The Geography of Open Source Software: Evidence from GitHub

Johannes Wachs
Vienna Univ. of Economics and Business
Complexity Science Hub Vienna
johannes.wachs@wu.ac.at

Mariusz Nitecki
Vienna Univ. of Economics and Business

William Schueller
Medical Univ. of Vienna
Complexity Science Hub Vienna

Axel Polleres
Vienna Univ. of Economics and Business
Complexity Science Hub Vienna

July 7, 2021

Open Source Software plays an important role in the digital economy. Yet although software production is amenable to remote collaboration and its end products are easily shared across distances, software development seems to cluster geographically in places such as Silicon Valley, London, or Berlin. And while recent work indicates that positive effects of open source software production accrue locally through knowledge spillovers and information effects, up-to-date data on the geographic distribution of active open source developers remains limited. Here we analyze the geographic distribution of more than half a million active contributors to GitHub located in early 2021 at various spatial scales. Comparing our data with results from before 2010, we find a significant increase in the relative share of developers based in Asia, Latin America and Eastern Europe, suggesting a more even spread of OSS developers globally. Within countries, however, we find significant concentration in regions, exceeding by some margin the concentration of workers in high-tech fields. We relate OSS activity to a number of social and technological indicators at both scales using a multiple regression framework. Despite the potential of OSS as a distributed mode of collaborative work, the data suggest that OSS activity remains highly localized.

***K*eywords** Geography, open source software, GitHub, innovation, future of work

The importance of software, both as a ubiquitous complement to other activities in the modern economy and as a key sector in its own right, is widely acknowledged [4, 50]. Software is also an especially global industry, in part because its end products are easily shared and distributed through the Web. Yet the notion that the software industry might transcend geographic constraints is inconsistent with anecdotal observations that its most coveted jobs and leading firms tend to cluster in particular places like Silicon Valley, London, or Berlin [30]. And if influence of software on our economy and society is growing [73], the extent and causes of geographic concentration in software production are likely to shape this influence.

Within the vast ecosystem of software, open source software (OSS) plays a distinguished role [16], and is sometimes described as the infrastructure of the digital society. OSS contributions are also thought to be intensely concentrated in space, despite the low cost of distributing software, the development of technologies for remote collaboration, and its inherently open nature. Research from 2010 estimated that 7.4% of global OSS activity at the time took place in or around San Francisco [64]. As OSS activity is known to impact local firm productivity [48, 50] and rates of technology entrepreneurship [76], its geographic clustering likely effects economic growth and inequality, hence should be of interest to policymakers [49]. Geography disparities may also influence who participates in OSS, and understanding them better can inform us about root causes of diversity issues in software [2, 56] and its role in innovation [13].

So while the geographic distribution of OSS contributors likely has important effects on our society and economy, we lack an up-to-date geographic mapping of OSS activity. Previous work was either carried out over ten years ago [26, 64], focuses on data at the country-level [49, 76], and/or on a subset of projects [56]. We suggest that there is need for a fresh look at the geographic distribution of OSS developers, including regional data. This gap is becoming more relevant as researchers take a greater interest in the influence of algorithms and software on society [73]. If algorithms do shape the fabric of our society, the hypothesis that software is created by a distinct group of people living in a few particular places merits testing. At the same time, software engineering researchers have gathered massive datasets on open source software contributions [28, 53, 23], presenting opportunities to study this question in greater depth.

In light of the explosive growth of data on software activity and the apparent scientific interest in the geography of this activity, we present and implement a pipeline to geolocate highly active contributors to open source software projects on GitHub. We share both country-level and sub-national counts of active OSS developers. Using this data, we show that open source activity correlates strongly with human and economic development, though with significant variation. Using multiple linear regression models, we find that social and political factors such as generalized trust, human development, and quality of government have a significant relationship with the number of active OSS contributors in a country and in regions. In contrast to studies from over ten years ago, we find a more even global distribution of OSS activity, with Asian, Latin American, and Eastern European countries contributing a greater overall share. Within countries, however, we find that OSS remains highly concentrated in particular regions. OSS activity is even more spatially concentrated than the university educated population and workers in high-tech sectors.

The rest of the paper is organized as follows. We first discuss related works on the importance of OSS in the digital economy both in general and locally. We then introduce our data collection process, describe how we geolocate developers, and share access to data on counts of active developers in various geographic units. We then analyze the distribution of developers at both the country and regional levels. We conclude with a discussion of limitations, potential uses for this dataset, and ideas for future work in this area.

Related Works

In this section we review related works on both OSS in particular and on the geography of innovation and technology in general. We outline the emerging evidence that OSS has a significant impact on local growth and innovation outcomes, and suggest theoretical reasons why software activity clusters in space.

OSS, Innovation, and Growth

Researchers are increasingly recognizing the importance of software and OSS activity in particular as a key component of growth in the digital economy [48]. OSS products contribute substantial amounts to global GDP and productivity [12, 29]. Besides the specific outputs of the OSS sector, software know-how itself is clearly an important input in emerging sectors including Industry 4.0 [6] and artificial intelligence, for instance via the intensely computational methods of deep-learning [39]. It also contributes to productivity gains in manufacturing [10], suggesting that software has a complementary role in many sectors of the modern economy [51], including for example biotechnology via bioinformatics [35]. OSS in particular has advantages over proprietary models of software development in several contexts [57, 41, 1], which have led to growing adoption of and participation in OSS, including by firms with a traditionally proprietary orientation such as Microsoft and Apple [5]. Companies increasingly release in-house projects as OSS, for example Google's TensorFlow library for machine learning and Facebook's React web framework. OSS also helps startups compete with established players in digital sectors [27].

But what are the mechanisms by which OSS fosters economic dynamism locally? Wright et al. [76] demonstrate a strong relationship between OSS activity and IT-adjacent entrepreneurship. They present three reasons why OSS facilitates innovation. First, *people who often contribute to OSS tend to have valuable know-how and human capital*. Beyond the selection effect of skilled developers into OSS, OSS developers are also constantly learning and developing both their technical and coordinating skills. This framing presents OSS development as an incubator for new software ventures, recalling the "doing-using-interacting" model of innovation [36, 3].

Second, Wright et al. highlight *the value of OSS contributions as a knowledge sharing phenomenon*. Sharing and re-using code facilitates the spread of better solutions. Bugs in public code are likely to be spotted more quickly [57], both because of increased visibility and widespread usage. The popularity of specific code can highlight trends and needs in the larger software market. Because of its open nature, OSS developers can effectively signal their specific areas of expertise, broadening the network of potential collaborators in a new venture. Public and transparent information about how people and teams work together makes it easier to attract outside investments [38], to get quality feedback [72], and to learn by observing [58].

Finally, *a large OSS footprint suggests that a location has complementary assets necessary for software entrepreneurship*. This presents OSS activity as an effective proxy measurement for a place's technological development and capacity for IT innovation. As we will demonstrate, OSS activity is only moderately correlated with economic development, and this correlation decreases among the most developed countries. Indeed, we will show that social and political development indicators explain a significant share of the variance between OSS activity between countries and regions, above and beyond what can be explained simply by economic development.

Geography and Collaboration in Software

Even though software is easily shared and distributed, software creation has long been intensely geographically clustered [8]. A 2015 study claimed that 40% of all jobs in the US software industry were clustered in merely nine cities [30]. Geographic proximity to customers, end-users, and other software developers improves developer productivity and software quality via spillover effects [75] including knowledge transfers [68]. Physical proximity is known to accelerate the formation of collaboration ties and network effects in knowledge sectors [37]. The primacy of Silicon Valley in the software industry can be thought of as the end result of such agglomerative effects. Indeed previous work on OSS [64] using data from the late 2000s finds a strong negative relationship between the distance separating two OSS developers and their likelihood of collaboration, replicating a classic result of economic geography [55, 11] in the OSS context. However, several of these mechanisms that explain why software companies cluster do not clearly transfer to the context of OSS. For example, users and competitors in the OSS space are likely more geographically dispersed than in traditional industries. That OSS still clusters geographically in spite of these factors suggests that these are very powerful forces. Ironically, knowledge spillovers and transfer in the virtual, global collaboration network of OSS developers [20] may result in intensely localized pockets of expertise and learning.

Data

We now describe our data collection and processing pipeline. Before gathering information about activity levels and positions of individual OSS contributors, one needs to define contributors and their activity. Developers share code using version control protocols – allowing them to track changes, compare, test and merge with modifications of others – on specific online platforms. The most widely used protocol is called *git*, and the most widely used public platform for projects using git is *GitHub*. Modifications to files are collected in *commits*, which can be seen as snapshots of code edits. Developers can commit small and large modifications, though often commit code before switching tasks or finishing a unit of work. *We use commits, as elemental contributions in OSS, to quantify the activity of individual developers*.

Data on commits contributed to public projects on GitHub is made available and dynamically updated on the GH Archive database[1]. We use this database rather than querying data from the GitHub API. The next step is to assign GitHub accounts of authors to their commits. Commits themselves contain plaintext names and email addresses of

[1] `https://www.gharchive.org/`

authors, which do not correspond directly to GitHub accounts. For instance, a GitHub account user may contribute commits from multiple computers, each linked to git via a different name and email. Merging these identities under one developer is a crucial part in geolocating GitHub users, as the clearest information about their geographic location is provided on their GitHub profile page. We therefore applied a method from the software engineering research community to link email addresses and specific commits to GitHub user accounts (which we assume correspond to individual developers) [46] using the GitHub API: for each email address, we select a random commit made by that email address and query the GitHub API to retrieve the specific account login associated to the commit. In case the API cannot resolve the account using this first commit, we try three additional commits submitted using this email. As we are interested in active GitHub contributors, we consider all email addresses with at least 100 commits over the two years of 2019 and 2020. This corresponds to an average of nearly one commit per week.

Having associated GitHub accounts to contributions, we access information about users via the GitHub API. In particular we access the *location* and *Twitter account* of individual users, when provided. Through the *Twitter API*, one can also retrieve location information of Twitter users, when provided. A third way of gathering information on location is through email suffixes, belonging either to a country or institutions such as universities. We describe our method of geolocating developers in the following section.

Geolocation

Given a collection of active GitHub contributors and their account information, our goal is to infer the location for as many users as possible. We first focused on the raw location fields provided on GitHub user profiles. Reflecting on the common pitfalls of geocoding online profiles, for instance that users may give unreal or sarcastic locations ("the moon") [33], we selected the Bing Maps API. This API resolves multiple input languages ("Vienna" and "Wien" refer to the same location) and can handle inputs at varying scales from country to geolocation precise to within meters.

In case a user did not share a location on GitHub, or the Bing API was not able to geolocate the string that the user did share, we check whether the user linked to a Twitter account. If so, we attempted to geolocate the Twitter account in a similar way, using the location field provided by the user on Twitter. In case we could not geolocate a user from their Twitter data, we considered the email suffixes under which they made commits. Email suffixes can suggest the location of individuals in two ways: by the country domain (i.e. .jp for Japan, .it for Italy) or by a university suffix. In the latter case we imitate previous work associating GitHub contributors to universities [71] using a list of universities, their locations and email domains maintained by Hipo[2]. We only infer user country from email suffix data.

Using this pipeline we could geolocate 587,852 active OSS contributors (out of 1,124,874 accounts with at least 100 commits) to at least the country level. 502,415 or 85% user locations were identified from their GitHub account information alone. In other words, by considering email suffixes and Twitter data we could increase our pool of geolocated developers by 15%. As country-identifying email domains and Twitter use vary significantly between countries, we share data on the number of users identified by each of the three methods in the iterative process. We note that a share of users could only be geocoded at the country level, for instance those classified using email-suffix data or giving only coarse geographic information in their location fields (i.e. "Austria"). Specifically, we could infer subnational locations for 415,783 users. Code to replicate our data mapping pipeline is available at: `https://github.com/n1tecki/Geography-of-Open-Source-Software`.

Data Availability

A primary goal of this work is to make geographic data on OSS developers accessible to researchers. We have uploaded both national and regional datasets to GitHub, available at: `https://github.com/johanneswachs/OSS_Geography_Data`. In order to protect user privacy, we only share geographically aggregated counts of active users. For example, one file includes the number of active developers located in each of the 50 US States in 2021. In particular we

[2] `https://github.com/Hipo/university-domains-list`

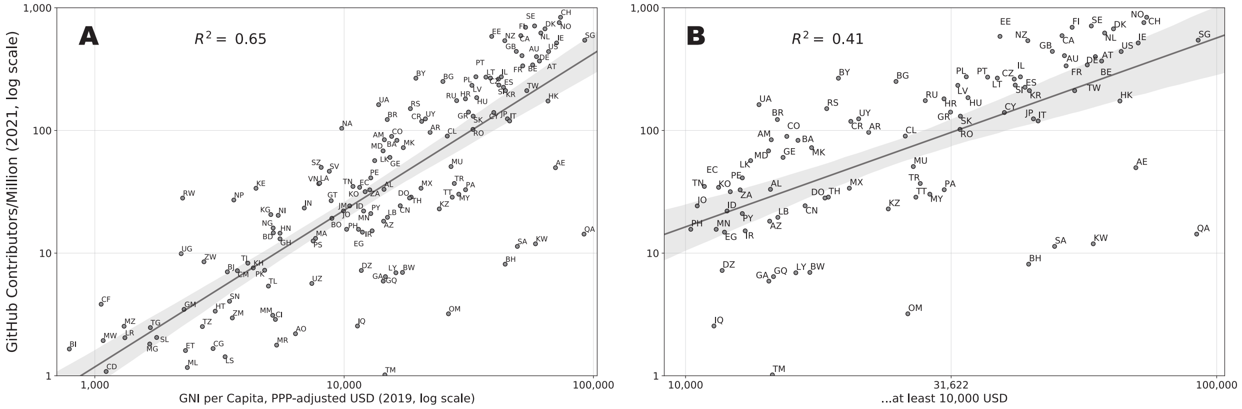

Figure 1: Country economic development (PPP-adjusted GNI per capita) and OSS contributors per capita, log-log scale. A) We observe a strong relationship between economic development and OSS activity, though some countries deviate significantly from the trend. B) Zooming in on wealthier countries only, the relationship weakens significantly, suggesting other factors play an important role in OSS activity. For sake of visualization, we exclude countries with fewer than 1 million inhabitants or less than 1 contributor per million inhabitants.

share data in CSV files comarping counts of active OSS contributors across Countries, European NUTS2 regions, and sub-national units (states/provinces) of the US, Japan, China, India, Russia, and Brazil.

Results

We begin by reporting summary statistics on the number of developers we located in various countries and regions. We highlight which countries host the most OSS developers both in raw terms and per capita. We compare our results with those from previous works published in 2008 and 2010, studying the geographic distribution of developers on the Sourceforge and GitHub platforms, respectively. We also present evidence that the count of active OSS developers correlates strongly with a variety of measures of development and quality of living indicators, above and beyond economic development. We replicate these findings at the regional scale using data from the European NUTS2 regions. We then present an analysis of the geographic concentration of OSS developers within countries, finding that they are in general highly concentrated.

International Comparison

In Table 1 we report the top thirty countries, ranked by overall share of active OSS developers, and compare data from 2021 with data from previous work. At first glance, our results indicate significant changes in the global distribution of OSS activity since 2010 [26, 64]. While North American and Western European countries are still leading locations for OSS, Asian, Eastern European, and Latin American countries are catching up. This finding aligns with recent work mapping national activity in AI and deep learning research, which highlights China's growth in ICT activity [39].

A more appropriate ranking of national activity in OSS would take into account both the population of each country and its level of economic development. We report the top fifty countries by OSS contributors per 100k inhabitants in the appendix (see Table 5). In Figure 1 we present the relationship between income per capita, sourced from the World Bank [66], and the number of active OSS developers per million inhabitants, both on logarithmic scales. We exclude countries with a population of less than one million people (for the sake of visualization). The regression fit explains roughly two-thirds of the variance among all countries, but only 40% for countries with an income per capita of at least $10,000. Countries above the regression line have more OSS developers per capita than expected for their level of economic development, while those below have less. Ukraine, Belarus, Namibia, Brazil, Bulgaria, and Estonia have more OSS activity than expected, while oil-rich states like Qatar, Kuwait, and Saudi Arabia are OSS laggards.

	Sourceforge 2008		GitHub 2010		**GitHub 2021**		
Rank	Country	Share	Country	Share	Country	Share	Rank Change vs. 2008
1	United States	36.1	United States	38.7	United States	24.6	-
2	Germany	8.1	UK	7.7	China	5.8	↑ 4
3	UK	5.1	Germany	6.2	Germany	5.6	↓ 1
4	Canada	4.2	Canada	4.3	India	5.4	↑↑ 7
5	France	3.8	Japan	3.9	UK	5.0	↓ 2
6	China	3.1	Brazil	3.6	Brazil	4.4	↑↑ 6
7	Australia	2.7	France	3.2	Russia	4.3	↑↑ 6
8	Italy	2.6	Australia	3.1	France	3.8	↓ 3
9	Netherlands	2.5	Russia	2.3	Canada	3.8	↓↓ 5
10	Sweden	2.0	Sweden	2.2	Japan	2.7	↑↑ 5
11	India	1.9			South Korea	1.9	↑↑↑ 14
12	Brazil	1.8			Netherlands	1.8	↓ 3
13	Russia	1.6			Spain	1.8	↑ 1
14	Spain	1.6			Poland	1.8	↑ 2
15	Japan	1.3			Australia	1.8	↓↓ 8
16	Poland	1.2			Sweden	1.2	↓↓ 6
17	Belgium	1.2			Italy	1.2	↓↓↓ 9
18	Switzerland	1.0			Ukraine	1.2	New
19	Austria	0.8			Switzerland	1.2	↓ 1
20	Denmark	0.8			Indonesia	1.0	New
21	Singapore	0.8			Taiwan	0.8	↑↑↑ 9
22	Finland	0.8			Colombia	0.8	New
23	Norway	0.7			Argentina	0.7	↑ 4
24	Mexico	0.7			Mexico	0.7	-
25	South Korea	0.7			Norway	0.7	↓ 2
26	Israel	0.6			Belgium	0.7	↓↓↓ 9
27	Argentina	0.6			Denmark	0.7	↓↓ 7
28	Hungary	0.6			Finland	0.6	↓↓ 6
29	Romania	0.5			Vietnam	0.6	New
30	Taiwan	0.5			Austria	0.6	↓↓↓ 11

Table 1: Country shares of active OSS contributors on GitHub in 2021. We include the top 30 countries and compare shares from previous work using Sourceforge (2008) [26] and GitHub (2010, only top 10 available) [64]. Across countries the distribution has become more uniform. South and East Asian and Latin American countries have seen the greatest relative increase in share of global OSS contributors.

What explains these residuals? The ability and decision to contribute to OSS projects is likely a complex and multifaceted process [24], but we can compare the relative importance of various environmental factors in a regression framework. Beyond the broad economic development of a country, measured by income per capita, internet penetration [65] likely plays an important role. The UN's Human Development Index (HDI) offers a broad measure of social development, including access to education and health services - likely important upstream factors facilitating OSS contributions [32].

One distinguishing aspect of OSS, compared to many other kinds of knowledge-intense activities, is that OSS outputs are essentially public goods - goods that can be used by anyone. Though a thorough review of the economics of OSS [42, 61] and individual motivations for contributing [7] is beyond the scope of this article, we do expect that OSS activity will be higher where people are more inclined to contribute to public goods. For instance, people living in areas with high levels of generalized trust, that is to say where individuals are more likely to report that in general "most people can be trusted", are known to be more likely to contribute to public goods [59]. We use data from the most recent wave of the World Values Survey to measure this concept of generalized trust [31]. Another feature of countries that strongly correlates with individual propensity to contribute to public goods is quality of government. We therefore also relate OSS outcomes to the Index of Public Integrity (IPI), a measure of the quality of public institutions in a country [47].

Country Feature	Spearman ρ	p-Value	Observations
PPP-adjusted GNI per capita (2019)	0.79	< 0.01	176
Human Development Index (HDI, 2019)	0.86	< 0.01	178
WVS share "most people can be trusted"	0.68	< 0.01	75
Index of Public Integrity (IPI, 2019)	0.87	< 0.01	117
Economic Complexity Index (ECI, 2019)	0.82	< 0.01	175
Deep-learning/AI publications/capita	0.67	< 0.01	183
Internet Penetration (2019)	0.78	< 0.01	180

Table 2: Spearman correlations between country-level social and economic development indicators and active OSS contributors per capita in 2021.

Lastly, we consider the overall economic focus and specialization of a country. As noted before, oil-producing states seem to have fewer OSS contributors relative to their economic development. The extent to which countries specialize in more complex, coordination-intensive industries, and in particular in software-adjacent fields, likely correlates significantly with OSS activity. To capture the former, we consider the Economic Complexity Index (ECI) [34], a measure of the sophistication of a country's export profile. To measure the latter, we use a count the number of academic research preprints on AI/deep-learning published by each country in the last decade [39].

Each of these social, political and economic features has a strong correlation with the local intensity of OSS contributors. We report the Spearman correlation of each variable with OSS contributors per capita in Table 2. Though it is not possible to disentangle the cause and effect relationships between these variables using observational data, we can observe how some of these variables mediate each other, and how they can explain a greater share of inter-country variance in OSS contributors. To do so, we regress the (log-transformed) number of OSS contributors per million inhabitants on a selection of these variables, reporting the results of several alternative specifications fit using ordinary-least-squares (OLS) in Table 3, reporting robust standard errors.

Our baseline model, including per capita income, internet penetration, and population, explains around two-thirds of variance. Adding HDI, IPI, or ECI as individual features increases the variance explained by over 10%. All three

	Active GitHub Contributors per Million Inhab. (log, 2021)				
	(1)	(2)	(3)	(4)	(5)
PPP GNI per Cap. ('000 USD, 2019)	0.017* (0.009)	-0.007 (0.007)	-0.002 (0.009)	0.004 (0.008)	-0.010* (0.006)
Internet Penetration (% of Pop., 2019)	0.043*** (0.005)	0.002 (0.008)	0.029*** (0.005)	0.026*** (0.006)	-0.003 (0.009)
Population (log, 2019)	-0.145*** (0.052)	-0.071 (0.044)	-0.016 (0.046)	-0.143** (0.056)	-0.038 (0.057)
Human Development Index (2019)		11.709*** (1.528)			9.327*** (1.553)
Index of Public Integrity (2019)			0.704*** (0.127)		
Economic Complexity Index (2019)				0.962*** (0.184)	0.683*** (0.153)
Observations	174	173	115	150	149
Adjusted R^2	0.631	0.747	0.819	0.740	0.804
Residual Std. Error	1.191	0.986	0.788	1.016	0.882
F Statistic	81.3***	175.2***	195.8***	142.4***	155.6***

*p<0.1; **p<0.05; ***p<0.01

Table 3: Regression models (1-5) relating country-level counts of GitHub contributors per million inhabitants (log-transformed) and socio-economic indicators. While income and internet penetration alone account for nearly two-thirds of variance in OSS activity (1), human development (2), quality of political institutions (3), and economic complexity (4) significantly improve model fit above and beyond that baseline. A combined model (5) explains over 80% of variance. We report robust standard errors.

variables have a significant positive relationship with great OSS activity in a country. A model combining the baseline model with HDI and ECI accounts for over 80% of variance in OSS activity. These models suggest that OSS activity is not merely the by-product of an advanced economy, but also depends to a significant degree on social, educational, and political institutions and the degree of technological sophistication in a country. As we will see in the following section, geographic variance in OSS activity becomes more difficult to explain with macro indicators at finer spatial scales.

Regional Variation

Comparing OSS activity between countries indicates that it has spread internationally to a significant extent in the last ten years. However, we know little about the distribution of activity within countries. As mentioned above, while most prior work has focused on international comparisons, Takhteyev and Hilts [64] reported data on local clusters in their work from 2010. They estimated that 7.4% of *global* contributors were in the San Francisco Bay Area, suggesting an immense local concentration of OSS activity. In this section we explore the local distribution of OSS developers in various countries. We focus first on European NUTS2 regions. We again relate socio-economic features to OSS activity, replicating our international findings at the regional scale. We then consider the *concentration* of developers within multiple countries including the EU, US, China, India, Japan, and Brazil. We find that OSS activity is significantly concentrated relative to the distribution of the general population in all countries we examine, though with significant heterogeneity. Repeating the calculation using university educated workers or workers employed in high-tech. fields instead of OSS contributors, we find far lower levels of concentration.

European Regions

We first zoom in on European NUTS2 regions. These regions are especially useful because we can compare regions from multiple countries with generally consistent statistics, sourced from Eurostat. In particular we use the regions defined in 2016. In Figure 2 we map the number of OSS developers per 100,000 inhabitants by European NUTS2 region using Jenks-Caspall bins. We note that the five London NUTS2 regions are merged into a single unit because developers tended to refer to their location as London rather than “Inner London”. We can observe several patterns. First we note that major hubs such as London, Amsterdam, Berlin, Prague, Zurich, Hamburg, Helsinki, Oslo and Stockholm tend to have significant OSS presence. Second, we can see significant variation between regions within countries. Some countries have regions in both the lowest and highest valued bins. In other countries, such as Italy, Spain, and France, the distribution seems to be more uniform. We will return to an investigation of the within-country spatial concentration of developers later.

Knowing the locations of OSS developers at the regional levels, we can attempt to replicate our earlier findings about the relationship between socio-economic development indicators and OSS activity. Again the goal is to show that OSS activity is related to more than just economic development outcomes, albeit this time at a sub-national scale. On top of a baseline set of features including internet penetration, GDP per capita, population and density, we consider the relationships between various indicators of social and technological development and OSS activity. In particular we consider general levels of social trust, measured by the European Values Survey [25], R&D Spending per capita, share of workers employed in high-tech industries, number of patents per 100k inhabitants (sourced from the OECD REGPAT database [43]), and share of the prime working-age population with tertiary education. Unless otherwise noted data are sourced from the 2017 QoG Basic dataset [14]. When data on any feature is only available at the coarser NUTS0 (country) level, we impute from the country level to the regions. Finally we also consider the number of patents filed in the Electrical Engineering (EE) sector in case the relationship between OSS activity and closed-source innovation is heterogeneous across sectors. The WIPO IPC Technology Concordance Tables categorize patents into five fields at its coarsest level, of which EE is most directly related to software[3].

We report the results of an exploratory multiple regression analysis in Table 4. We again fit the models using OLS, with the (log-transformed) number of active GitHub contributors in a region per 100k inhabitants as the dependent

[3] The other categories are: Instruments, Chemistry, Mechanical Engineering, and Other.

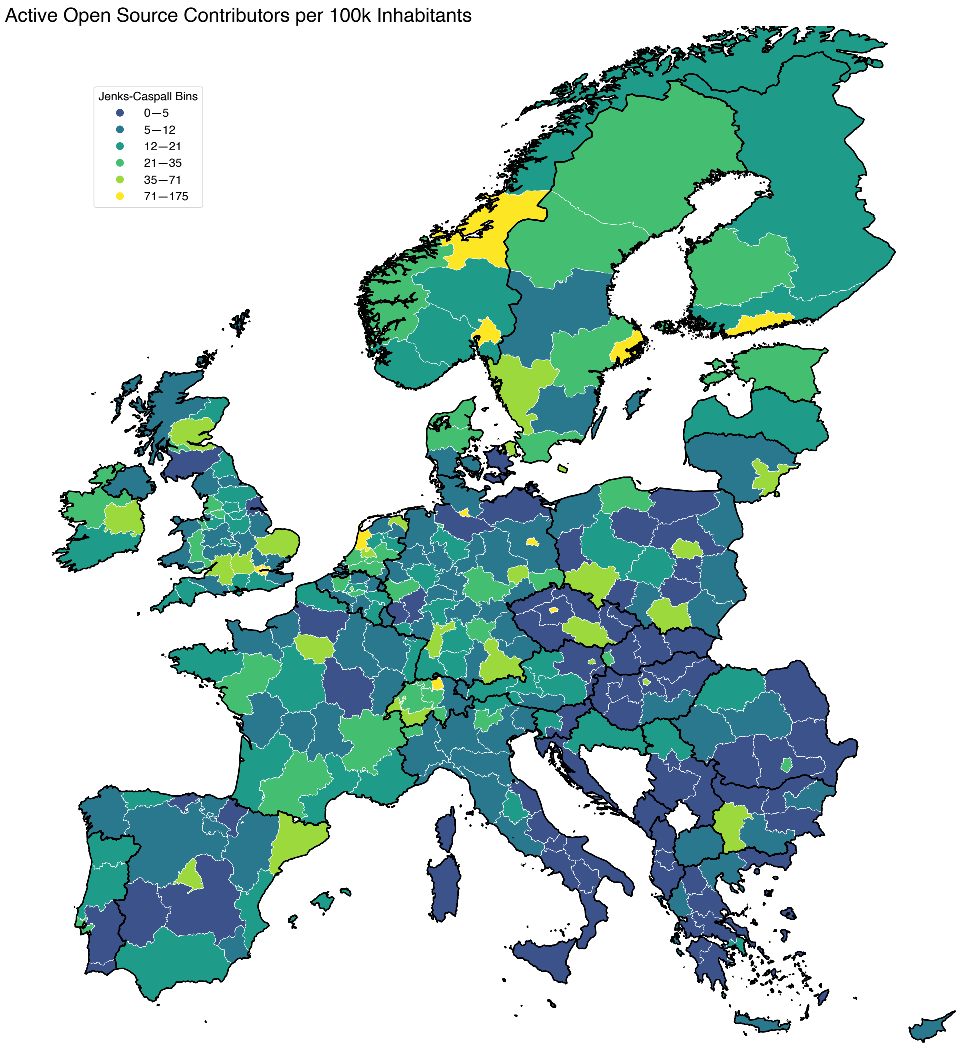

Figure 2: Active OSS developers concentrations in early 2021 per 100,000 inhabitants, NUTS2 regions. We observe significant within country variation.

variable and report robust standard errors. The baseline model again indicates that economic development and internet access have a significant positive relationship with OSS activity in regions. However, at this spatial scale the model has a significantly less accurate fit (Adj. $R^2 \approx .39$) than a similar model predicting OSS activity at the country level. This observation applies also to the feature-rich models. Though generalized trust, R&D spending per capita, share of employment in high-tech industries, share of population with tertiary education, and EE patents are significant predictors of greater OSS activity, the overall model fit does not improve significantly. Only when we include share of the working age population with tertiary education or share of workers employed in high-tech sectors does the adjusted R^2 exceed 50%. While factors like the presence of technologically advanced industries and an educated workforce clearly relate to OSS activity, it seems that at the regional level, more idiosyncratic forces determine local participation in OSS.

The relationships between the two patenting variables and OSS activity also merit comment. Patents are awarded to protect intellectual property and to block the uncompensated use of a creator's ideas. In a naive sense, patenting would appear to be a substitute for open source activity. In practice, however, we see that patenting in EE, and to lesser extent all fields, has a significant positive relationship with OSS activity in regions. If OSS activity were to crowd-out

	Active GitHub Contributors/100k Inhab. European NUTS2 (log, 2021)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Internet Penetration	0.010**	0.007	0.012**	0.011***	0.011**	0.011**	0.004
—(% of Pop. 2017)	(0.005)	(0.006)	(0.005)	(0.004)	(0.004)	(0.004)	(0.004)
GDP per Cap.	0.754***	0.694***	0.014	0.450***	0.274	0.342	0.426***
—(log Eur, 2017)	(0.147)	(0.167)	(0.262)	(0.152)	(0.181)	(0.211)	(0.155)
Population	0.270***	0.381***	0.486***	0.266***	0.169**	0.216**	0.285***
—(log, 2020)	(0.083)	(0.095)	(0.136)	(0.087)	(0.085)	(0.085)	(0.079)
Population Dens.	0.043*	0.052**	0.048	-0.019	0.064**	0.058*	0.027
—(log, 2017)	(0.026)	(0.026)	(0.029)	(0.023)	(0.029)	(0.030)	(0.023)
EVS Trust		0.415**					
—(2017)		(0.176)					
R&D Spend. per Cap.			0.128**				
—(log, 2017)			(0.052)				
% Empl. High-Tech				0.089***			
—(2019/20)				(0.011)			
Patents Elec-Eng./100k					0.072***		
—(log, 2017)					(0.023)		
Patents/100k						0.050*	
—(log, 2017)						(0.027)	
% with Tertiary Edu.							0.022***
—(2019/20)							(0.002)
Observations	276	198	258	262	258	258	276
Adjusted R^2	0.388	0.428	0.410	0.503	0.406	0.395	0.530
Residual Std. Error	0.371	0.354	0.347	0.328	0.348	0.352	0.325
F Statistic	36.8***	27.1***	34.5***	50.0***	33.0***	31.0***	52.8***

*p<0.1; **p<0.05; ***p<0.01

Table 4: Regression models relating EU NUTS2 counts of GitHub contributors per 100k inhabitants (log-transformed) and socio-economic indicators. We report robust standard errors.

patenting, we would expect to see the opposite relationship. This suggests that OSS plays a complementary role in the innovation process, likely via knowledge spillovers discussed earlier in the paper. Hybrid outcomes are also possible, in which software that accompanies a proprietary product is made open source. This finding also suggests how OSS activity can serve as a proxy of useful skills cognitively close to those involved in closed-source innovation [9].

In regions, OSS activity is positively related with economic development, activity in technology intensive sectors, and the presence of an educated and trusting population. At the same time, models including these factors fail to explain over half of the observed variance in OSS activity between regions in Europe. These findings motivate our analysis of the concentration of OSS activity within countries, which will demonstrate that this variance is also large in size.

Concentration

Many kinds of knowledge-intensive activities are generally known to cluster in specific areas within countries. One of the primary goals of our study is to examine the extent to which this is true of OSS. Though it may be especially conducive to remote collaboration and decentralization, Figure 2 provides qualitative evidence that these aspects of OSS development do not outweigh the tendency of knowledge-intensive activities to cluster. We now introduce a measure to quantify this phenomenon, and compare the degree of geographic concentration of OSS activity in countries.

There are many measures of the dispersion or concentration of people or things across geographic regions [17]. As we are interested in comparing the relative concentration of OSS developers between countries, we need a measure which considers population heterogeneity between regions within countries. For example, in a hypothetical country with two regions, A and B containing 80% and 20% of the country's population, respectively, an 80%-20% distribution of OSS developers between regions A and B should not be interpreted as concentration.

Measures like the Herfindahl Index depend on the number of regions, while the Ellison-Glaeser measure is sensitive to variance in population between regions [17], and Gini-like measures quantify inequality, which is distinct from concentration. The *Adjusted Geographic Concentration* (AGC), developed by the OECD [63], measures concentration that is comparable between countries with different numbers of regions and different distributions of the underlying population between them [60].

Consider a country C with a population P split into N regions. The regions $i \in \{1, 2, \ldots N\}$ have shares of the population p_i. Denoting by m_i the share of OSS developers in a country living in the region C_i, we define the *Geographic Concentration* (GC) of developers in country C as:

$$GC(C) = \sum_{i \in C}^{N} |m_i - p_i| \, .$$

This measure sums the absolute differences in shares between the general population and the subpopulation of interest (in our case, active OSS contributors). This statistic tends to underestimate concentration in regions with a larger share of the population, and the validity of comparisons between countries with different numbers of regions is unclear. To address these issues the GC is usually scaled by its maximum possible value in each country: specifically, by the value it would take if all OSS developers were located in the least populated region of the country. In our previous notation:

$$GC_{max}(C) = (1 - p_{min}) + \sum_{i \in C, p_i \neq min}^{N} p_i = 2(1 - p_{min})$$

Dividing $GC(C)$ by $GC_{max}(C)$, we obtain the *Adjusted Geographic Concentration* (AGC) of a country C:

$$AGC(C) = \frac{GC(C)}{GC_{max}(C)}$$

The AGC varies between 0 and 1: a country in which the population of OSS developers is distributed in precisely the same proportions across region as the population, would have an AGC score of 0. A country in which all of OSS developers live in the region with the smallest population would have an AGC score of 1.

We calculated the AGC score for various countries, including European countries with at least 2 NUTS2 regions, and the US (states + DC), China (provinces and municipalities, excluding Hong Kong and Taiwan), India (states and union territories), Russia (federal subjects) and Brazil (federal states + the Federal District). We report our estimates of developer concentration by country in Figure 3.

We see that in all countries we examine, OSS development is concentrated regionally, relative to the general population. There is however significant variation between countries. For instance we can say that Brazilian, Portuguese and Italian OSS contributors are more evenly distributed amongst regions in those countries, than developers in Czechia, Hungary and Lithuania.

This analysis, however, does not make it clear how concentrated OSS developers are compared to other kinds of knowledge workers. In general regional statistics on such workers are not internationally comparable. However, among the European NUTS countries, we can make this comparison. We therefore recalculated the AGC of each country in this group, substituting the share of workers in high-tech sectors and with tertiary education, respectively, for OSS contributors in the calculation. If OSS contributors are more dispersed in a country than the university educated or high-tech workforce, we would expect the AGC to be higher under these alternative specifications.

In Figure 3 we observe the opposite effect: OSS contributors are significantly *more* concentrated in particular regions than either university educated or high-tech workers. This finding holds with remarkable regularity across the countries we analyze (only Greece is an exception). We provide the full table in the appendix, see Table 6. These estimates of concentration also provide useful perspective for the previous more global analysis. The AGC scores for workers with higher education vary between .02 and .19 (mean: .09, stdev.: .05), for workers in high-tech sectors between .1 and .52 (mean: .23, stdev.: .10), and for OSS contributors in European countries between .22 and .67 (mean: .41, stdev.: .

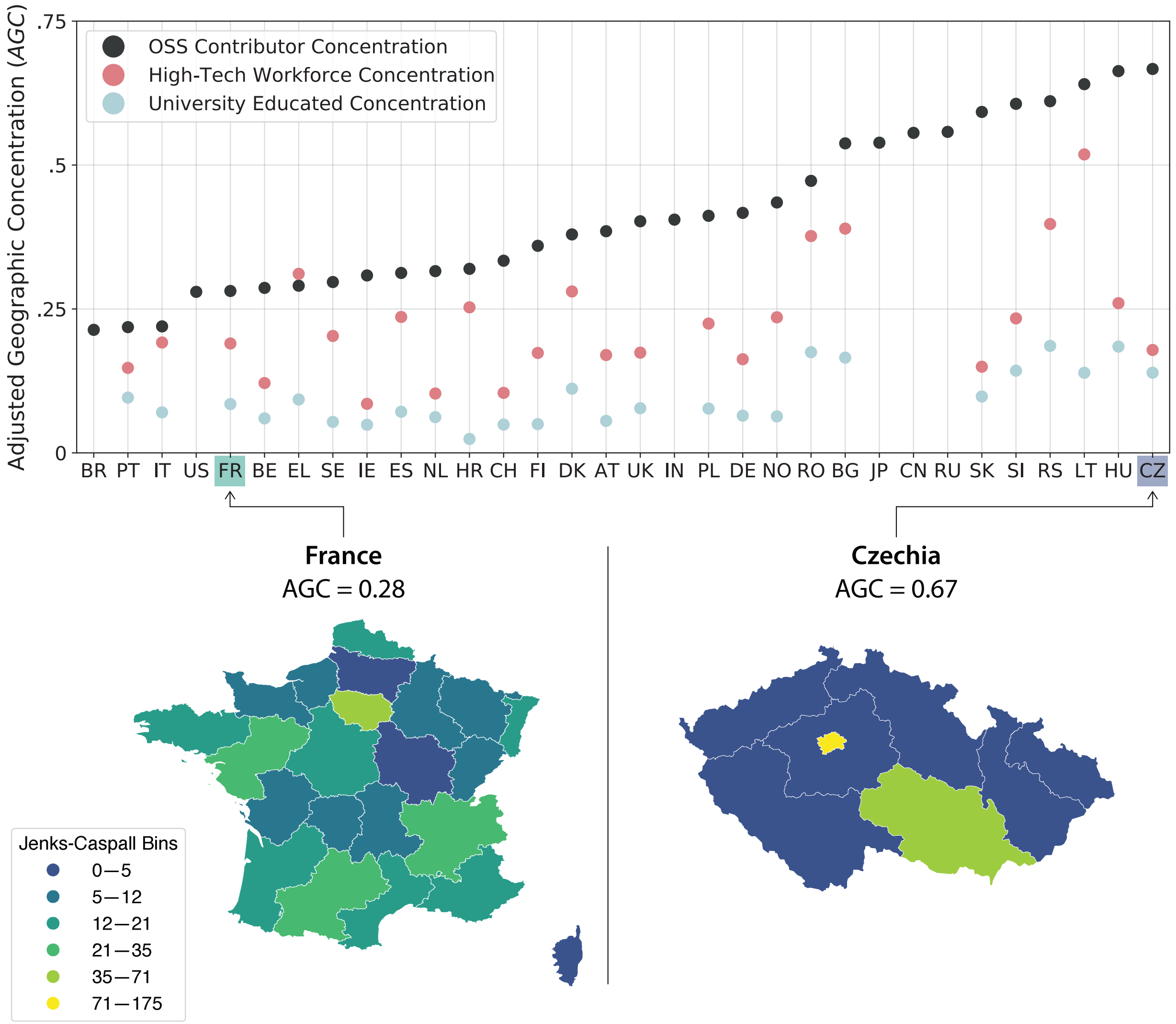

Figure 3: OSS contributor regional concentration within countries, measured using Adjusted Geographic Concentration (AGC). A score of zero indicates that OSS contributors are distributed across a country's regions in proportion to population. AGC equals 1 if all OSS contributors are concentrated in the least populated region of the country. For European countries with NUTS regions we compare the AGC of OSS contributors with the AGC calculated for university educated workers and people employed in high-tech sectors. Below: Examples of countries with relatively low (France) and high (Czechia) OSS AGC scores. Brighter NUTS2 regions indicate more OSS contributors per capita, with bins as in Figure 2.

14). The AGC of OSS developers in all countries in our sample exceeds the average AGC of high-tech workers in European countries. Overall, these results are strong evidence that OSS developers cluster to a significant degree in all countries in our analysis.

Discussion

In this paper we presented an analysis of a novel dataset on the geography of open source software developers. We found that while the overall share of active developers has become more evenly distributed between countries, within-country regional differences remain strong. As a technologically and intellectually demanding area of knowledge creation, the local conditions for strong OSS activity are particular. The social and economic spillovers of local strength in OSS are likely to reinforce local strength in software and technology. This suggests how the data presented in this paper could be a useful proxy for local expertise in software, complementing recent work on, for instance, local exposure to AI and other emerging technologies or innovations [52, 19]. If indeed OSS is a driver of, and not merely a proxy for, innovation outcomes, we need to better understand the role of actors such as universities, research institutes, and governments in promoting OSS activity [62, 49], for example in a mission-oriented context [74] or through procurement policy [70]. Geographic heterogeneities could also be used to better understand diversity deficits in software [44, 56, 2]. In the context of the broader global economy, OSS reflects at least one mainstream trend: activity is increasingly spread around the world, while within-country concentration remains high [45].

Though digitalization facilitates collaborations across distances [22], continued regional clustering suggests that location matters as much as ever. The network effects present in places such as Silicon Valley are strong enough to overcome other obstacles including higher tax rates and cost of living. In software knowledge spillovers and transfers still happen locally and within firms [77]. The developers in our dataset were geocoded in early 2021, roughly one year after the Covid-19 crisis went global. It remains to be seen whether proliferation of remote work and decentralization will be reflected in a change in the geographic distribution of OSS developers. If knowledge workers are going to permanently decamp to smaller cities in the post-Covid era, one would except to see the first signs of this in OSS with its advanced infrastructure for remote collaboration. Our results, early in the post-Covid era, suggest that the winner-take-all dynamics of economic geography persist [21].

Some limitations of our study highlight potential future work and extensions. The most obvious extension is to continue collecting data into the future to observe trends as they unfold, for instance to better understand whether OSS activity follows or predicts entrepreneurial activity and innovation. Given the apparent interest in the future of work vis-á-vis Covid-19 and remote collaboration, this swould be a valuable and important extension. The movement of individual developers could provide additional level of detail into how the tech world is adapting to OSS. Movements likely predict future inter-regional linkages, as those who arrive to a new place connect people in their new homes with those from their previous one [40]. Such data also presents the opportunity to study how software innovations diffuse geographically for instance via collaboration ties [67].

Our approach could be extended to cover alternative platforms for open source contributions including GitLab and Bitbucket, the absence of which may bias our results [69]. Other possibilities for geolocating developers, including by geocoding the companies and organizations they work for, should be explored to add breadth. However, such extensions may bias results as developers in different countries may link to organizations at different rates. Nor does the use of an email address with a country top-level domain guarantee that an individual in question lives in that country.

Another potential extension would be to measure the success or impact of developers and regions, which we hypothesize is still centralized among the core regions of OSS identified over ten years ago [26, 64]. Yet regional clusters of creativity are known to rise and fall [15] over long periods of time. More work is needed to understand how creative clusters of OSS developers can be nurtured and kept healthy [49]. In this context, it is crucial to better understand also exactly how OSS and software more generally relates to other industries and sectors [18, 54].

Acknowledgements

JW, MN, and AP acknowledge support from the City of Vienna WU Jubilee Fund (Jubiläumsfonds der Stadt Wien). The authors thank Frank Neffke, Gergö Tóth, and Balázs Lengyel for helpful discussions and Jeremias Brändle for assistance with data.

Conflict of interest

The authors declare that they have no conflict of interest.

Appendix

	Country	ISO2	# GitHub	# Twitter	# Email Suf.	Total Contributors.	Pop. (mm)	Contribs./100k
1	Iceland	IS	249	10	162	421	0.4	105
2	Switzerland	CH	4978	55	2164	7197	8.6	84
3	Norway	NO	3137	29	846	4012	5.3	76
4	Sweden	SE	6076	38	1209	7323	10.3	71
5	Finland	FI	3086	20	707	3813	5.5	69
6	Denmark	DK	2798	22	1086	3906	5.8	67
7	Netherlands	NL	8843	110	1820	10773	17.3	62
8	Canada	CA	19267	219	2783	22269	37.6	59
9	Estonia	EE	600	6	154	760	1.3	58
10	Luxembourg	LU	280	4	40	324	0.6	54
11	New Zealand	NZ	2299	28	315	2642	4.9	54
12	Singapore	SG	2818	33	251	3102	5.7	54
13	Ireland	IE	2224	25	282	2531	4.9	52
14	United States	US	128526	1831	14014	144371	328.2	44
15	United Kingdom	GB	24493	443	4516	29452	66.8	44
16	Australia	AU	8970	120	1247	10337	25.4	41
17	Germany	DE	25027	276	7909	33212	83.1	40
18	Austria	AT	2472	53	751	3276	8.9	37
19	France	FR	17474	202	4875	22551	67.1	34
20	Belgium	BE	3134	43	758	3935	11.5	34
21	Israel	IL	2131	43	314	2488	9.1	27
22	Belarus	BY	2375	8	149	2532	9.5	27
23	Portugal	PT	2485	32	285	2802	10.3	27
24	Lithuania	LT	683	5	60	748	2.8	27
25	Poland	PL	8864	51	1491	10406	38.0	27
26	Czechia	CZ	2771	34	0	2805	10.7	26
27	Bulgaria	BG	1510	11	234	1755	7.0	25
28	Slovenia	SI	411	6	75	492	2.1	23
29	Latvia	LV	371	4	68	443	1.9	23
30	Spain	ES	9091	157	1345	10593	47.1	22
31	Malta	MT	100	0	12	112	0.5	22
32	Taiwan	TW	4293	66	620	4979	23.6	21
33	South Korea	KR	10025	35	861	10921	51.7	21
34	Hungary	HU	1616	13	184	1813	9.8	18
35	Croatia	HR	666	3	73	742	4.1	18
36	Russia	RU	15543	108	9620	25271	144.4	18
37	Hong Kong	HK	1151	13	139	1303	7.5	17
38	Ukraine	UA	6941	29	234	7204	44.4	16
39	Serbia	RS	953	5	81	1039	6.9	15
40	Cyprus	CY	157	4	7	168	1.2	14
41	Greece	GR	1338	21	151	1510	10.7	14
42	Slovakia	SK	620	6	93	719	5.5	13
43	Japan	JP	12181	277	3248	15706	126.3	12
44	Uruguay	UY	397	11	27	435	3.5	12
45	Brazil	BR	24021	299	1571	25891	211.0	12
46	Costa Rica	CR	501	7	85	593	5.0	12
47	Italy	IT	5728	107	1369	7204	60.3	12
48	Romania	RO	1820	11	148	1979	19.4	10
49	Namibia	NA	250	9	1	260	2.5	10
50	Argentina	AR	3864	50	418	4332	44.9	10

Table 5: Countries ranked by number of GitHub contributors (located via GitHub or Twitter location, or email suffix data), per 100k inhabitants. We exclude countries with fewer than 300k inhabitants and Montenegro, because the ".me" domain suffix is popular world-wide.

Country	$AGC_{TertEdu}$	AGC_{HiTech}	AGC_{OSS}
PT	0.10	0.15	0.22
IT	0.07	0.19	0.22
FR	0.08	0.19	0.28
BE	0.06	0.12	0.29
EL	0.09	0.31	0.29
SE	0.05	0.20	0.30
ES	0.07	0.24	0.31
IE	0.05	0.09	0.31
HR	0.02	0.25	0.32
NL	0.06	0.10	0.32
CH	0.05	0.10	0.33
FI	0.05	0.17	0.36
DK	0.11	0.28	0.38
AT	0.06	0.17	0.39
UK	0.08	0.17	0.40
PL	0.08	0.22	0.41
DE	0.06	0.16	0.42
NO	0.06	0.24	0.44
RO	0.17	0.38	0.47
BG	0.17	0.39	0.54
SK	0.10	0.15	0.59
RS	0.19	0.40	0.61
SI	0.14	0.23	0.61
LT	0.14	0.52	0.64
HU	0.18	0.26	0.66
CZ	0.14	0.18	0.67

Table 6: The concentration of workers with higher education, workers in high-tech industries, and OSS contributors across NUTS2 regions of countries, measured by the Adjusted Geographic Concentration (AGC). The concentration of OSS contributors in particular regions is consistently higher than the concentration of both alternative populations.

References

[1] Aksulu, A., Wade, M.R.: A comprehensive review and synthesis of open source research. Journal of the Association for Information Systems **11**(11), 6 (2010)

[2] Albusays, K., Bjorn, P., Dabbish, L., Ford, D., Murphy-Hill, E., Serebrenik, A., Storey, M.A.: The diversity crisis in software development. IEEE Software **38**(2), 19–25 (2021)

[3] Alhusen, H., Bennat, T., Bizer, K., Cantner, U., Horstmann, E., Kalthaus, M., Proeger, T., Sternberg, R., Töpfer, S.: A new measurement conception for the 'doing-using-interacting' mode of innovation. Res. Pol. **50**(4) (2021)

[4] Andreessen, M.: Why software is eating the world. Wall Street Journal **20**(2011), C2 (2011)

[5] Anthes, G.: Open source software no longer optional. Communications of the ACM **59**(8), 15–17 (2016)

[6] Balland, P.A., Boschma, R.: Mapping the potentials of regions in europe to contribute to new knowledge production in industry 4.0 technologies. Regional Studies pp. 1–15 (2021)

[7] Bessen, J.: Open source software: Free provision of complex public goods. In: The economics of open source software development, pp. 57–81. Elsevier (2006)

[8] Bettencourt, L.A., Ostrom, A.L., Brown, S.W., Roundtree, R.I.: Client co-production in knowledge-intensive business services. California management review **44**(4), 100–128 (2002)

[9] Boschma, R.: Proximity and innovation: a critical assessment. Regional studies **39**(1), 61–74 (2005)

[10] Branstetter, L.G., Drev, M., Kwon, N.: Get with the program: Software-driven innovation in traditional manufacturing. Management Science **65**(2), 541–558 (2019)

[11] Cunningham, S.W., Werker, C.: Proximity and collaboration in european nanotechnology. Papers in Regional Science **91**(4), 723–742 (2012)

[12] Daffara, C.: Estimating the economic contribution of open source software to the European economy. In: The First Openforum Academy Conference Proceedings, pp. 11–14 (2012)

[13] Dahlander, L., Gann, D.M., Wallin, M.W.: How open is innovation? a retrospective and ideas forward. Research Policy **50**(4), 104218 (2021)

[14] Dahlberg, S., Sundström, A., Holmberg, S., Rothstein, B., Alvarado Pachon, N., Dalli, C.M.: The Quality of Government Basic Dataset. University of Gothenburg: The Quality of Government Institute (2021)

[15] Doehne, M., Rost, K.: Long waves in the geography of innovation: The rise and decline of regional clusters of creativity over time. Research Policy **50**(9), 104298 (2021)

[16] Eghbal, N.: Working in public: the making and maintenance of open source software. Stripe Press (2020)

[17] Ellison, G., Glaeser, E.L.: Geographic concentration in us manufacturing industries: a dartboard approach. Journal of political economy **105**(5), 889–927 (1997)

[18] Essletzbichler, J.: Relatedness, industrial branching and technological cohesion in us metropolitan areas. Regional Studies **49**(5), 752–766 (2015)

[19] Felten, E., Raj, M., Seamans, R.: Occupational, industry, and geographic exposure to artificial intelligence: A novel dataset and its potential uses. Strategic Management Journal (2021)

[20] Fershtman, C., Gandal, N.: Direct and indirect knowledge spillovers: the "social network" of open-source projects. The RAND Journal of Economics **42**(1), 70–91 (2011)

[21] Florida, R., Rodríguez-Pose, A., Storper, M.: Cities in a post-covid world. Urban Studies (2021)

[22] Forman, C., van Zeebroeck, N.: Digital technology adoption and knowledge flows within firms: Can the internet overcome geographic and technological distance? Research Policy **48**(8), 103697 (2019)

[23] Fry, T., Dey, T., Karnauch, A., Mockus, A.: A dataset and an approach for identity resolution of 38 million author ids extracted from 2b git commits. In: Proc. of the 17th Int. Conf. on Mining Software Repositories (2020)

[24] Gerosa, M., Wiese, I., Trinkenreich, B., Link, G., Robles, G., Treude, C., Steinmacher, I., Sarma, A.: The shifting sands of motivation: Revisiting what drives contributors in open source. In: 2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE) (2021)

[25] GESIS Data Archive: European values study 2017: Integrated dataset (2017). DOI 10.4232/1.13560

[26] Gonzalez-Barahona, J.M., Robles, G., Andradas-Izquierdo, R., Ghosh, R.A.: Geographic origin of libre software developers. Information Economics and Policy **20**(4), 356–363 (2008)

[27] Goth, G.: Open source business models: ready for prime time. IEEE Software **22**(6), 98–100 (2005)

[28] Gousios, G., Spinellis, D.: Ghtorrent: Github's data from a firehose. In: 2012 9th IEEE Working Conference on Mining Software Repositories (MSR), pp. 12–21. IEEE (2012)

[29] Greenstein, S., Nagle, F.: Digital dark matter and the economic contribution of Apache. Res. Pol. **43**(4) (2014)

[30] Grier, D.A.: The tyranny of geography. IEEE Annals of the History of Computing **48**(02), 100–100 (2015)

[31] Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J.: Round seven-country-pooled datafile. World Values Survey (2017)

[32] HDI: Human Development Index. United Nation Development Program (2019)

[33] Hecht, B., Hong, L., Suh, B., Chi, E.H.: Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In: Proc. of the SIGCHI conference on Human Factors in Computing Systems, pp. 237–246 (2011)

[34] Hidalgo, C.A.: Economic complexity theory and applications. Nature Reviews Physics pp. 1–22 (2021)

[35] Hu, T., Li, J., Zhou, H., Li, C., Holmes, E.C., Shi, W.: Bioinformatics resources for sars-cov-2 discovery and surveillance. Briefings in bioinformatics (2021)

[36] Jensen, M.B., Johnson, B., Lorenz, E., Lundvall, B.Å., Lundvall, B.: Forms of knowledge and modes of innovation. The learning economy and the economics of hope **155** (2007)

[37] Juhász, S., Lengyel, B.: Creation and persistence of ties in cluster knowledge networks. Journal of Economic Geography **18**(6), 1203–1226 (2018)

[38] Kaminski, J., Hopp, C., Tykvová, T.: New technology assessment in entrepreneurial financing–does crowdfunding predict venture capital investments? Technological Forecasting and Social Change **139**, 287–302 (2019)

[39] Klinger, J., Mateos-Garcia, J., Stathoulopoulos, K.: Deep learning, deep change? mapping the evolution and geography of a general purpose technology. Scientometrics pp. 1–33 (2021)

[40] Kunczer, V., Lindner, T., Puck, J.: Benefitting from immigration: The value of immigrants' country knowledge for firm internationalization. Journal of International Business Policy **2**(4), 356–375 (2019)

[41] Lakhani, K.R., Von Hippel, E.: How open source software works:"free" user-to-user assistance. In: Produktentwicklung mit virtuellen Communities, pp. 303–339. Springer (2004)

[42] Lerner, J., Tirole, J.: Some simple economics of open source. The Journal of Industrial Economics **50**(2) (2002)

[43] Maraut, S., Dernis, H., Webb, C., Spiezia, V., Guellec, D.: The oecd regpat database: a presentation (2008)

[44] May, A., Wachs, J., Hannák, A.: Gender differences in participation and reward on stack overflow. Empirical Software Engineering **24**(4), 1997–2019 (2019)

[45] Milanovic, B.: Global inequality and the global inequality extraction ratio: the story of the past two centuries. World Bank Policy Research Working Paper (5044) (2009)

[46] Montandon, J.E., Silva, L.L., Valente, M.T.: Identifying experts in software libraries and frameworks among GitHub users. In: 2019 IEEE/ACM 16th Int. Conf. on Mining Software Repositories (MSR) (2019)

[47] Mungiu-Pippidi, A., Dadašov, R.: Measuring control of corruption by a new index of public integrity. European Journal on Criminal Policy and Research **22**(3), 415–438 (2016)

[48] Nagle, F.: Learning by contributing: Gaining competitive advantage through contribution to crowdsourced public goods. Organization Science **29**(4), 569–587 (2018)

[49] Nagle, F.: Government technology policy, social value, and national competitiveness. Harvard Business School Strategy Unit Working Paper (19-103) (2019)

[50] Nagle, F.: Open source software and firm productivity. Management Science **65**(3), 1191–1215 (2019)

[51] Neffke, F.M.: The value of complementary co-workers. Science advances **5**(12), eaax3370 (2019)

[52] Ozgun, B., Broekel, T.: The geography of innovation and technology news- An empirical study of the German news media. Technological Forecasting and Social Change **167**, 120692 (2021)

[53] Pietri, A., Spinellis, D., Zacchiroli, S.: The Software Heritage graph dataset: public software development under one roof. In: 2019 IEEE/ACM 16th Int. Conf. on Mining Software Repositories, pp. 138–142. IEEE (2019)

[54] Pintar, N., Scherngell, T.: The complex nature of regional knowledge production: Evidence on european regions. Res. Pol. (2021)

[55] Ponds, R., Van Oort, F., Frenken, K.: The geographical and institutional proximity of research collaboration. Papers in regional science **86**(3), 423–443 (2007)

[56] Prana, G.A.A., Ford, D., Rastogi, A., Lo, D., Purandare, R., Nagappan, N.: Including everyone, everywhere: Understanding opportunities and challenges of geographic gender-inclusion in oss. IEEE Trans. Soft. Eng. (2021)

[57] Raymond, E.: The cathedral and the bazaar. Knowledge, Technology & Policy **12**(3), 23–49 (1999)

[58] Riedl, C., Seidel, V.P.: Learning from mixed signals in online innovation communities. Organization Science **29**(6), 1010–1032 (2018)

[59] Rothstein, B., Uslaner, E.M.: All for all: Equality, corruption, and social trust. World Pol. **58**, 41 (2005)

[60] Rovolis, A., Tragaki, A.: Ethnic characteristics and geographical distribution of immigrants in greece. European Urban and Regional Studies **13**(2), 99–111 (2006)

[61] Sahay, S.: Free and open source software as global public goods? what are the distortions and how do we address them? The Electronic Journal of Information Systems in Developing Countries **85**(4), e12080 (2019)

[62] Secundo, G., Perez, S.E., Martinaitis, Ž., Leitner, K.H.: An intellectual capital framework to measure universities' third mission activities. Technological Forecasting and Social Change **123**, 229–239 (2017)

[63] Spiezia, V.: Measuring regional economies. Statistics Directorate of the OECD (2003)

[64] Takhteyev, Y., Hilts, A.: Investigating the geography of open source software through github. Manuscript (2010)

[65] The International Telecommunication Union: Internet penetration (2019)

[66] The World Bank: GNI per Capita, World Development Indicators (2019)

[67] Tóth, G., Juhász, S., Elekes, Z., Lengyel, B.: Repeated collaboration of inventors across european regions. European Planning Studies pp. 1–21 (2021)

[68] Trippl, M., Tödtling, F., Lengauer, L.: Knowledge sourcing beyond buzz and pipelines: evidence from the vienna software sector. Economic geography **85**(4), 443–462 (2009)

[69] Trujillo, M.Z., Hébert-Dufresne, L., Bagrow, J.P.: The penumbra of open source: projects outside of centralized platforms are longer maintained, more academic and more collaborative (2021)

[70] Uyarra, E., Zabala-Iturriagagoitia, J.M., Flanagan, K., Magro, E.: Public procurement, innovation and industrial policy: Rationales, roles, capabilities and implementation. Research Policy **49**(1), 103844 (2020)

[71] Valiev, M., Vasilescu, B., Herbsleb, J.: Ecosystem-level determinants of sustained activity in open-source projects: A case study of the pypi ecosystem. In: Proc. of the 2018 26th ACM Joint Meeting ESEC/FSE (2018)

[72] Wachs, J., Vedres, B.: Does crowdfunding really foster innovation? evidence from the board game industry. Technological Forecasting and Social Change **168** (2021). DOI https://doi.org/10.1016/j.techfore.2021.120747

[73] Wagner, C., Strohmaier, M., Olteanu, A., Kiciman, E., Contractor, N., Eliassi-Rad, T.: Measuring algorithmically infused societies. Nature (2021)

[74] Wanzenböck, I., Wesseling, J.H., Frenken, K., Hekkert, M.P., Weber, K.M.: A framework for mission-oriented innovation policy: Alternative pathways through the problem–solution space. Sci. and Pub. Pol. **47**(4) (2020)

[75] Weterings, A., Boschma, R.: The impact of geography on the innovative productivity of software firms in the netherlands. Regional Development in the Knowledge Economy. Routledge, London and New York (2006)

[76] Wright, N., Nagle, F., Greenstein, S.M.: Open source software and global entrepreneurship. Harvard Business School Technology & Operations Mgt. Unit Working Paper (20-139), 20–139 (2020)

[77] Wu, L., Jin, F., Hitt, L.M.: Are all spillovers created equal? a network perspective on information technology labor movements. Management Science **64**(7), 3168–3186 (2018)