated using skill tests. Previous work has administered skill tests to determine Java expertise [12].

## 7 CONCLUSION

We present DISKO, a tool to detect OSS-related skills which identifies *signals* (i.e., measurable activities or cues associated with the skill) and then computes them from GitHub data. DISKO detects the following skills: *Teaches others to be involved in the OSS project*, *Shows commitment towards the OSS project*, *Has knowledge in specific programming languages*, and *Is familiar with OSS practices*. We demonstrate this approach yields positive results, as DISKO detects the presence of OSS skills with precision scores between 77% to 97%. Additionally, a near majority of participants find the detected skills important to OSS. We expect to improve the tool and perform more rigorous evaluations in the future. Future work could design tools to augment existing OSS experiences or improve upon our current approach. Our supplemental materials are publicly available at [26].

## ACKNOWLEDGMENTS

We thank our survey participants for their insight and Christian Bird, Mala Kumar, Victor Grau Serrat, and Lucy Harris for their feedback. Jenny T. Liang conducted this work for an internship at Microsoft Research’s Software Analysis and Intelligence Group.

## REFERENCES

- [1] 2021. The 2020 State of the Octoverse. Retrieved September 20, 2021 from <https://octoverse.github.com>.
- [2] 2021. Creating a default community health file. Retrieved September 20, 2021 from <https://docs.github.com/en/communities/setting-up-your-project-for-healthy-contributions/creating-a-default-community-health-file>.
- [3] 2021. GitHub Copilot - Your AI pair programmer. Retrieved September 20, 2021 from <https://copilot.github.com/>.
- [4] 2021. GitHub GraphQL API. Retrieved September 20, 2021 from <https://docs.github.com/en/graphql>.
- [5] 2021. GitHub REST API. Retrieved September 20, 2021 from <https://docs.github.com/en/rest>.
- [6] 2021. List of software package management systems. Retrieved September 20, 2021 from [https://en.wikipedia.org/wiki/List\\_of\\_software\\_package\\_management\\_systems](https://en.wikipedia.org/wiki/List_of_software_package_management_systems).
- [7] 2021. Stack Overflow Developer Survey. Retrieved September 20, 2021 from <https://insights.stackoverflow.com/survey/2020>.
- [8] 2022. All Contributors: Recognize all contributors. Retrieved January 25, 2022 from <https://allcontributors.org/>.
- [9] Faheem Ahmed, Luiz Fernando Capretz, and Piers Campbell. 2012. Evaluating the demand for soft skills in software development. *IT Professional* 14 (2012), 44–49.
- [10] Sebastian Baltes and Stephan Diehl. 2018. Towards a theory of software development expertise. In *ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 187–200.
- [11] Lingfeng Bao, Xin Xia, David Lo, and Gail C. Murphy. 2019. A large scale study of long-time contributor prediction for GitHub projects. *IEEE Transactions on Software Engineering* (2019).
- [12] Gunnar R. Bergersen, Dag I.K. Sjøberg, and Tore Dybå. 2014. Construction and validation of an instrument for measuring programming skill. *IEEE Transactions on Software Engineering* 40 (2014), 1163–1184.
- [13] Amreeta Chatterjee, Mariam Guizani, Catherine Stevens, Jillian Emard, Mary Evelyn May, Margaret Burnett, Iftekhar Ahmed, and Anita Sarma. 2021. AID: An automated detector for gender-inclusivity bugs in OSS project pages. In *IEEE/ACM International Conference on Software Engineering*, 1423–1435.
- [14] Moataz Chouchen, Ali Ouni, Raula Gaikovina Kula, Dong Wang, Patanamom Thongtanunam, Mohamed Wiem Mkaouer, and Kenichi Matsumoto. 2021. Antipatterns in modern code review: Symptoms and prevalence. In *IEEE International Conference on Software Analysis, Evolution and Reengineering*, 531–535.
- [15] Edson Dias, Paulo Meirelles, Fernando Castor, Igor Steinmacher, Igor Wiese, and Gustavo Pinto. 2021. What makes a great maintainer of open source projects?. In *IEEE/ACM International Conference on Software Engineering*, 982–994.
- [16] Vijaya Kumar Eluri, Thomas A. Mazzuchi, and Shahram Sarkani. 2021. Predicting long-time contributors for GitHub projects using machine learning. *Information and Software Technology* 138 (2021), 106616.
- [17] Georgios Gousios and Diomidis Spinellis. 2012. GHTorrent: GitHub's data from a firehose. In *IEEE Working Conference on Mining Software Repositories*, 12–21.
- [18] Georgios Gousios, Andy Zaidman, Margaret-Anne Storey, and Arie Van Deursen. 2015. Work practices and challenges in pull-based development: The integrator's perspective. In *IEEE/ACM International Conference on Software Engineering*, 358–368.
- [19] Gillian J. Greene and Bernd Fischer. 2016. CVExplorer: Identifying candidate developers by mining and exploring their open source contributions. In *IEEE/ACM International Conference on Automated Software Engineering*, 804–809.
- [20] Hideaki Hata, Nicole Novielli, Sebastian Baltes, Raula Gaikovina Kula, and Christoph Treude. 2021. GitHub Discussions: An exploratory study of early adoption. *arXiv preprint arXiv:2102.05230* (2021).
- [21] Claudia Hauff and Georgios Gousios. 2015. Matching GitHub developer profiles to job advertisements. In *IEEE/ACM Conference on Mining Software Repositories*, 362–366.
- [22] Yu Huang, Denae Ford, and Thomas Zimmermann. 2021. Leaving my fingerprints: Motivations and challenges of contributing to OSS for social good. In *IEEE/ACM International Conference on Software Engineering*, 1020–1032.
- [23] Barbara A. Kitchenham and Shari Lawrence Pfleeger. 2008. Personal Opinion Surveys. In *Guide to Advanced Empirical Software Engineering*, Forrest Shull, Janice Singer, and Dag I. K. Sjøberg (Eds.). Springer, 63–92. [https://doi.org/10.1007/978-1-84800-044-5\\_3](https://doi.org/10.1007/978-1-84800-044-5_3)
- [24] Amanda Lee and Jeffrey C. Carver. 2017. Are one-time contributors different? A comparison to core and periphery developers in floss repositories. In *ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 1–10.
- [25] Paul Luo Li, Amy J. Ko, and Andrew Begel. 2020. What distinguishes great software engineers? *Empirical Software Engineering* 25 (2020), 322–352.
- [26] Jenny T. Liang, Thomas Zimmermann, and Denae Ford. 2021. Supplemental Materials to “Towards Mining OSS Skills from GitHub Activity”. <https://doi.org/10.6084/m9.figshare.16814665>.
- [27] Cristina V. Lopes, Petr Maj, Pedro Martins, Vaibhav Saini, Di Yang, Jakub Zitny, Hitesh Sajani, and Jan Vitek. 2017. DéjàVu: a map of code duplicates on GitHub. *Proceedings of the ACM on Programming Languages* 1, OOPSLA (2017), 1–28.
- [28] Jennifer Marlow and Laura Dabbish. 2013. Activity traces and signals in software developer recruitment and hiring. In *ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 145–156.
- [29] Audris Mockus and James D. Herbsleb. 2002. Expertise Browser: A quantitative approach to identifying expertise. In *IEEE/ACM International Conference on Software Engineering*, 503–512.
- [30] Joao Eduardo Montandon, Luciana Lourdes Silva, and Marco Tulio Valente. 2019. Identifying experts in software libraries and frameworks among GitHub users. In *IEEE/ACM International Conference on Mining Software Repositories*, 276–287.
- [31] João Eduardo Montandon, Marco Tulio Valente, and Luciana L Silva. 2021. Mining the technical roles of GitHub users. *Information and Software Technology* 131 (2021), 106485.
- [32] Maria Papoutsoglou, Apostolos Ampatzoglou, Nikolaos Mittas, and Lefteris Angelis. 2019. Extracting knowledge from online sources for software engineering labor market: A mapping study. *IEEE Access* 7 (2019), 157595–157613.
- [33] Maria Papoutsoglou, Nikolaos Mittas, and Lefteris Angelis. 2017. Mining people analytics from StackOverflow job advertisements. In *IEEE EuroMicro Conference on Software Engineering and Advanced Applications*, 108–115.
- [34] Naveen Raman, Minxuan Cao, Yulia Tsvetkov, Christian Kästner, and Bogdan Vasilescu. 2020. Stress and burnout in open source: Toward finding, understanding, and mitigating unhealthy interactions. In *ACM/IEEE International Conference on Software Engineering*, 57–60.
- [35] Janet Siegmund, Christian Kästner, Jörg Liebig, Sven Apel, and Stefan Hanenberg. 2014. Measuring and modeling programming experience. *Empirical Software Engineering* 19, 5 (2014), 1299–1334.
- [36] Igor Steinmacher, Tayana Conte, Marco Aurélio Gerosa, and David Redmiles. 2015. Social barriers faced by newcomers placing their first contribution in open source software projects. In *ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1379–1392.
- [37] Bianca Trinkenreich, Mariam Guizani, Igor Wiese, Anita Sarma, and Igor Steinmacher. 2020. Hidden figures: Roles and pathways of successful OSS contributors. *ACM on Human-Computer Interaction* (2020).
- [38] Jason Tsay, Laura Dabbish, and James Herbsleb. 2014. Influence of social and technical factors for evaluating contribution in GitHub. In *IEEE/ACM International Conference on Software Engineering*, 356–366.
- [39] Sri Lakshmi Vadlamani and Olga Baysal. 2020. Studying software developer expertise and contributions in Stack Overflow and GitHub. In *IEEE International Conference on Software Maintenance and Evolution*, 312–323.
- [40] Minghui Zhou and Audris Mockus. 2012. What make long term contributors: Willingness and opportunity in OSS community. In *IEEE/ACM International Conference on Software Engineering*, 518–528.