# Gravity in Online Collaborations

Thomas Fackler[a], Michael Hofmann[b], Nadzeya Laurentsyeva[c]

January 30, 2022

Abstract

To which extent are geographic and cultural barriers relevant for modern, knowledge-intensive production processes that can be completely virtual? Using microdata from GitHub, the world's largest online platform for software projects, we estimate a gravity model at the city-level and show that distance, country borders, and language matter for virtual collaborations. Collaborations are significantly more likely to originate within the same city. Conditional on distance, the country border and language barriers remain economically important. During the COVID-19 pandemic, the importance of geographic distance decreased. Further project-level analysis suggests that ex-ante distributed teams were relatively less affected by the pandemic than colocated teams.

Keywords: gravity model, open source, remote work, online labour markets, COVID-19

JEL: J01, M54, O30, F14

[a]Contact: fackler@ifo.de, ifo Institute, LMU Munich and CESifo.
[b]Contact: michael.hofmann@econ.lmu.de, LMU Munich.
[c]Contact: nadzeya.laurentsyeva@econ.lmu.de, LMU Munich and CESifo.

## 1 Introduction

Growing importance of immaterial goods and increasing digitization have enabled virtual production processes and have made virtual teamwork possible. Digital technologies do not only reduce communication costs, but also help to create powerful environments with apparently no physical barriers for collaboration and for exchange of knowledge and ideas. Does it mean that traditional obstacles, such as bilateral distance, country borders, language barriers or cultural differences do not matter in the virtual production processes? Did the COVID-19 pandemic still alter online production and organization of work among knowledge workers? What types of teams turned out to be more resilient during the pandemic?

We try to answer these questions by using detailed data from GitHub - the world's largest online platform for software development. Our analysis proceeds in two steps. First, we estimate the gravity model for collaborations on GitHub at the city-pair level. We obtain estimates for the effect of distance, country borders and language on collaborations and benchmark them against those found for trade. We then study whether the role of these barriers changed during the COVID-19 pandemic. Finally, we zoom in on the level of individual projects and analyse how the effect of the pandemic on activity differed between colocated and distributed teams.

Our results show that online collaborations on GitHub are still subject to barriers, which hinder other economic interactions. The role of geographic barriers is weaker than in trade, but statistically and economically significant, despite the fact that both the production process and the output of programmers are immaterial. Distance elasticity for online collaborations is almost two thirds of the size of distance elasticity found for trade flows, while the role of sharing a common language is even stronger than that for trade. The effect of distance between locations is non-linear: an additional kilometer decreases collaboration more when distance is low than when owner and committer are already far apart. This is in line with the idea that different modes of transport are used, such that moving from what may be a commuting distance to one that is usually travelled by plane changes the cost of an additional kilometer.

We further analyse whether the role of geographic and cultural barriers for online collaborations has changed over time. While the role of borders and common language for online collaborations seems to have increased during the pandemic, we note that the effect of distance decreased.

Finally we analyse the effects of the pandemic at the project level. We focus on productivity patterns of preexisting small teams of up to 3 members, which are either all from the same location or distributed across multiple locations. Our baseline results suggest that members of distributed teams made about twice as many contributions than members of colocated teams. To address potential concerns of systematic differences across teams we apply Coarsened Exact Matching and show that our results hold. These findings are consistent with the idea that distributed teams were already used to online collaboration before Covid and thus better prepared for fully remote work during the pandemic.

This project relates to several strands of literature. First, the gravity models are well established in the Economic literature and have helped to identify the determinants of bilateral trade in goods and services or of migration flows between different geographical units. The gravity equation models bilateral interactions between geographic units where economic size and distance effects enter multiplicatively. In particular, the scope of interactions is positively related to partners' size, which could be measured by GDP, income and population size, and negatively related to bilateral distance. Such models have been used as a workhorse for understanding the determinants of bilateral trade flows for over 50 years since being introduced first by Tinbergen (1962); see Head and Mayer (2014) for a recent survey. They have also been widely applied to study the determinants of migration flows, see Beine et al. (2016) and Ramos (2017) for reviews of modelling approaches and Mayda (2010) and Migali et al. (2018) for applications to international migration. By applying the gravity model to an online setting, we can identify the determinants of virtual collaborations and benchmark them against those established in trade or migration literature. We can also benefit from finer geographical data and identify drivers of collaboration at the city, rather than country level.

Second, studying cross-city and cross-country code contributions are not only related to trade, but also to the literature on knowledge flows and knowledge production. Knowledge has been shown to be more localised than what would be expected from agglomeration effects alone (Jaffe et al. 1993). Furthermore, knowledge spillovers to other countries have been shown to take time (Hu and Jaffe 2003; Jaffe and Trajtenberg 1999) and the effect of international localisation has turned out to be more robust over time than within-country localisation (Thompson and Fox-Kean 2005). While a large body of this literature draws on analysing patent data and thus focusing on inventors, we provide new evidence on collaboration and knowledge flows among software engineers.

Third, our work relates to the emerging literature that studies how the COVID-19 pandemic has changed work practices, knowledge flows and collaborations across and within organizations (Barrero et al. 2021; DeFilippis et al. 2020; Gibbs et al. 2021; McDermott and Hansen 2021).

## 2 Context and Data

GitHub is a platform for software development that was launched in 2007 and hosts a collaborative version control system. Projects can be started by individual users and companies. The repositories cover a wide variety of (mostly) software projects, some of which are aimed at other developers and some at a wider audience. GitHub allows users to have private and public repositories for the project's code. Our data contain only the latter. These public repositories are usually licensed under common open source licenses such as the GNU General Public License.

To contribute to projects or create new ones, users have to set up an account and can provide their real name, location (usually city) and additional biographical information. Each project has only one owner. The owner may invite other users to contribute and become project members. Users can also initiate and contribute to a project before being invited (McDonald and Goggins 2013). Users who are not project members can not only report issues but also suggest modifications to the code, which the project members can review and accept to the project.

In public projects, all of these activities can be observed by everyone. This makes collaborative software development a unique setting that gives researchers a detailed and, in terms of code, comprehensive view of workers' interactions. Users' profile pages on GitHub show their contributions to different projects, while project pages reveal which users have contributed. Thanks to the version control system, the development history of a project is recorded down to the addition of each line of code. In addition to tools for software development, GitHub also shares some features of social networks, giving users the ability to get updates about each others activities, as well as watch projects and give ”stars” to the ones they like. Motivations of open source contributors have been the subject of economic research and include paid work at software companies, career concerns (showcasing skills), as well as writing software for one’s own needs or to help others (Belenzon and Schankerman 2008; Hergueux and Jacquemet 2015; Lerner and Tirole 2001, 2005).

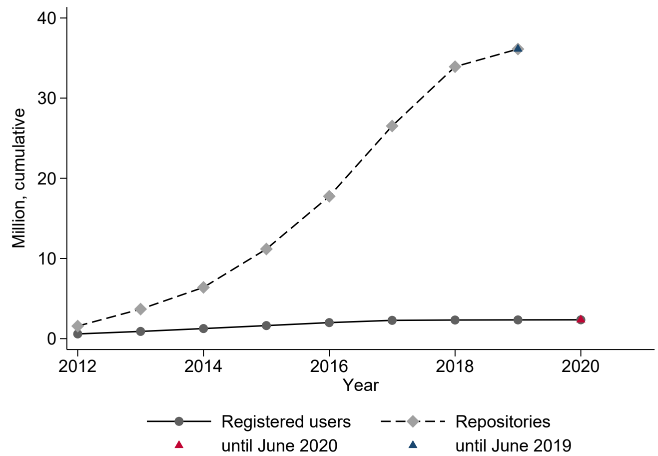
For this study, we mainly look at ”push events”, i.e. submissions of commits to a repository, and here in particular the ones involving project owners and committers from different cities and countries.

We use a snapshot from GitHub Torrents (Gousios 2013) and the GitHub Archive Dataset, as well as a Gravity dataset from CEPII (Version 202102). Both Torrents and Archive datasets provide a mirror of the GitHub public event stream from 2012 on. Both are publicly available in the Google Cloud Platform. We use the two datasets in a complementary way. We take the event stream data from GitHub Archive as it is updated in real time and allows us to incorporate the most up-to-date activity data. We then merge the events to data on users (in particular, their reported geographic locations), which is available in the Torrents dataset. We use two snapshots of GitHub Torrents from June 2019 and June 2020. Our event data spans from 2012 to August 2021, conditional on the involved users (project owners and project committers) being registered on GitHub as of June 2020.

Our final dataset has several features. First, it contains the available information from the public repositories only, as we cannot observe the activity of private projects stored on GitHub. Second, given our research question, we have to limit the data to events where we can identify the location of project owners and project committers. As Figure 1 shows, that leaves us with about 2.4 million registered users and about 36 million repositories.[1] Third, to focus on the collaborative work, we keep only those events on GitHub where a project committer is different from the project owner.

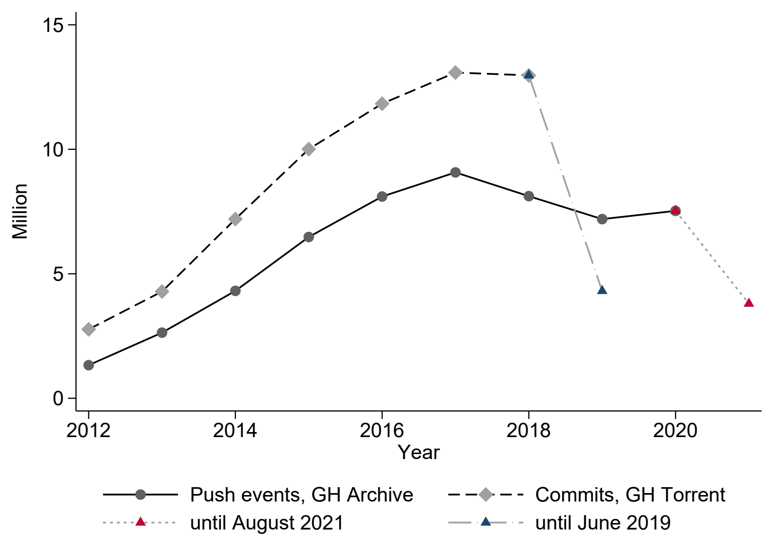
[1] In total, as of June 2019 there were 32 million registered users on GitHub and 125 million repositories.

Figure 1: Cumulative Number of Registered Users and Repositories on GitHub




Note: Only users with reported location; only repositories owned by users with reported location. Source: Data from GitHub Torrent project (snapshots from June 2019 for repositories and from June 2020 for users); publicly available on Google Cloud

Figure 2: Number of Contributions on GitHub




Note: Commits from users with reported location and to repositories owned by users with reported location. Source: Own calculations using GitHub Torrent (June 2019 and June 2020 Snapshots) and GitHub Archive (August 2021)

## 3 Geography of the Activity and Collaborations on GitHub: Descriptive Data

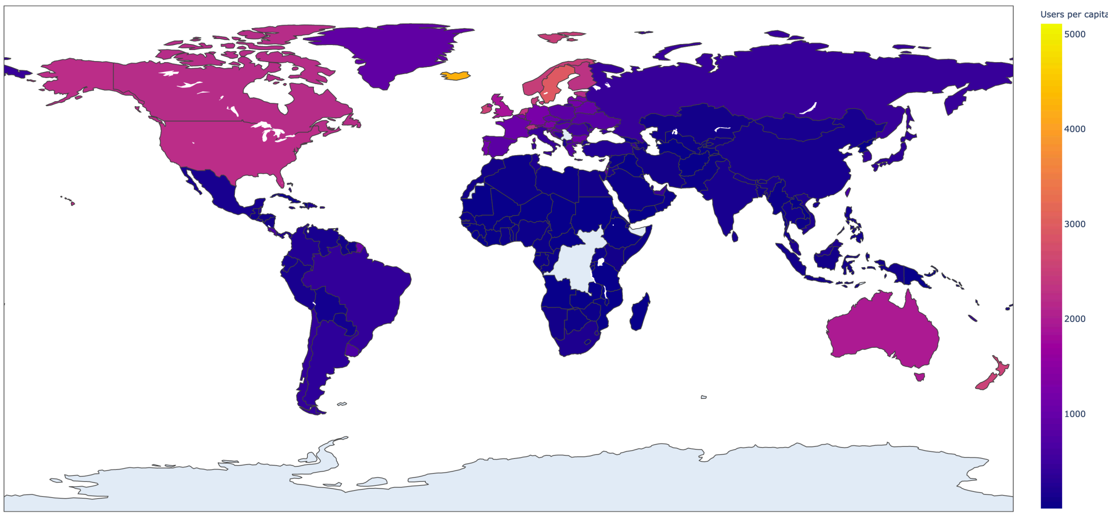
Since its start in 2007, GitHub has become popular with users around the world. Figure 3 shows the number of GitHub users in our data relative to a country's population (in millions). Overall, more advanced countries have a higher share of registered users. It should be noted that even though per capita activity is highest in North America, Europe and Oceania, populous countries such as India and China have sizable user bases on GitHub as well.

Figure 4 depicts the flows of contributions between the eight most active countries on GitHub in our data in terms of international contributions (US, Great Britain, Germany, France, Canada, Netherlands, China, Japan), as well as to and from the set of all other countries. Within-country contributions are excluded. The circle shows how the international contributions between the illustrated countries are divided by committers' countries. The flows go towards the project owner's country.

The largest flow between two countries is from committers in France to projects whose owners are in the US (about 700,000 cross-border events), closely followed by commits from Great Britain (600,000), Germany (600,000) and Canada (500,000) to the US. The next largest flow between countries is from the US (committers) to Great Britain (owner location), which is about two thirds the size of the reverse flow. Among the shown countries, the top three countries by inflow of contributions to projects owned in the country are the US, Great Britain and Germany. The top three by "outflows" are the same countries, but Germany is in second place and Great Britain is third instead.

If the total of "outflows" (contributions to foreign projects) is divided by the total of "inflows" (contributions to local projects by foreigners), Germany has the highest ratio (about 2.4) and the US the lowest (0.6). This is interesting in view of the discussion about Germany's scarcity of technology startups relative to the US, despite the availability of local engineering talent. It is also in line with the political debate about Germany's export strength and American concerns about the trade balance, even though we are analysing numbers that do not enter trade statistics. Japan and China, however, are the other two among the shown countries with a

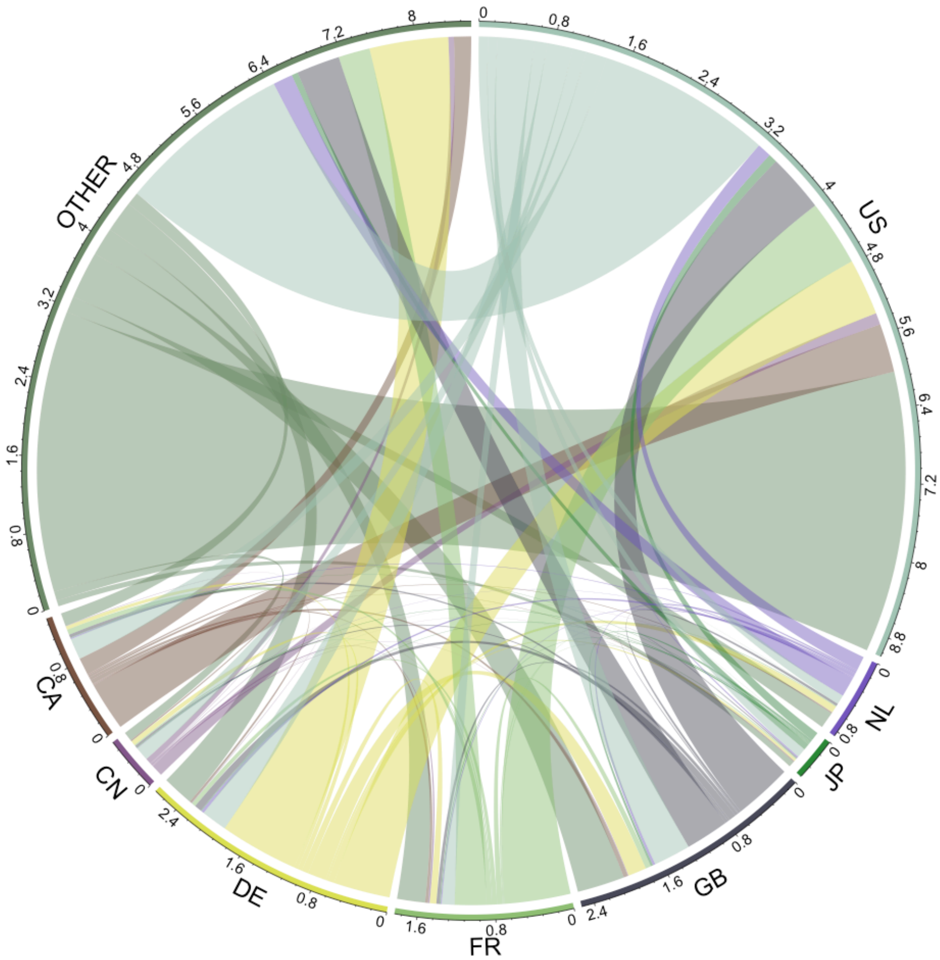
Figure 3: Number of GitHub Users per Capita




Note: This map shows the number of GitHub users per capita (population in millions, i.e. users per one million inhabitants).
Only users with reported location; only repositories owned by users with reported location. Source: Data from GitHub Torrents project (snapshot from June 2019)

ratio smaller than one, despite their export strength.

Figure 4: Bilateral Flows on GitHub




Note: Number of contributions in millions
Only users with reported location; only repositories owned by users with reported location.
Source: Data from GitHub Torrent project (snapshot from June 2019); publicly available on Google Cloud

# 4 Gravity on GitHub

## 4.1 Estimation of the Gravity Equation

We adapt the standard gravity model to estimate the gravity of collaborations on Github at the city-pair level. We aggregate the data at a city-pair and year level. We further restrict our dataset to about 700 of the most active cities on GitHub (as proxied by the number of registered users as of June 2020).[2] These cities together account for 76% of all users and for 82% of all commits by users with reported locations. We construct a strongly balanced annual panel dataset by forming all possible city pairs from our sample for a period between 2012 and 2021, which results in about 5.3 million observations. Only 194,630 of our observation cells, however, are greater than zero.

Our baseline specification is the following:

$$c_{ijt} = \beta_0 D_{ij}^{\beta_1} \{X\}^{\{\gamma\}} \tau_{c_it} \tau_{c_jt} \varepsilon_{ijt}$$

where

$c_{ijt}$ - number of collaborations between a city pair $ij$ in a year $t$. It is measured by the number of contributions by users from a city $i$ and submitted to a project owned by users from a city $j$. In our setting, direction matters: collaborations between city pairs $ij$ and $ji$ are treated as two different observations. To make an analogy to the trade and migration literature, we think of a city of committers as an origin (e.g. origin of service providers - exporters) and a city of the project owner as a destination (e.g. destination of services - importers).

$D_{ij}$ - geographic distance between two cities. We calculate it as the shortest path (in km) between two cities, using their geographic coordinates.

$\{X\}$ - a vector of controls. Economic size (mass analogy in the gravity equation) is proxied by the number of registered GitHub users in an "origin" city (city of a committer) and a "destination" city (city of a project's owner). In addition, we add a dummy for foreign country and a dummy for common language (for cross-border collaborations). Conditional on distance, these dummies capture the effects of state borders and language barriers. In line with the trade literature, we also control for

[2] We set a cutoff of at least 400 registered users per city as of June 2020, resulting in 727 cities. Our results are robust to setting a smaller cutoff.

remoteness of committers and users, which we measure as the user-weighted average distance to other cities.

$\tau_{c_i t}$ and $\tau_{c_j t}$ account for country of committers and country of owners time-specific fixed effects.[3] We conservatively cluster standard errors at the country-pair level.

For most regressions, we apply a poisson pseudo-maximum likelihood regressions with multi-way fixed effects, which is consistent for models with count data in the presence of zero-outcomes.[4] We estimate the effects jointly at the intensive and extensive margins and separately at the intensive and at the extensive margin. For the latter, we use a linear probability model.

## 4.2 Results

Table 1 presents our main results. Columns (1) and (4) estimate the effects jointly at the intensive and extensive margins. Column (2) focuses on the intensive margin, while Column (3) presents the results for the extensive margin. The dependent variable is either a number of commits between a given city pair in a given year or a dummy equal to one if there any commits in a given city pair and year. Distance, economic size variables, and remoteness are measured in natural logarithms.

Our results show that geographic and cultural barriers matter even in virtual environments. The effect of geographic distance on online collaborations is negative and statistically significant with an estimated elasticity of 0.493: if a distance doubles, the number of commits drops by about 50%; conditional, on non-zero collaboration, the number of commits drops by about 40%. On the extensive margin, the linear probability results show that if distance doubles, the probability of having a cross-city collaboration drops by about 2.4 percentage points ($-0.034 * log(2)$). Column (4) uses distance bins instead of a continuous distance measures to capture non-linear distance effects. The reference category corresponds to collaborations within the same city. The results highlight non-linearity in the distance effect and suggest that interactions on GitHub are substantially more likely to happen within the same city, i.e. between people who know each other personally and/or can collaborate in an offline setting. Beyond the distance of 300km, the effect stays at about the same level.

[3] The estimates of the main parameters of interest are robust to including city of committers and city of owners fixed effects.

[4] The estimates are robust to the zero-inflated poisson model.

Our results also show that similar to the trade and migration literature, conditional on distance, the state borders reduce virtual collaborations, while a common language mitigates this negative effect. The effects are economically significant. Conditional on the same geographic distance, a country border reduces the collaboration almost by half (percent drop in commits: $e^{-2.839} - 1 = -94.2\%$), however having a common language cancels out the border effect (percent increase in commits: $e^{0.667} - 1 = 94.8\%$).

Table 2 benchmarks the effect of geographic and cultural barriers on GitHub to their role in shaping modern tradeflows. To that end, we use 2012-19 cross-country export data and aggregate GitHub data at the country-pair level for the same period. Consistent with the trade literature, the estimated distance elasticity for trade equals to -0.8. Trade increases if two countries share a border and a common language. Distance elasticity for GitHub collaborations is almost two thirds compared to that for trade, even though the role of trade costs (e.g. transportation, legal costs, search costs for partners) should be negligible on the platform. Sharing a common language appears to be even more important for online collaboration than for trade, which points to the importance of communication in the production of knowledge-intensive goods, such as software.

Table 1: Gravity Model for Collaborations on GitHub

| VARIABLES | (1)<br>Contributions | (2)<br>Contributions<br>> 0 | (3)<br>Contributions<br>yes/no | (4)<br>Contributions<br>distance dummies |
|---|---|---|---|---|
| Distance | -0.493*** | -0.391*** | -0.034*** | |
| | (0.093) | (0.059) | (0.007) | |
| 1-50km | | | | -0.195 |
| | | | | (0.380) |
| 50-100km | | | | -1.845*** |
| | | | | (0.324) |
| 100-300km | | | | -1.781** |
| | | | | (0.815) |
| 300-700km | | | | -3.151*** |
| | | | | (0.524) |
| >700km | | | | -3.791*** |
| | | | | (0.474) |
| Foreign country | -2.839*** | -1.281*** | -0.053 | -2.978*** |
| | (0.613) | (0.359) | (0.040) | (0.315) |
| Common language | 0.667*** | 0.177 | 0.010* | 0.607*** |
| | (0.174) | (0.147) | (0.006) | (0.162) |
| Users, owner | 0.378*** | 0.212*** | 0.027*** | 0.439*** |
| | (0.041) | (0.056) | (0.005) | (0.047) |
| Users, committer | 1.094*** | 0.825*** | 0.032*** | 1.092*** |
| | (0.068) | (0.049) | (0.005) | (0.074) |
| Remoteness, committer | 2.233 | 2.271 | 0.104*** | 1.459 |
| | (1.866) | (1.953) | (0.031) | (2.039) |
| Remoteness, owner | -1.027 | -1.892 | 0.076*** | -1.588 |
| | (1.071) | (1.189) | (0.025) | (1.225) |
| Observations | 5,347,967 | 194,630 | 5,372,810 | 5,347,967 |
| Clusters | 8464 | 3394 | 8464 | 8464 |
| R-squared | | | 0.109 | |

Notes: The dependent variable in Columns (1), (2), and (4) is the number of contributions between a given city pair. Column (2) limits the sample to city pairs with non-zero contributions. The dependent variable in Column (3) is a dummy equal to one if there are any contributions in a given city pairs. Distance, number of users, and remoteness are in natural logarithms. All specifications include country of committers and country of project owners time-specific fixed effects. Standard errors are clustered at a country pair level. Estimation method: PPML in Columns (1), (2) and (4) and OLS in Column (3).
Source: Own calculations using data from GH Torrent (June 2020) and GitHub Archive (2012-August 2021).

Table 2: Country-level Gravity Regressions: Benchmarking to Trade

| VARIABLES | (1) Tradeflows | (2) Tradeflows > 0 | (3) Contributions | (4) Contributions > 0 |
|---|---|---|---|---|
| Distance | -0.803*** | -0.807*** | -0.488*** | -0.449*** |
| | (0.037) | (0.037) | (0.078) | (0.081) |
| GitHub users, origin | | | 1.381*** | 1.154*** |
| | | | (0.318) | (0.326) |
| GitHub users, destination | | | -0.172 | -0.421 |
| | | | (0.241) | (0.257) |
| GDP, origin | 0.543*** | 0.535*** | -0.363 | -0.322 |
| | (0.077) | (0.078) | (0.320) | (0.323) |
| GDP, destination | 0.259*** | 0.271*** | 0.479* | 0.504* |
| | (0.064) | (0.063) | (0.272) | (0.276) |
| Remoteness, origin | -0.073 | -0.098 | -1.036* | -0.563 |
| | (0.498) | (0.498) | (0.558) | (0.593) |
| Remoteness, destination | -0.912* | -0.928** | -0.073 | 0.066 |
| | (0.466) | (0.466) | (0.724) | (0.755) |
| Contiguity | 0.485*** | 0.485*** | -0.129 | -0.089 |
| | (0.069) | (0.069) | (0.171) | (0.171) |
| Common language | 0.148** | 0.150** | 0.262** | 0.214* |
| | (0.066) | (0.066) | (0.113) | (0.115) |
| Regional trade aggrement | 0.314*** | 0.309*** | -0.228** | -0.263** |
| | (0.054) | (0.054) | (0.107) | (0.110) |
| Observations | 268,590 | 197,779 | 225,208 | 20,057 |
| Clusters | 37419 | 31922 | 30797 | 5283 |

Notes: The dependent variables: tradeflows (exports from origin to destination) in Columns (1)-(2); GitHub contributions (from origin to destination) in Columns (3)-(4) between a given country pair. Columns (2) and (4) limit the sample to pairs with non-zero tradeflows/contributions.
All specifications include year, country of origin, and country of destination fixed effects. Standard errors are clustered at a country pair level. Estimation method: PPML.
Source: Own calculations using data from GH Torrent (June 2020), GitHub Archive (2012-2019) and CEPII Gravity data (2012-2019).

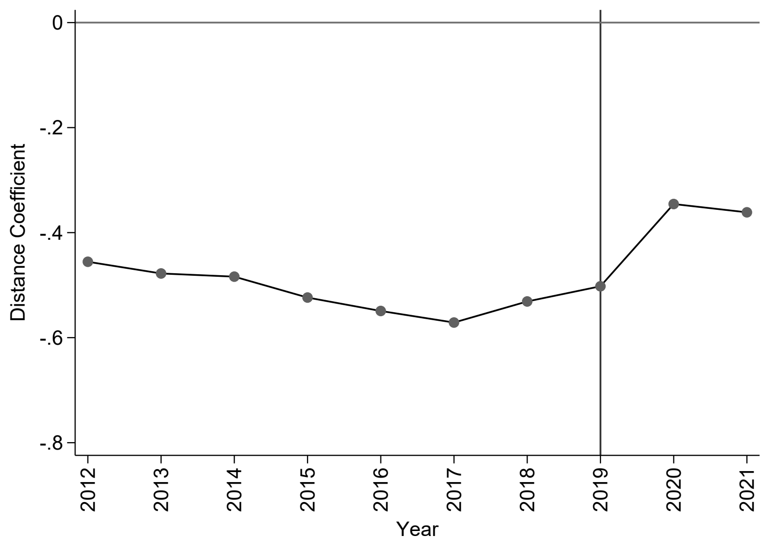
### 4.3 The Effect of COVID-19 Pandemic on Gravity Estimates

Given the emerging evidence that the COVID-19 pandemic has transformed organization of work, we are also interested in tracing whether the role of barriers to online collaboration has changed between 2012 and 2021. In Figure 5 below and in Figures 7 and 6 in the Appendix, we show how the coefficients for distance, common language and country border have changed over time. We estimate the same regression model as in Table 1, Column (1) but separately for each year in our sample. We observe that the role of country borders and common language, if anything, increased with the COVID-19 pandemic. While further analysis is needed to establish the mechanisms, one possible explanation is that it has become harder to start new collaborations abroad with travel bans in place.

Interestingly, we observe an opposite effect for distance. In 2020, distance elasticity changed to -0.3 (while being below -0.4 in 2019). This suggests that the role of geographic distance for online collaborations has weakened during the pandemic. It is likely that the pandemic and associated lockdowns represented a stronger shock for ex-ante colocated teams relative to already distributed ones, because the former had to adjust their work practices under time pressure. It could also be in line with the fact that moving to remote work for a large part of knowledge workers made geographically closer projects lose their comparative advantage in terms of easier off-line communication.

In the next section, we explore the COVID-19 effects on GitHub collaborations in more detail, by using project-level data and comparing performance of colocated vs. distributed teams.

Figure 5: Distance Elasticity of Collaborations on GitHub in 2012-21




Notes: Distance elasticity is calculated by running separate regressions as in Column (1) Table 1 for each year.
Source: Data from GH Torrent project (June 2020) and GitHub Archive publicly available on Google Cloud; authors' own calculations

# 5 The Effect of COVID-19 Pandemic: Team-level Analysis

The analyses in the previous section have shown that international online collaboration patters have changed as a result of the pandemic. This section moves from the aggregate cross-city and cross-country perspective to the level of individual projects. How was the activity of existing projects affected by Corona and how did this effect differ by (co-)location of team members?

## 5.1 Empirical specification

Our first estimation uses a sample of projects started between 2015 and 2018. This analysis thus focuses on the number of contributions to existing projects and not on new projects. Given the size of our sample, using earlier years would not add much power, but might introduce projects that are less comparable, if the type of projects publicly developed on GitHub has shifted over time. Using projects that were started later would limit our ability to observe how active projects performed before Corona.

In an alternative estimation, we restrict the sample further and use only the projects started in 2018. Using multiple years has the advantage that we observe projects being ”treated” by the pandemic shock at different ages, whereas exclusively using projects started in 2018 implies that their age differs only by less than 12 months.

On the one hand, if a shift in GitHub’s user base and project types is a concern, using only one cohort of projects should mitigate this problem. On the other hand, if projects differ more by the month they were started in than by founding year, then the first sample yields more credible results. We show estimations with both samples below and compare the results.

To account for differences between projects with colocated and distributed teams, we match projects to each other. More precisely, we use Coarsened Exact Matching (CEM; cf. Iacus et al. 2012) to find similar projects along multiple dimensions. Our matching approach is somewhat different from the standard CEM applications in that our ”treatment”, the pandemic, affects all projects. We use matching nonetheless because projects were affected differently by the shock depending on whether members were colocated or not. Hence, for teams of three members, we define fully colocated teams as a treatment group, for which we then find two suitable "control groups", one with two distinct locations and another with three, i.e. all three members have stated different cities in their profiles.

Our matching variables are:

- Number of commits in the first year of the projects existence
- Number of projects members in the first two years
- Stars received in the first year
- Year in which the project was started
- Country

For all three samples, we estimate the following equation using a Poisson regression to account for the fact that the number of commits is count data. We cluster at the project level.

$$
\begin{aligned}
c_{it} = {} & \beta_0 + \beta_1 \text{Covid} \\
& + \beta_2 \text{D(Locations=2)} + \beta_3 \text{D(Locations=3)} \\
& + \beta_4 \text{Covid} \times \text{D(Locations=2)} + \beta_5 \text{Covid} \times \text{D(Locations=3)} \\
& + \beta_6 X_{it} + \varepsilon_{it}
\end{aligned}
$$

The dependent variable is the number of commits to project $i$ in month $t$. We account for the difference in activity that affected all projects after the start of the pandemic (coefficient $\beta_1$, where the regressor is a dummy that is equal to 1 starting March 2021 and 0 before), as well as for the differences between colocated projects (reference category) and those with two or three different locations in the team. The coefficients of interest are $\beta_4$ and $\beta_5$, which estimate the differential effect of the pandemic on projects with two and three different locations respectively. Controls include repository age (and quadratic repository age), month fixed effects to capture seasonality in commits, and repository-year $\times$ number-locations fixed effects.

## 5.2 Results

The results for the full sample of projects started between 2015 and 2018 with three members within the first year of existence are presented in Table 3. The regression in Column (1) shows that with the start of the pandemic, activity seems to have dropped overall. The coefficient of interest, however, shows that teams with multiple locations fared significantly better. The coefficients of the interaction effect of Covid and the dummies for the number of locations are both positive (0.599 and 0.937 for two and three distinct locations in the three member teams, respectively). In the regression in Column (2) with additional controls and fixed effects, these coefficients only change slightly. Hence, the results suggest that the distributed projects in the full sample were around twice as productive as teams with all three members in the same location. This is in line with the idea that the less offline collaboration was possible already before Covid, the better they were prepared for fully remote work.

The impact of the pandemic on productivity might be due to differences between colocated and distributed projects. To minimize such differences, we match projects with two and three different locations to colocated projects with three members, based on the dimensions described above using CEM. Qualitatively, the results in Columns (3) and (4), which starts from the same sample as above but adds the matching step, remain the same as in the corresponding Columns (1) and (2) without matching, although the effect size is a little smaller.

In addition to control variables and matching, we can restrict the sample to projects started in 2018 to avoid any influence of shifts in the types of projects started on GitHub over time. Table 4 in the Appendix shows the results of the same regressions with the limited sample containing only projects started in 2018. As described above, this should ensure that projects are more comparable, but may introduce a different type of selection problem if projects started in different months of the year differ systematically, e.g. because of the academic year and term projects started by students.[5] The fact that the results are qualitatively very similar to those estimated with the full sample suggests that selection problems are not a major concern. Quantitatively the results are slightly larger for both teams with two and three different locations. Also, the difference between teams with two and three locations is smaller in this sample.

Table 5 in the Appendix extends the sample to include teams of two as well as three members. "Distributed" teams include all teams with at least two different locations. The results confirm that teams with different locations performed better

[5]More precisely, this would only be a problem for the coefficients of interest if different type of projects reacted differently to the pandemic. Month fixed effects should already control for differences in levels.

during the pandemic. The effect is slightly smaller, but still significant and positive for all specifications. Table 6 estimates the specifications with a sample that contains only projects with two members in the first year. The results are qualitatively similar as well.

Overall, the results consistently show that projects with different locations fared better than colocated teams during the pandemic.

Table 3: Collaborations by team distribution (Full sample)

| | Dependent: N. contributions | | | |
|---|---|---|---|---|
| | Unmatched | | Matched | |
| | (1) | (2) | (3) | (4) |
| Constant | 0.572*** | 2.503*** | 0.565*** | 2.691*** |
| | (0.023) | (0.064) | (0.023) | (0.064) |
| Covid | −2.879*** | 0.067 | −2.916*** | 0.855*** |
| | (0.150) | (0.166) | (0.157) | (0.166) |
| D(Locations=2) | −0.100*** | −0.114 | −0.006 | −0.013 |
| | (0.026) | (0.070) | (0.028) | (0.075) |
| D(Locations=3) | −0.066** | −0.022 | 0.021 | −0.012 |
| | (0.026) | (0.071) | (0.027) | (0.074) |
| Covid × D(Locations=2) | 0.599*** | 0.629*** | 0.458** | 0.517** |
| | (0.174) | (0.178) | (0.215) | (0.220) |
| Covid × D(Locations=3) | 0.937*** | 0.925*** | 0.584*** | 0.661*** |
| | (0.164) | (0.169) | (0.182) | (0.187) |
| Controls | | X | | X |
| Pseudo $R^2$ | 0.073 | 0.341 | 0.087 | 0.400 |
| N | 1,859,152 | 1,859,152 | 1,687,234 | 1,687,234 |
| Clusters | 30,721 | 30,721 | 28,002 | 28,002 |

Notes: The dependent variable is the monthly number of contributions. Included are repositories created between 2015 and 2018 which had 3 members within the first year of existence. Controls include repository age (quadratic), month fixed effects and repository-year × number-locations fixed effects. Columns (1) and (2) show unweighted poisson-regression results and Columns (3) and (4) show weighted poisson-regression results after applying coarsened exact matching. Standard errors are in parentheses and clustered at repository level. ***, **, * indicate significance at 1%, 5% and 10%.

## 6 Conclusion

To sum up, results in Table 1 and Figure 5 highlight that standard barriers found to affect trade and migration flows also matter in a virtual environment. This is particularly interesting given that (monetary) search costs for a relevant project, technology, or a potential partner on GitHub are zero. There are neither the usual "trade" costs, such as tariffs or quotas, nor any travel costs. Moreover in a transparent setting such as GitHub, the information about the quality of a potential project or a contributor is easy to observe for all the actors. This finding is consistent with Singh and Marx (2013) who show that advances in communication technologies and lower costs of travelling hardly reduce the localisation of knowledge over time.

There could be several explanations behind the effect of distance and country borders on GitHub. First, it could be driven by motivation of programmers working on GitHub. If programmers' main motivation to contribute to a certain project is career-driven and if they consider mainly geographically close labour markets, they might focus their activity on local projects. Second, it is likely that personal contact and offline communication between co-workers matter even for online production processes. While GitHub offers infrastructure for virtual collaboration, certain problems (especially those related to the strategic development of a project) require personal interaction. Third, while software products and programming languages are relatively standard, substantial geographic differences in the contents, available technologies, and approaches to work are likely to exist, making projects from different cities and countries non-compatible. From a non-technical perspective, cultural differences could also play a role. For instance, Lyons (2017) use data from an online contract labour market and shows that team organization improves outcomes when workers are from the same country. She argues that the effect is driven by easier communication between team members. Laurentsyeva (2019) uses GitHub data and provides evidence that political conflicts (which are completely exogenous to the functioning of GitHub) increase ingroup-outgroup biases among programmers from the affected countries and decrease their collaboration.

Analysing activity at the project level, we find that teams with members in different locations were less affected by the pandemic than colocated teams. We match projects to mitigate systematic differences in colocated and distributed teams. The results are consistent with better performance for teams already used to fully remote work and online communication. Future work should move from the project to the individual level and analyse the drivers of these effects.

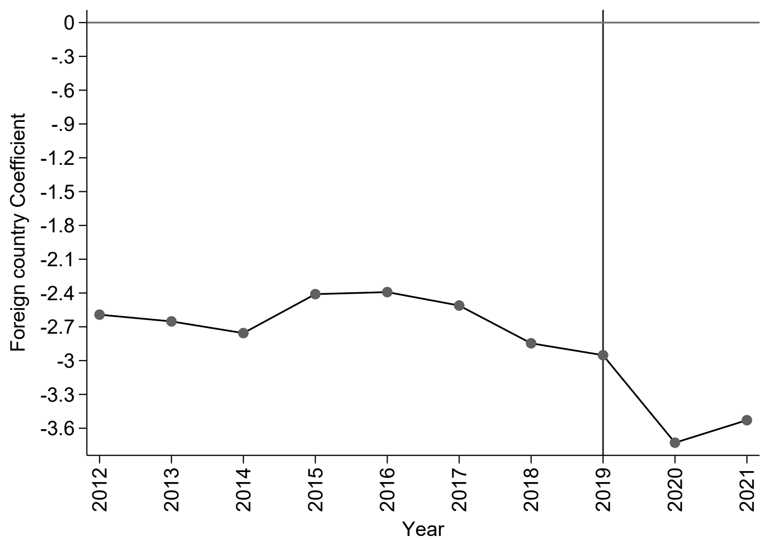
# 7 Appendix

Table 4: Collaborations by team distribution (2018 Repositories)

| | Dependent: N. contributions | | | |
|---|---|---|---|---|
| | Unmatched | | Matched | |
| | (1) | (2) | (3) | (4) |
| Constant | 1.092*** | 2.880*** | 1.082*** | 2.964*** |
| | (0.035) | (0.050) | (0.036) | (0.055) |
| Covid | −3.326*** | 0.880*** | −3.414*** | 1.416*** |
| | (0.170) | (0.186) | (0.183) | (0.206) |
| D(Locations=2) | −0.090** | −0.065 | −0.049 | −0.025 |
| | (0.043) | (0.043) | (0.046) | (0.046) |
| D(Locations=3) | 0.011 | 0.053 | −0.024 | 0.021 |
| | (0.043) | (0.042) | (0.046) | (0.046) |
| Covid × D(Locations=2) | 0.833*** | 0.832*** | 0.760*** | 0.758*** |
| | (0.219) | (0.219) | (0.254) | (0.254) |
| Covid × D(Locations=3) | 1.054*** | 1.053*** | 0.798*** | 0.798*** |
| | (0.194) | (0.195) | (0.234) | (0.236) |
| Controls | | X | | X |
| Adj./Pseudo $R^2$ | 0.135 | 0.396 | 0.147 | 0.441 |
| N | 290,112 | 290,112 | 272,071 | 272,071 |
| Clusters | 6,782 | 6,782 | 6,365 | 6,365 |

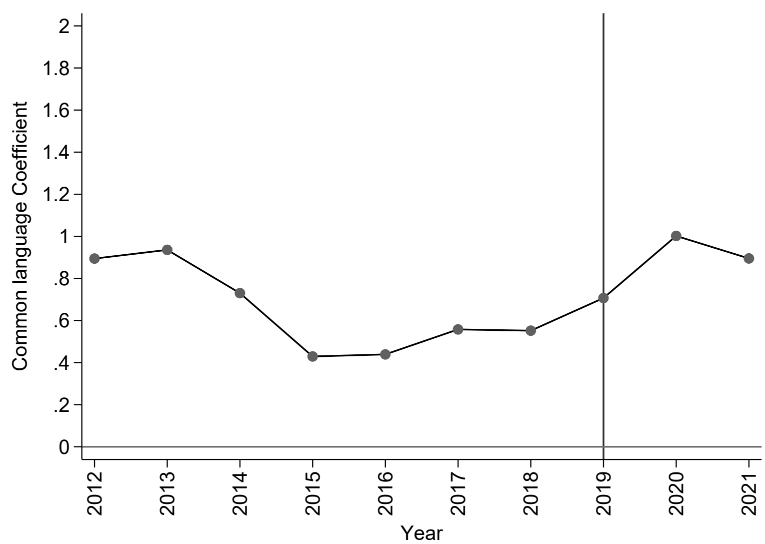
Notes: The dependent variable is the monthly number of contributions. Included are repositories created in 2018 which had 3 members within the first year of existence. Controls include repository age (quadratic), month fixed effects, repository-year fixed effects and repository-year × number-locations fixed effects. Columns (1) and (2) show unweighted poisson-regression results and Columns (3) and (4) show weighted poisson-regression results after applying coarsened exact matching. Standard errors are in parentheses and clustered at repository level. ***, **, * indicate significance at 1%, 5% and 10%.

Figure 6: Country Border Effect on Collaborations on GitHub in 2012-21




Notes: Country border effect is calculated by running separate regressions as in Column (1) Table 1 for each year.
Source: Data from GH Torrent project (June 2020) and GitHub Archive publicly available on Google Cloud; authors' own calculations

Figure 7: Common Language Effect on Cross-Country Collaborations on GitHub in 2012-21




Notes: Common language effect is calculated by running separate regressions as in Column (1) Table 1 for each year.
Source: Data from GH Torrent project (June 2020) and GitHub Archive publicly available on Google Cloud; authors' own calculations

Table 5: Collaborations by team distribution (Full sample)

| | Dependent: N. contributions | | | |
|---|---|---|---|---|
| | Unmatched | | Matched | |
| | (1) | (2) | (3) | (4) |
| Constant | 0.204*** | 2.162*** | 0.156*** | 2.185*** |
| | (0.006) | (0.019) | (0.008) | (0.023) |
| Covid | −2.619*** | 0.308*** | −2.441*** | 0.468*** |
| | (0.047) | (0.062) | (0.062) | (0.073) |
| Distributed | −0.002 | 0.049** | 0.028*** | 0.007 |
| | (0.007) | (0.021) | (0.009) | (0.025) |
| Covid × Distributed | 0.663*** | 0.656*** | 0.422*** | 0.453*** |
| | (0.057) | (0.058) | (0.070) | (0.072) |
| Controls | | X | | X |
| Pseudo $R^2$ | 0.068 | 0.324 | 0.068 | 0.328 |
| N | 16,714,531 | 16,714,531 | 16,555,389 | 16,555,389 |
| Clusters | 279,076 | 279,076 | 276,545 | 276,545 |

Notes: The dependent variable is the monthly number of contributions. Included are repositories created between 2015 and 2018 which had 2 or 3 members within the first year of existence. Controls include repository age (quadratic), month fixed effects and repository-year × number-locations fixed effects. Columns (1) and (2) show unweighted poisson-regression results and Columns (3) and (4) show weighted poisson-regression results after applying coarsened exact matching. Standard errors are in parentheses and clustered at repository level. ***, **, * indicate significance at 1%, 5% and 10%.

Table 6: Collaborations by team distribution (Full sample)

| | Dependent: N. contributions | | | |
|---|---|---|---|---|
| | Unmatched | | Matched | |
| | (1) | (2) | (3) | (4) |
| Constant | 0.175*** | 2.138*** | 0.109*** | 2.114*** |
| | (0.007) | (0.020) | (0.009) | (0.023) |
| Covid | −2.598*** | 0.327*** | −2.442*** | 0.394*** |
| | (0.049) | (0.067) | (0.062) | (0.076) |
| Distributed | −0.022*** | 0.031 | 0.042*** | 0.033 |
| | (0.008) | (0.023) | (0.010) | (0.026) |
| Covid × Distributed | 0.663*** | 0.655*** | 0.496*** | 0.523*** |
| | (0.061) | (0.062) | (0.072) | (0.074) |
| Controls | | X | | X |
| Pseudo $R^2$ | 0.068 | 0.322 | 0.066 | 0.318 |
| N | 14,855,379 | 14,855,379 | 14,818,087 | 14,818,087 |
| Clusters | 248,355 | 248,355 | 247,761 | 247,761 |

Notes: The dependent variable is the monthly number of contributions. Included are repositories created between 2015 and 2018 which had 2 members within the first year of existence. Controls include repository age (quadratic), month fixed effects and repository-year × number-locations fixed effects. Columns (1) and (2) show unweighted poisson-regression results and Columns (3) and (4) show weighted poisson-regression results after applying coarsened exact matching. Standard errors are in parentheses and clustered at repository level. ***, **, * indicate significance at 1%, 5% and 10%.

# References


Barrero, Jose Maria, Nicholas Bloom, and Steven J Davis (2021). Why working from home will stick. Tech. rep. National Bureau of Economic Research.

Beine, Michel, Simone Bertoli, and Jesús Fernández-Huertas Moraga (2016). "A practitioners' guide to gravity models of international migration". The World Economy 39(4), pp. 496–512.

Belenzon, Sharon and Mark Schankerman (Oct. 2008). Motivation and Sorting in Open Source Software Innovation. CEPR Discussion Papers 7012. C.E.P.R. Discussion Papers.

DeFilippis, Evan, Stephen Michael Impink, Madison Singell, Jeffrey T Polzer, and Raffaella Sadun (2020). Collaborating during coronavirus: The impact of COVID-19 on the nature of work. Tech. rep. National Bureau of Economic Research.

Gibbs, Michael, Friederike Mengel, and Christoph Siemroth (2021). "Work from Home & Productivity: Evidence from Personnel & Analytics Data on IT Professionals". University of Chicago, Becker Friedman Institute for Economics Working Paper (2021-56).

Gousios, Georgios (2013). "The GHTorrent dataset and tool suite". Proceedings of the 10th Working Conference on Mining Software Repositories. MSR '13. San Francisco, CA, USA: IEEE Press, pp. 233–236.

Head, Keith and Thierry Mayer (2014). "Gravity equations: Workhorse, toolkit, and cookbook". Handbook of international economics. Vol. 4. Elsevier, pp. 131–195.

Hergueux, Jérôme and Nicolas Jacquemet (2015). "Social preferences in the online laboratory: a randomized experiment". Experimental Economics 18(2), pp. 251–283.

Hu, Albert G. Z. and Adam B. Jaffe (2003). "Patent citations and international knowledge flow: the cases of Korea and Taiwan". International Journal of Industrial Organization 21(6), pp. 849–880.

Iacus, Stefano M, Gary King, and Giuseppe Porro (2012). "Causal inference without balance checking: Coarsened exact matching". Political analysis 20(1), pp. 1–24.

Jaffe, Adam B and Manuel Trajtenberg (1999). "International knowledge flows: evidence from patent citations". Economics of Innovation and New Technology 8(1-2), pp. 105–136.

Jaffe, Adam B, Manuel Trajtenberg, and Rebecca Henderson (1993). "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations". The Quarterly Journal of Economics 108(3), pp. 577–98.

Laurentsyeva, Nadzeya (2019). "From Friends to Foes: National Identity and Collaboration in Diverse Teams". CRC TRR 190 Discussion Paper 226.

Lerner, Josh and Jean Tirole (2001). "The open source movement: Key research questions". European Economic Review 45(4-6), pp. 819–826.

— (2005). "The Economics of Technology Sharing: Open Source and Beyond". Journal of Economic Perspectives 19(2), pp. 99–120.

Lyons, Elizabeth (2017). "Team production in international labor markets: Experimental evidence from the field". American Economic Journal: Applied Economics 9(3), pp. 70–104.

Mayda, Anna Maria (2010). "International migration: A panel data analysis of the determinants of bilateral flows". Journal of Population Economics 23(4), pp. 1249–1274.

McDermott, Grant R and Benjamin Hansen (2021). Labor Reallocation and Remote Work During COVID-19: Real-time Evidence from GitHub. Tech. rep. National Bureau of Economic Research.

McDonald, Nora and Sean Goggins (2013). "Performance and Participation in Open Source Software on GitHub". CHI '13 Extended Abstracts on Human Factors in Computing Systems. CHI EA '13. Paris, France: ACM, pp. 139–144.

Migali, Silvia, Fabrizio Natale, Guido Tintori, Sona Kalantaryan, Sara Grubanov-Boskovic, Marco Scipioni, Fabio Farinosi, Cristina Cattaneo, Barbara Benandi, Marco Follador, et al. (2018). International Migration Drivers.

Ramos, Raul (2017). "Modelling migration". The Econometrics of Multi-Dimensional Panels. Springer, pp. 377–395.

Singh, Jasjit and Matt Marx (2013). "Geographic constraints on knowledge spillovers: political borders vs. spatial proximity". Management Science 59(9), pp. 2056–2078.

Thompson, Peter and Melanie Fox-Kean (2005). "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment". American Economic Review 95(1), pp. 450–460.

Tinbergen, Jan (1962). "Shaping the world economy; suggestions for an international economic policy".