Geographic diversity in public code contributions

Geographic Diversity in Public Code Contributions

An Exploratory Large-Scale Study Over 50 Years

Davide Rossi
daviderossi@unibo.it
University of Bologna
Bologna, Italy

Stefano Zacchiroli
stefano.zacchiroli@telecom-paris.fr
LTCI, Télécom Paris, Institut Polytechnique de Paris
Paris, France

ABSTRACT

We conduct an exploratory, large-scale, longitudinal study of 50 years of commits to publicly available version control system repositories, in order to characterize the geographic diversity of contributors to public code and its evolution over time. We analyze in total 2.2 billion commits collected by Software Heritage from 160 million projects and authored by 43 million authors during the 1971–2021 time period. We geolocate developers to 12 world regions derived from the United Nation geoscheme, using as signals email top-level domains, author names compared with names distributions around the world, and UTC offsets mined from commit metadata.

We find evidence of the early dominance of North America in open source software, later joined by Europe. After that period, the geographic diversity in public code has been constantly increasing. We also identify relevant historical shifts related to the UNIX wars, the increase of coding literacy in Central and South Asia, and broader phenomena like colonialism and people movement across countries (immigration/emigration).

KEYWORDS

geography, diversity, open source, commit, version control systems, social coding, software heritage

ACM Reference Format:
Davide Rossi and Stefano Zacchiroli. 2022. Geographic Diversity in Public Code Contributions: An Exploratory Large-Scale Study Over 50 Years. In *19th International Conference on Mining Software Repositories (MSR '22), May 23–24, 2022, Pittsburgh, PA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3524842.3528471

1 INTRODUCTION

Gender diversity, or more often its lack thereof, among participants to software development activities has been thoroughly studied in recent years. In particular, the presence of, effects of, and countermeasures for *gender bias* in Free/Open Source Software (FOSS) have received a lot of attention over the past decade [2, 13, 16, 18, 21, 24, 30, 31, 33]. *Geographic diversity* is on the other hand the kind of diversity that stems from participants in some global activity coming from different world regions and cultures.

MSR '22, May 23–24, 2022, Pittsburgh, PA, USA

ACM ISBN 978-1-4503-9303-4/22/05…$15.00
https://doi.org/10.1145/3524842.3528471

Geographic diversity in FOSS has received relatively little attention in scholarly works. In particular, while seminal survey-based and point-in-time medium-scale studies of the geographic origins of FOSS contributors exist [2, 6, 7, 25, 29, 32], large-scale longitudinal studies of the geographic origin of FOSS contributors are still lacking. Such a quantitative characterization would be useful to inform decisions related to global development teams [8] and hiring strategies in the information technology (IT) market, as well as contribute factual information to the debates on the economic impact and sociology of FOSS around the world.

Contributions. With this work we contribute to close this gap by conducting **the first longitudinal study of the geographic origin of contributors to public code over 50 years.** Specifically, we provide a preliminary answer to the following research question:

RQ 1. From which world regions do authors of publicly available commits come from and how has it changed over the past 50 years?

We use as dataset the Software Heritage archive [3] and analyze from it 2.2 billion commits archived from 160 million projects and authored by 43 million authors during the 1971–2021 time period. We geolocate developers to 12 world regions, using as signals email country code top-level domains (ccTLDs) and author (first/last) names compared with name distributions around the world, and UTC offsets mined from commit metadata.

We find evidence of the early dominance of North America in open source software, later joined by Europe. After that period, the geographic diversity in public code has been constantly increasing. We also identify relevant historical shifts related to the end of the UNIX wars and the increase of coding literacy in Central and South Asia, as well as of broader phenomena like colonialism and people movement across countries (immigration/emigration).

Data availability. A replication package for this paper is available from Zenodo at https://doi.org/10.5281/zenodo.6390355 [26].

2 RELATED WORK

Both early and recent works [2, 6, 18, 25] have characterized the geography of Free/Open Source Software (FOSS) using *developer surveys*, which provide high-quality answers but are limited in size (2-5 K developers) and can be biased by participant sampling.

In 2008 Barahona et al. [7] conducted a seminal large-scale (for the time) study on FOSS *geography using mining software repositories (MSR) techniques*. They analyzed the origin of 1 M contributors using the SourceForge user database and mailing list archives over the 1999–2005 period, using as signals information similar to ours: email domains and UTC offsets. The studied period (7 years) in [7]

is shorter than what is studied in the present paper (50 years) and the data sources are largely different; with that in mind, our results show a slightly larger quote of European v. North American contributions.

Another empirical work from 2010 by Takhteyev and Hilts [29] harvested self-declared geographic locations of GitHub accounts recursively following their connections, collecting information for ≈ 70 K GitHub users. A very recent work [32] by Wachs et al. has geolocated half a million GitHub users, having contributed at least 100 commits each, and who self-declare locations on their GitHub profiles. While the study is point-in-time as of 2021, the authors compare their findings against [7, 29] to characterize the evolution of FOSS geography over the time snapshots taken by the three studies.

Compared with previous empirical works, our study is much larger scale—having analyzed 43 million authors of 2.2 billion commits from 160 million projects—longitudinal over 50 years of public code contributions rather than point in time, and also more fine-grained (with year-by-year granularity over the observed period). Methodologically, our study relies on Version Control System (VCS) commit data rather than platform-declared location information.

Other works—in particular the work by Daniel [1] and, more recently, Rastogi et al. [20, 22, 23]—have studied geographic *diversity and bias*, i.e., the extent to which the origin of FOSS developers affect their collaborative coding activities. In this work we characterized geographic diversity in public code for the first time at this scale, both in terms of contributors and observation period. We do not tackle the bias angle, but provide empirical data and findings that can be leveraged to that end as future work.

Global software engineering [8] is the sub-field of software engineering that has analyzed the challenges of scaling developer collaboration globally, including the specific concern of how to deal with geographic diversity [5, 9]. Decades later the present study provides evidence that can be used, in the specific case of public code and at a very large scale, to verify which promises of global software engineering have borne fruit.

3 METHODOLOGY

Dataset. We retrieved from Software Heritage [19] all commits archived until 2021-07-07. They amount to 2 198 808 389 commits, unique by SHA1 identifier, harvested from about 160 million public projects coming from major development forges (GitHub, GitLab, etc.) and package repositories (Debian, PyPI, NPM, etc.). Commits in the dataset are by 43 381 366 authors, unique by ⟨name, email⟩ pairs. The dataset came as two relational tables, one for commits and one for authors, with the former referencing the latter via a foreign key. For each entry in the author table we have author full name and email as two separate strings of raw bytes.

We removed implausible or unusable names that: are not decodable as UTF-8 (4127 author names removed), are email addresses instead of names (84 954 "names"), consist of only blank characters (25 055), contain more than 10% non-letters (9 915 884), are longer than 100 characters (46). After filtering, about 33 M authors (77% of the initial dataset) remained for further analysis.

Note that the amount of public code commits (and authors) contained in the initial dataset grows exponentially over time [28],

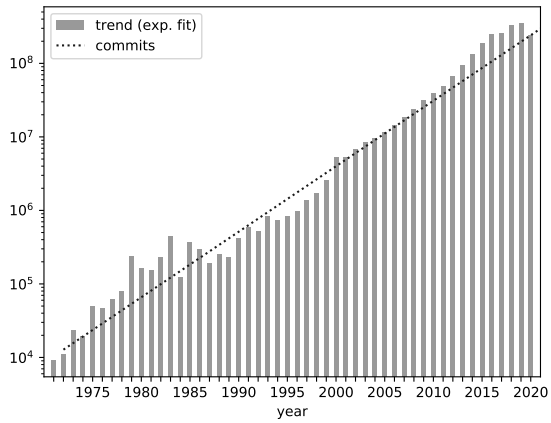

Figure 1: Yearly public commits over time (log scale).

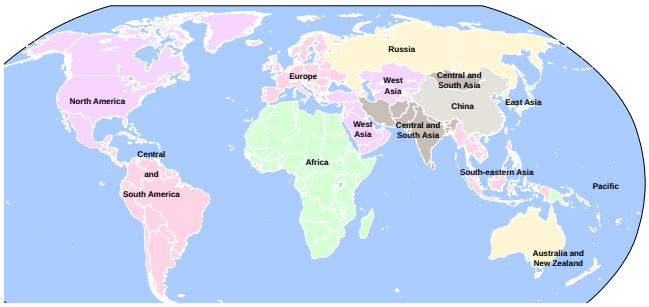

Figure 2: The 12 world regions used as geolocation targets.

as shown for commits in Figure 1. As a consequence the observed trends tend to be more stable in recent decades than in 40+ year-old ones, due to statistics taken on exponentially larger populations.

Geolocation. As geolocation targets we use macro world regions derived from the United Nations geoscheme [17]. To avoid domination by large countries (e.g., China or Russia) within macro regions, we merged and split some regions based on geographic proximity and the sharing of preeminent cultural identification features, such as spoken language. Figure 2 shows the final list of 12 world regions used as geolocation targets in this study.

Geolocation of commit authors to world regions uses the two complementary techniques introduced in [27], briefly recalled below. The first one relies on the country code top-level domain (ccTLD) of email addresses extracted from commit metadata, e.g., `.fr`, `.ru`, `.cn`, etc. We started from the IANA list of Latin character ccTLDs [11] and manually mapped each corresponding territory to a target world region.

The second geolocation technique uses the UTC offset of commit timestamps (e.g., UTC-05:00) and author names to determine the most likely world region of the commit author. For each UTC offset we determine a list of compatible places (country, state, or dependent territory) in the world that, at the time of that commit, had that UTC offset; commit time is key here, as country UTC offsets vary

over time due to timezone changes. To make this determination we use the IANA time zone database [10].

Then we assign to each place a score that captures the likelihood that a given author name is characteristic of it. To this end we use the Forebears dataset of the frequencies of the most common first and family names which, quoting from [4]: *"provides the approximate incidence of forenames and surnames produced from a database of* 4 044 546 938 *people (55.5% of living people in 2014). As of September 2019 it covers* 27 662 801 *forenames and* 27 206 821 *surnames in 236 jurisdictions."* As in our dataset authors are full name strings (rather than split by first/family name), we first tokenize names (by blanks and case changes) and then lookup individual tokens in both first and family names frequency lists. For each element found in name lists we multiply the place population[1] by the name frequency to obtain a measure that is proportional to the number of persons bearing that name (token) in the specific place. We sum this figure for all elements to obtain a place score, ending up with a list of ⟨place, score⟩ pairs. We then partition this list by the world region that a place belongs to and sum the score for all the places in each region to obtain an overall score, corresponding to the likelihood that the commit belongs to a given world region. We assign the starting commit as coming from the world region with the highest score.

The email-based technique suffers from the limited and unbalanced use of ccTLDs: most developers use generic TLDs such as `.com`, `.org`, or `.net`. Moreover this does not happen uniformly across zones: US-based developers, for example, use the `.us` ccTLD much more seldomly than their European counterparts. On the other hand the offset/name-based technique relies on the UTC offset of the commit timestamps. Due to tool configurations on developer setups, a large number of commits in the dataset has an UTC offset equal to zero. This affects less recent commits (22% of 2020s commits have a zero offset) than older ones (96% in 2000). As a result the offset/name-based technique could end up detecting a large share of older commits as authored by African developers, and to a lesser extent Europeans.

To counter these issues we combine the two geolocation techniques together by applying the offset/name-based techniques to all commits with a non-zero UTC offset, and the email-based on to all other commits.

4 RESULTS AND DISCUSSION

To answer RQ 1 we gathered the number of commits and distinct authors per year and per world zone. We present the obtained results in Figure 3 as two stacked bar charts, showing yearly breakdowns for commits and authors respectively. Every bar represents a year and is partitioned in slices showing the commit/author ratio for each of the world regions of Figure 2 in that year. To avoid outliers due to sporadic contributors, in the author chart we only consider authors having contributed at least 5 commits in a given year.

While observing trends in the charts remember that the total numbers of commits and authors grow exponentially over time. Hence for the first years in the charts, the number of data points in some world regions can be extremely small, with negative consequences on the stability of trends.

Geographic diversity over time. Overall, the general trend appears to be that the **geographic diversity in public code is increasing**: North America and Europe alternated their "dominance" until the middle of the 90s; from that moment on most other world regions show a slow but steady increment. This trend of increased participation into public code development includes Central and South Asia (comprising India), Russia, Africa, Central and South America, Notice that also zones that do not seem to follow this trend, such as Australia and New Zealand, are also increasing their participation, but at a lower speed with respect to other zones. For example, Australia and New Zealand incremented the absolute number of their commits by about 3 orders of magnitude from 2000 to present days.

Another interesting phenomenon that can be appreciated in both charts is the sudden contraction of contributions from North America in 1995; since the charts depict ratios, this corresponds to other zones, and Europe in particular, increasing their share. An analysis of the main contributions in the years right before the contraction shows that nine out of ten have `ucbvax.Berkeley.EDU` as author email domain, and the tenth is Keith Bostic, one of the leading Unix BSD developers, appearing with email `bostic`. No developer with the same email domain appears anymore within the first hundred contributors in 1996. This shows the relevance that BSD Unix and the Computer Systems Research Group at the University of California at Berkeley had in the history of open source software. The group was disbanded in 1995, partially as a consequence of the so-called UNIX wars [12], and this contributes significantly—also because of the relatively low amount of public code circulating at the time—to the sudden drop of contributions from North America in subsequent years. Descendant UNIX operating systems based on BSD, such as OpenBSD, FreeBSD, and NetBSD had smaller relevance to world trends due to (i) the increasing amount of open source code coming from elsewhere and (ii) their more geographically diverse developer community.

Another time frame in which the ratios for Europe and North America are subject to large, sudden changes is 1975–79. A preliminary analysis shows that these ratios are erratic due to the very limited number of commits in those time period, but we were unable to detect a specific root cause. Trends for those years should be subject to further studies, in collaboration with software historians.

Colonialism. Another trend that stands out from the charts is that Africa appears to be well represented. To assess if this results from a methodological bias, we double-checked the commits detected as originating from Africa for timezones included in the $[0, 3]$ range using both the email- the offset/name-based methods. The results show that the offset/name-based approach assigns 22.7% of the commits to Africa whereas the email-based one only assigns 2.7% of them. While a deeper investigation is in order, it is our opinion that the phenomenon we are witnessing here is a consequence of colonialism, specifically the adoption of Europeans names in African countries. For example the name Eric, derived from Old Norse, is more popular in Ghana than it is in France or in the UK. This challenges the ability of the offset/name-based method to correctly differentiate between candidate places. Together with the fact that several African countries are largely populated, the

[1] To obtain population totals—as the notion of "place" is heterogeneous: full countries v. slices of large countries spanning multiple timezones—we use a mixture of primary sources (e.g., government websites), and non-primary ones (e.g., Wikipedia articles).

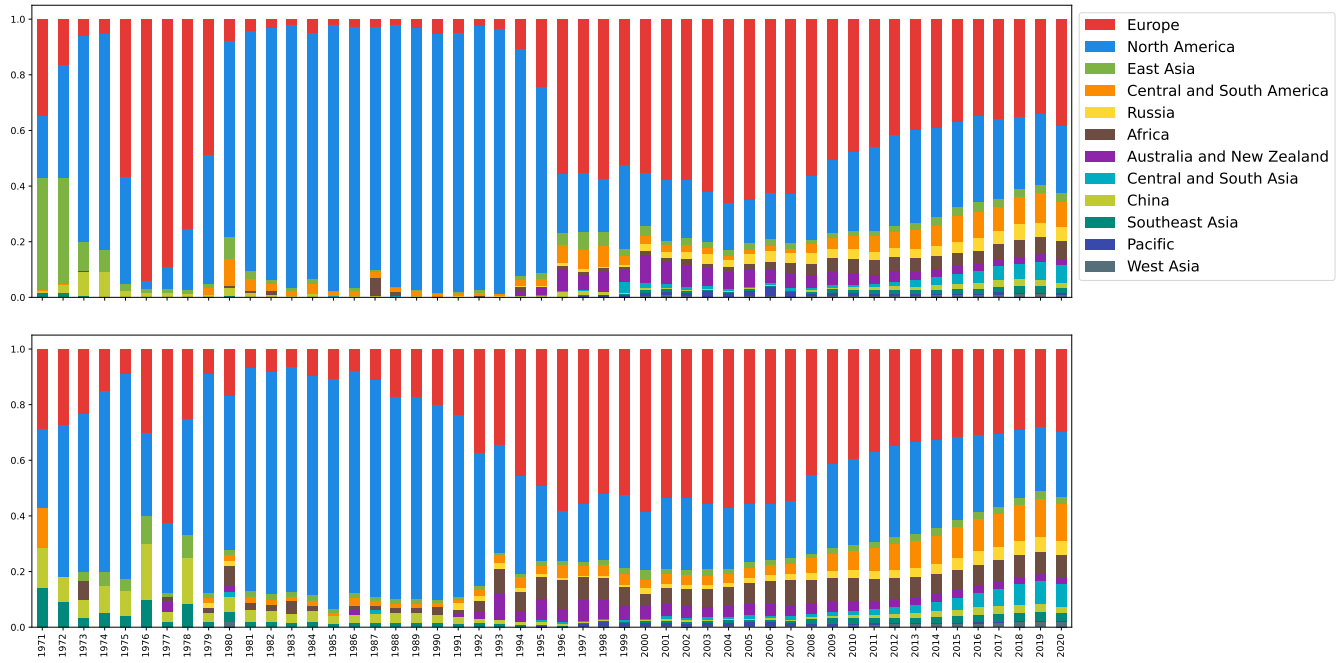

Figure 3: Ratio of commits (above) and active authors (below) by world zone over the 1971–2020 period.

offset/name-based method could detect European names as originating from Africa. While this cuts both way, the likelihood of a random person contributing to public code is very different between European countries, all having a well-developed software industry, and African countries that do not all share this trait.

Immigration/emigration. Another area where a similar phenomenon could be at play is the evolution of Central and South America. Contribution from this macro region appears to be growing steadily. To assess if this is the result of a bias introduced by the name-based detection we analyzed the evolution of offset/name-based assignment over time for authors whose email domain is among the top-ten US-based entities in terms of overall contributions (estimated in turn by analyzing the most frequent email domains and manually selecting those belonging to US-based entities). In 1971 no author with an email from top US-based entities is detected as belonging to Central and South America, whereas in 2019 the ratio is 12%. Nowadays more than one tenth of the people email-associated to top US-based entities have popular Central and South American names, which we posit as a likely consequence of immigration into US (emigration from Central and South America). Since immigration has a much longer history than what we are studying here, what we are witnessing probably includes long-term consequences of it, such as second and third generation immigrants employed in white-collar jobs, such as software development.

5 LIMITATIONS AND FUTURE WORK

We have performed an exploratory, yet very large scale, empirical study of the geographic diversity in public code commits over time. We have analyzed 2.2 billion public commits covering the 1971–2021 time period. We have geolocated developers to 12 world regions using as signals email domains, timezone offsets, and author names. Our findings show that the geographic diversity in public code is increasing over time, and markedly so over the past 20–25 years. Observed trends also co-occur with historical events and macro phenomena like the end of the UNIX wars, increase of coding literacy around the world, colonialism, and immigration.

Limitations. This study relies on a combination of two geolocation methods: one based on email domains, another based on commit UTC offsets and author names. We discussed some of the limitations of either method in Section 3, motivating our decision of restricting the use of the email-based method to commits with a zero UTC offset. As a consequence, for most commits in the dataset the offset/name-based method is used. With such method, the frequencies of forenames and surnames are used to rank candidate zones that have a compatible UTC offset at commit time.

A practical consequence of this is that for commits with, say, offset UTC+09:00 the candidate places can be Russia, Japan and Australia, depending on the specific date due to daylight saving time. Popular forenames and surnames in these regions tend to be quite different so the likelihood of the method to provide a reliable detection is high. For other offsets the set of popular forenames and surnames from candidate zones can exhibit more substantial overlaps, negatively impacting detection accuracy. We have discussed some of these cases in Section 4, but other might be lingering in the results impacting observed trends.

The choice of using the email-based method for commits with zero UTC offset, and the offset/name-based method elsewhere, has allowed us to study all developers not having a country-specific email domain (ccTLD), but comes with the risk of under-representing

the world zones that have (in part and in some times of the year) an actual UTC offset of zero.

A potential bias in this study could be introduced by the fact that the name database used for offset/name-based geolocation only contains names formed using Latin alphabet characters. We looked for names containing Chinese, Japanese, and Korean characters in the original dataset, finding only a negligible amount of authors who use non-Latin characters in their VCS names, which leads us to believe that the impact of this issue is minimal.

We did not apply identity merging (e.g., using state-of-the-art tools like SortingHat [15]), but we do not expect this to be a significant issue because: (a) to introduce bias in author trends the distribution of identity merges around the world should be uneven, which seems unlikely; and (b) the observed commit trends (which would be unaffected by identity merging) are very similar to observed author trends.

We did not systematically remove known bot accounts [14] from the author dataset, but we did check for the presence of software bots among the top committers of each year. We only found limited traces of continuous integration (CI) bots, used primarily to automate merge commits. After removing CI bots from the dataset the observed global trends were unchanged, therefore this paper presents unfiltered data.

Future work. To some extent the above limitations are the price to pay to study such a large dataset: there exists a trade-off between large-scale analysis and accuracy. We plan nonetheless to further investigate and mitigate them in future work. Multi-method approaches, merging data mining with social science methods, could be applied to address some of the questions raised in this exploratory study. While they do not scale to the whole dataset, multi-methods can be adopted to dig deeper into specific aspects, specifically those related to social phenomena. Software is a social artifact, it is no wonder that aspects related to sociocultural evolution emerge when analyzing its evolution at this scale.

REFERENCES

[1] Sherae L. Daniel, Ritu Agarwal, and Katherine J. Stewart. 2013. The Effects of Diversity in Global, Distributed Collectives: A Study of Open Source Project Success. *Inf. Syst. Res.* 24, 2 (2013), 312–333. https://doi.org/10.1287/isre.1120.0435

[2] Paul A David and Joseph S Shapiro. 2008. Community-based production of open-source software: What do we know about the developers who participate? *Information Economics and Policy* 20, 4 (2008), 364–398.

[3] Roberto Di Cosmo and Stefano Zacchiroli. 2017. Software Heritage: Why and How to Preserve Software Source Code. In *Proceedings of the 14th International Conference on Digital Preservation, iPRES 2017.* https://hal.archives-ouvertes.fr/hal-01590958/

[4] Forebears. 2021. World Forename & Surname Distribution Maps. Online at https://forebears.io/about/name-distribution-and-demographics, accessed during April 2021.

[5] Steven Fraser, Dennis Mancl, Aki Namioka, Roberto Salama, and Allen Wirfs-Brock. 2014. East meets west: the influences of geography on software production. In *Conference on Systems, Programming, and Applications: Software for Humanity, SPLASH '14, Portland, OR, USA, October 20-24, 2014 - Companion Volume*, Andrew P. Black (Ed.). ACM, 41–42. https://doi.org/10.1145/2660252.2661293

[6] Rishab Aiyer Ghosh. 2005. Understanding free software developers: Findings from the FLOSS study. In *Perspectives on free and open source software*. Vol. 28. MIT Press, 23–47.

[7] Jesús M. González-Barahona, Gregorio Robles, Roberto Andradas-Izquierdo, and Rishab Aiyer Ghosh. 2008. Geographic origin of libre software developers. *Inf. Econ. Policy* 20, 4 (2008), 356–363. https://doi.org/10.1016/j.infoecopol.2008.07.001

[8] James D. Herbsleb. 2007. Global Software Engineering: The Future of Socio-technical Coordination. In *International Conference on Software Engineering, ISCE 2007, Workshop on the Future of Software Engineering, FOSE 2007, May 23-25, 2007, Minneapolis, MN, USA*, Lionel C. Briand and Alexander L. Wolf (Eds.). IEEE Computer Society, 188–198. https://doi.org/10.1109/FOSE.2007.11

[9] Helena Holmström, Eoin Ó Conchúir, Pär J. Ågerfalk, and Brian Fitzgerald. 2006. Global Software Development Challenges: A Case Study on Temporal, Geographical and Socio-Cultural Distance. In *1st IEEE International Conference on Global Software Engineering, ICGSE 2006, Florianopolis, Brazil, October 2006.* IEEE Computer Society, 3–11. https://doi.org/10.1109/ICGSE.2006.261210

[10] IANA. 2017. Time Zone Database. https://data.iana.org/time-zones/releases/ Retrieved 2021-09-28.

[11] IANA. 2021. Country code top-level domains. Mirrored at https://en.wikipedia.org/wiki/Country_code_top-level_domain#Latin_Character_ccTLDs, accessed 2021-10-06.

[12] Brian W Kernighan. 2019. *UNIX: A History and a Memoir.* Independently published. https://www.cs.princeton.edu/~bwk/memoir.html, accessed 2022-01-10.

[13] Victor Kuechler, Claire Gilbertson, and Carlos Jensen. 2012. Gender Differences in Early Free and Open Source Software Joining Process. In *8th International Conference on Open Source Systems, OSS 2012 (IFIP Advances in Information and Communication Technology, Vol. 378)*. Springer, 78–93. https://doi.org/10.1007/978-3-642-33442-9_6

[14] Carlene Lebeuf, Margaret-Anne D. Storey, and Alexey Zagalsky. 2018. Software Bots. *IEEE Softw.* 35, 1 (2018), 18–23. https://doi.org/10.1109/MS.2017.4541027

[15] David Moreno, Santiago Dueñas, Valerio Cosentino, Miguel Angel Fernández, Ahmed Zerouali, Gregorio Robles, and Jesús M. González-Barahona. 2019. SortingHat: wizardry on software project members. In *Proceedings of the 41st International Conference on Software Engineering: Companion Proceedings, ICSE 2019, Montreal, QC, Canada, May 25-31, 2019*, Joanne M. Atlee, Tevfik Bultan, and Jon Whittle (Eds.). IEEE / ACM, 51–54. https://doi.org/10.1109/ICSE-Companion.2019.00036

[16] Dawn Nafus. 2012. 'Patches don't have gender': What is not open in open source software. *New Media & Society* 14, 4 (2012), 669–683.

[17] United Nations. 1999. *Standard country or area codes for statistical use.* Technical Report. United Nations. https://unstats.un.org/unsd/methodology/m49/ Retrieved 2021-09-27.

[18] Mathieu O'Neil, Mahin Raissi, Molly de Blanc, and Stefano Zacchiroli. 2017. Preliminary Report on the Influence of Capital in an Ethical-Modular Project: Quantitative data from the 2016 Debian Survey. *Journal of Peer Production* 10 (2017).

[19] Antoine Pietri, Diomidis Spinellis, and Stefano Zacchiroli. 2019. The Software Heritage graph dataset: public software development under one roof. In *Proceedings of the 16th International Conference on Mining Software Repositories, MSR 2019, 26-27 May 2019, Montreal, Canada.*, Margaret-Anne D. Storey, Bram Adams, and Sonia Haiduc (Eds.). IEEE / ACM, 138–142. https://dl.acm.org/citation.cfm?id=3341907

[20] Gede Artha Azriadi Prana, Denae Ford, Ayushi Rastogi, David Lo, Rahul Purandare, and Nachiappan Nagappan. 2021. Including Everyone, Everywhere: Understanding Opportunities and Challenges of Geographic Gender-Inclusion in OSS. *IEEE Transactions on Software Engineering* (2021). https://doi.org/10.1109/TSE.2021.3092813 to appear.

[21] Yixin Qiu, Katherine J. Stewart, and Kathryn M. Bartol. 2010. Joining and Socialization in Open Source Women's Groups: An Exploratory Study of *KDE-Women*. In *6th International Conference on Open Source Systems, OSS 2010 (IFIP Advances in Information and Communication Technology, Vol. 319)*. Springer, 239–251. https://doi.org/10.1007/978-3-642-13244-5_19

[22] Ayushi Rastogi. 2016. Do biases related to geographical location influence work-related decisions in GitHub?. In *Proceedings of the 38th International Conference on Software Engineering, ICSE 2016, Austin, TX, USA, May 14-22, 2016 - Companion Volume*, Laura K. Dillon, Willem Visser, and Laurie A. Williams (Eds.). ACM, 665–667. https://doi.org/10.1145/2889160.2891035

[23] Ayushi Rastogi, Nachiappan Nagappan, Georgios Gousios, and André van der Hoek. 2018. Relationship between geographical location and evaluation of developer contributions in GitHub. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, ESEM 2018, Oulu, Finland, October 11-12, 2018*, Markku Oivo, Daniel Méndez Fernández, and Audris Mockus (Eds.). ACM, 22:1–22:8. https://doi.org/10.1145/3239235.3240504

[24] Gregorio Robles, Laura Arjona Reina, Jesús M. González-Barahona, and Santiago Dueñas Domínguez. 2016. Women in Free/Libre/Open Source Software: The Situation in the 2010s. In *12th International Conference on Open Source Systems, OSS 2016 (IFIP Advances in Information and Communication Technology, Vol. 472)*. Springer, 163–173. https://doi.org/10.1007/978-3-319-39225-7_13

[25] Gregorio Robles, Laura Arjona Reina, Alexander Serebrenik, Bogdan Vasilescu, and Jesús M. González-Barahona. 2014. FLOSS 2013: a survey dataset about free software contributors: challenges for curating, sharing, and combining. In *11th Working Conference on Mining Software Repositories, MSR 2014, Proceedings, May 31 - June 1, 2014, Hyderabad, India*, Premkumar T. Devanbu, Sung Kim, and Martin Pinzger (Eds.). ACM, 396–399. https://doi.org/10.1145/2597073.2597129

[26] Davide Rossi and Stefano Zacchiroli. 2022. *Geographic Diversity in Public Code Contributions - Replication Package.* https://doi.org/10.5281/zenodo.6390355

[27] Davide Rossi and Stefano Zacchiroli. 2022. Worldwide Gender Differences in Public Code Contributions (and How They Have Been Affected by the COVID-19 Pandemic). In *44th International Conference on Software Engineering (ICSE 2022) - Software Engineering in Society (SEIS) Track*. ACM. https://doi.org/10.1145/3510458.3513011

[28] Guillaume Rousseau, Roberto Di Cosmo, and Stefano Zacchiroli. 2020. Software Provenance Tracking at the Scale of Public Source Code. *Empirical Software Engineering* 25, 4 (2020), 2930–2959. https://doi.org/10.1007/s10664-020-09828-5

[29] Yuri Takhteyev and Andrew Hilts. 2010. Investigating the geography of open source software through GitHub. https://flosshub.org/sites/flosshub.org/files/Takhteyev-Hilts-2010.pdf.

[30] Josh Terrell, Andrew Kofink, Justin Middleton, Clarissa Rainear, Emerson Murphy-Hill, Chris Parnin, and Jon Stallings. 2017. Gender differences and bias in open source: Pull request acceptance of women versus men. *PeerJ Computer Science* 3 (2017), e111.

[31] Bogdan Vasilescu, Andrea Capiluppi, and Alexander Serebrenik. 2014. Gender, representation and online participation: A quantitative study. *Interacting with Computers* 26, 5 (2014), 488–511.

[32] Johannes Wachs, Mariusz Nitecki, William Schueller, and Axel Polleres. 2021. The Geography of Open Source Software: Evidence from GitHub. *CoRR* abs/2107.03200 (2021). arXiv:2107.03200 https://arxiv.org/abs/2107.03200

[33] Stefano Zacchiroli. 2021. Gender Differences in Public Code Contributions: A 50-Year Perspective. *IEEE Softw.* 38, 2 (2021), 45–50. https://doi.org/10.1109/MS.2020.3038765