RATIONALITY
& COMPETITION
CRC TRR 190

Defying Gravity: What Drives Productivity in Remote Teams?

Thomas Fackler (ifo Institute, LMU Munich, CESifo, Laboratory for Innovation Science at Harvard)
Michael Hofmann (LMU Munich)
Nadzeya Laurentsyeva (LMU Munich, CESifo)

Discussion Paper No. 427

September 19, 2023

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Georg Weizsäcker, Ph.D., Humboldt University Berlin, 10117 Berlin, Germany
info@rationality-and-competition.de

Defying Gravity: What Drives Productivity in Remote Teams?[a]

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

September 18, 2023

Abstract

How can teams organize for productive online collaboration? The coronavirus pandemic has led to a large and persistent shift toward remote work. Using fine-grained data from the world's largest platform for open-source software development, we find that the pandemic reduced the productivity of previously co-located teams substantially, whereas similar teams with remote work experience remained resilient. While access to remote talent and experience are important for overall team success, our results highlight the crucial role of communication for productive online collaboration. We find suggestive evidence that, with their peers shifting to online work, remote workers become better integrated into their teams' communication. We conclude that while teams' performance may suffer from the shift to remote work, setting up systems for effective online communication can help mitigate productivity loss.

JEL Codes: J01, M54, O30, F14

Keywords: gravity model, open source, knowledge workers, knowledge flows, remote work, online labor markets, COVID-19

---

[a]We thank Prithwiraj Choudhury, Florian Englmaier, Oliver Falck and Chris Stanton for valuable comments and suggestions. We also thank conference participants at the IIOC 2022, SIOE 2022, CESifo Economic of Digitization Area Conference 2022, ECLAC-ifo-PILLARS Workshop 2023, Workshop on the Economics of Firms and Labor 2023, 1st Organizational Economics Summer Symposium 2023. We thank Cemre Dane for excellent research assistance.
Support by the Deutsche Forschungsgemeinschaft through CRC TRR 190 (project number 280092119) and by the bidt Think Tank project "Changing workplaces: Patterns and determinants of technology and skill adoption by firms and individuals" is gratefully acknowledged. Thomas Fackler thanks the Laboratory for Innovation Science at Harvard for their hospitality while writing parts of this paper.

[b]Contact: fackler@ifo.de, ifo Institute, LMU Munich, CESifo, and Laboratory for Innovation Science at Harvard, Harvard University.

[c]Contact: michael.hofmann@econ.lmu.de, LMU Munich.

[d]Contact: nadzeya.laurentsyeva@econ.lmu.de, LMU Munich and CESifo.

1 Introduction

The growing importance of immaterial goods and increasing digitalization have enabled virtual production processes and have made virtual teamwork possible. Digital technologies not only reduce communication costs over long distances but also help create powerful environments with no physical barriers to collaboration and the exchange of knowledge and ideas. The COVID-19 pandemic has accelerated the adoption of technologies and new work practices by forcing teams to work remotely, at least during the lockdowns. In the (post-pandemic) future of work, remote work will likely be much more frequent than in the past, ranging from entirely remote companies to hybrid models that blend working from the office with working from anywhere (Barrero et al. 2021).

Do remote and hybrid work modes represent long-term viable solutions for the organization of knowledge teams? On the one hand, remote-work arrangements allow firms to access a larger talent pool of skilled workers and, hence, to address skill shortages, foster innovation and expand geographically. Simultaneously, many individual knowledge workers value flexibility and express their preference for a remote or hybrid job (Aksoy et al. 2022; Bloom et al. 2022). On the other hand, the existing research has stressed the importance of co-location for idea generation, complex problem-solving, knowledge transfer, and coordination (Bahar et al. 2022; Emanuel et al. 2022; Gibbs et al. 2021; Hu and Jaffe 2003; Yang et al. 2022).

This project aims at identifying characteristics that make remote collaboration in knowledge teams more or less productive. Our empirical setting is GitHub — the world's largest online platform for software development and code sharing. We analyze collaborations among open-source software engineers using the COVID-19 pandemic as a natural experiment, which forced GitHub teams (as many other knowledge teams worldwide) to work remotely with limited opportunities for offline interactions.

To motivate our analysis, we provide general evidence that geographic proximity still plays an important role in online collaborations on GitHub despite the available digital infrastructure for remote work and the fact that both the production process and the output of GitHub users are immaterial. By applying a standard gravity model to a city-pair panel data of global activities on GitHub between 2012 and 2021, we find that once distance doubles, the flow of expected collaborations between a given city pair drops by about 50%. The distance penalty is present at both extensive and intensive margins and is robust to controlling for technological differences. We note, however, that the role of geographic barriers has decreased with the start of the COVID-19 pandemic.

Our main analysis zooms in on the level of individual projects and focuses on their performance before and during the pandemic.[1] We benefit from rich data that allows us to observe teams and their spatial distribution before the pandemic. We can thus estimate how the productivity effect of COVID-19 varied between ex-ante *co-located* teams, in which all members shared the same location, ex-ante *fully distributed* teams, in which all members worked from distinct locations, and ex-ante *mixed* teams. While all types of teams had to work remotely during the pandemic, the prior reliance on face-to-face vs. remote collaboration and, hence, the actual exposure to the shock varied. It is plausible that co-located teams were hit the hardest as their team members had to adjust to the new work mode on top of other consequences of the pandemic.

First, we estimate the effect of the COVID-19 pandemic on performance, separately for each team type. To that end, we compare teams' performance during the pandemic to their hypothetical performance in the absence of the shock. We find that co-located teams are significantly less active during COVID-19 when compared to an artificial

[1] We refer to a group of individuals working on the same project as a team.

control group of earlier projects at similar maturity that did not experience the pandemic. However, we do not observe any negative effects among the fully distributed and mixed teams. If anything, the productivity of mixed teams seems to increase during the pandemic relative to a control group of earlier projects. These results confirm our intuition that co-located teams were negatively affected by the pandemic and show an interesting heterogeneity.

In order to investigate this heterogeneity more thoroughly, our second specification compares team performance *across*, rather than within team types. This allows us to better isolate the differential impact of the COVID-19 pandemic on different teams. Members of distributed teams made about 45% more code contributions than members of co-located teams, conditional on a rich set of pre-shock team characteristics. These results are robust to matching and placebo tests. The findings are consistent with the idea that distributed teams were already used to online collaboration before COVID-19 and thus were better prepared for fully remote work during the pandemic.

Next, we explore potential mechanisms and address concerns that co-located teams could be inherently different from distributed ones. The differential impact of COVID-19 on teams is not driven by differences in experience[2] or a better ability of distributed teams to source workers from the open-source community. Online communication, as proxied by comments on code contributions, is more abundant in distributed teams and plays an increased role in team productivity after the onset of COVID-19, which supports our idea that prior experience with remote work was beneficial during the pandemic. Note, that even combined these additional factors cannot (fully) explain the performance differences between co-located and distributed teams during the pandemic.

At the individual member level, we find suggestive evidence that remote workers

[2]Teams with more experienced members are more resilient to the pandemic on average, but this does not explain differences between co-located and distributed teams.

in mixed teams have become relatively more productive. This is consistent with the idea that remote workers benefited from better integration into their teams as all communication had to move online, such that they no longer missed out on offline exchanges.

Taken together our results show that organization matters for team productivity, highlighting, in particular, the crucial role of communication and of the ability to involve all team members regardles of their location.

This project relates to several strands of literature. First, it contributes to the growing body of studies on the determinants of team productivity, in particular, under remote and hybrid work arrangements. Many of these studies use the COVID-19 pandemic as a natural experiment and, in line with our conclusions, underline the importance of communication for the productivity of remote teams. Gibbs et al. (2021) investigate the difference in productivity before and during the work-from-home period of COVID-19. Using personnel data from over 10,000 skilled professionals at a large Asian IT services company, the study finds that working hours increased by 18%, while the average output decreased by 8-19%. The authors suggest that one of the reasons for the decline in productivity are higher communication and coordination costs associated with remote work. DeFilippis et al. (2020) examine how the COVID-19 pandemic has affected employees' digital communication patterns by conducting an event study of lockdowns in 16 large metropolitan areas across North America, Europe, and the Middle East. Using de-identified and aggregated meeting and email metadata from over three million users, the study reveals that compared to the pre-pandemic levels, the number of meetings per person and the number of attendees per meeting increased by 12.9% and 13.5%, respectively, while the average length of meetings decreased by 20.1%. Additionally, the study finds that the average workday increased by 48.5 minutes. Yang et al. (2022) analyze data from over 60,000 Microsoft employees in the US over the first six months of 2020. The study finds that remote work caused the collaboration network among workers to become more static and isolated with a decrease in synchronous communication, and an increase in asynchronous communication. According to the researchers, these changes could make it more difficult for employees to share new information across the network. Similarly, Emanuel et al. (2022) show that online interactions complement rather than substitute face-to-face interactions. By examining communication patterns of software engineers at a Fortune 500 company, the authors highlight the importance of geographic proximity for the knowledge exchange among coworkers, even if the latter happens online. Bojinov et al. (2021) conduct a field experiment to determine effective ways to onboard organizational newcomers who are working remotely. The results indicate the positive role of virtual water cooler talks. Their effect increases with participants' demographic proximity (in age, gender, and ethnicity), which facilitates online communication.

Several recent studies work with data on online collaborations from GitHub. McDermott and Hansen (2021) examine the impact of COVID-19 on labor activity. By analyzing data from GitHub, they show that the pandemic led to a significant shift in the pattern of labor allocation, with a higher likelihood for users to work on weekends and outside of regular working hours. Closely related to our research question, Lu et al. (2023) examine how remote GitHub teams were affected by the COVID-19 pandemic. The authors compare the productivity and sizes of the projects with what would have been expected in the absence of the pandemic. The study finds that the productivity and number of active members of GitHub teams varied considerably during different phases of the pandemic. Also, the resilience of a team under shock is closely tied to specific team characteristics before the pandemic, such as the country diversity, multitasking level, member experience and prestige, as well as emotions in team communication. However, Lu et al. (2023) focus on teams that already relied on remote work before the pandemic. In contrast, our study specifically investigates the differential impact of the pandemic on distributed teams and co-located teams, which allows us to analyze the productivity effects of moving from offline collaboration to an entirely virtual production process.

Second, by estimating a gravity model for online collaborations we contribute to the established literature that has helped 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. 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. Horton et al. (2017) use the gravity model to study contracting patterns on an outsourcing platform. By applying the gravity model to online collaborations in progress, we can identify the determinants of code contributions and benchmark them against findings established in trade literature. We can also benefit from finer geographical data and identify drivers of collaboration at the city, rather than country level.

Third, studying cross-city and cross-country code contributions, our paper is not only related to trade, but also to the literature on knowledge flows and knowledge production. Knowledge has been shown to be more localized 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 localization has turned out to be more robust over time than within-country localization (Thompson and Fox-Kean

2005). While a large body of this literature draws on analyzing patent data and thus focuses on inventors, we provide new evidence on collaboration and knowledge flows among software engineers.

The rest of the paper is organized as follows. Section 2 provides details on the GitHub dataset used for the analysis and presents estimations of the gravity model on GitHub to highlight the role of geographic proximity on the platform. Section 3 focuses on the performance of individual projects and remote workers. Section 4 concludes.

2 GitHub data and geography of online collaborations

2.1 Context and data

GitHub is a software development platform featuring a collaborative version control system and was launched in April 2008. As of early 2023, GitHub hosts the world's largest community of software developers comprising almost 100 million users and over 300 million repositories.[3]

GitHub projects cover a wide variety of (mostly) software applications, some of which provide tools for other developers, while some serve a wider audience. Projects can be started by both individual users and companies. GitHub allows its users to choose between private and public repositories for their projects. The latter are usually licensed under common open-source licenses such as the GNU General Public License, MIT License or Apache 2.0 License. Many open-source software repositories hosted on GitHub are used in thousands of other projects, including academic research, proprietary software, as well as projects in governmental and nonprofit organizations.

[3]The platform is constantly growing. For instance, in 2022 alone, 20.5 million new developers joined GitHub. Source: https://octoverse.github.com

Motivations of open source contributors have been the subject of economic research and include knowledge seeking and creation, career concerns (showcasing skills), paid work at software companies, 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).

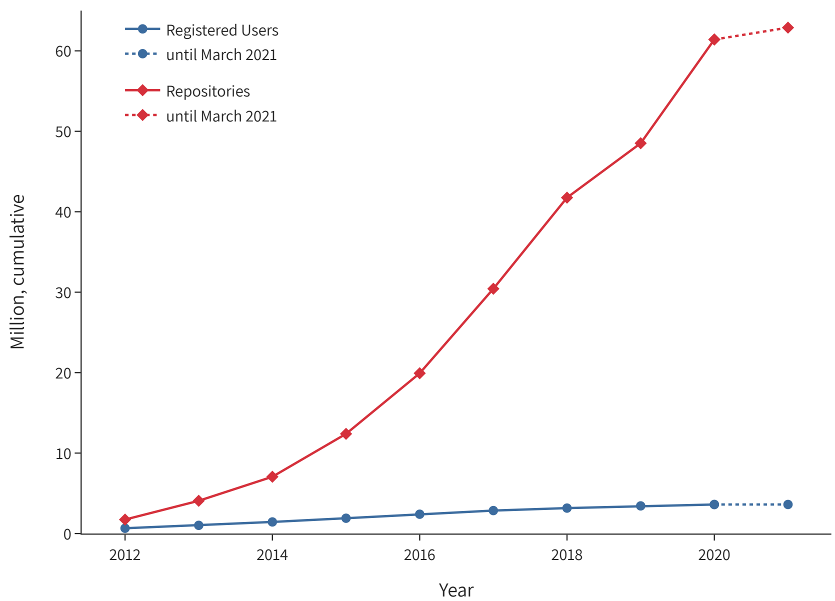
Our data cover the activities in GitHub's public repositories for which we can observe all code contributions to open-source projects and other user interactions, such as networking, code borrowing, bug reporting or commenting. To contribute to a public project or to create a new one, users have to set up an account (unless they already have one) where they can share their real name, location (usually a current city) and additional biographical information. Each project has only one owner. The owner may invite other users to contribute and to become project members. Users can also initiate a collaboration and contribute to a project even before being invited (McDonald and Goggins 2013). Users who are not project members can suggest modifications to the code, which the project members can review and either reject or accept. Users' profile pages on GitHub show their contributions to different public 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 each contribution. Along with tools for software development, GitHub shares features of social networks, allowing users to receive updates about each other's activities, follow projects and give "stars" to the ones they like.

We use two publicly available GitHub datasets for our analysis: a snapshot from GitHub Torrents (GHT) (Gousios 2013) and the GitHub Archive dataset (GHA).[4] Both datasets provide a mirror of the GitHub public event stream from 2012 on. We use the two datasets in a complementary way. We take the event stream data from

[4]https://www.gharchive.org

GHA because it is updated in real time and allows us to incorporate the up-to-date activity data. We then merge GHA events with data on users (in particular, their reported geographic locations), which is available in the GHT dataset. We use the latest available snapshot of the GHT dataset from March 2021. Our event data from GHA spans from 2012 to 2021. Given our research question, we have to limit the data to events where we can identify the geographic location of project owners and project committers. As Figure 1 shows, that leaves us with about 3.6 million registered users and about 62.6 million repositories.[5] Section A in the Appendix provides technical details on merging the GHA and GHT datasets.

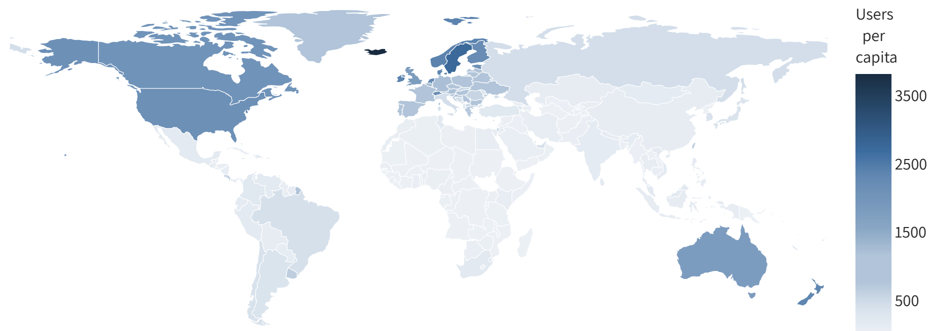
Figure 1 *Cumulative number of registered users and public repositories on GitHub*



Notes: This graph shows the cumulative number of GitHub users with reported locations and the number of public repositories owned by users with reported locations.

[5]In total, as of March 2021 there were about 36 million registered users working on GitHub public repositories and 183.5 million public repositories.

Figure 2 *Number of contributions on GitHub*



Notes: This map shows the number of GitHub users per capita (population in millions, i.e. users per one million inhabitants). It is based on users with reported location and repositories owned by users with reported location.

2.2 Importance of geographic proximity on GitHub

Since its start in 2008, GitHub has become popular with users worldwide. Figure 2 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. As of March 2021, the top five countries in terms of the absolute number of GitHub users were the United States, India, China, Great Britain and Brazil.

The main purpose of GitHub has been to ensure smooth collaboration and knowledge exchange among users regardless of their location. Unlike many other settings in which remote work is a relatively new phenomenon, GitHub teams have had the necessary technological infrastructure to function virtually ever since 2008. Therefore, the first question we ask in this paper is, to what extent geographic proximity still matters for collaboration in GitHub teams.

To quantify the role of geographic barriers, we adapt the standard gravity model from the trade literature and estimate the gravity of collaborations on GitHub at the city-pair level. We aggregate the combined GHT-GHA dataset (2012–2021) at a city-pair and year level. We further restrict our data to about 700 of the most active cities on GitHub (as proxied by the number of registered users as of March 2021).[6] 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. Section B in the Appendix provides details of the empirical specification.

Our results in Table 1 show that geographic proximity matters even in virtual environments on GitHub. The effect of geographic distance on online collaborations is negative and statistically significant with an estimated elasticity of 0.493: if distance doubles, the number of commits drops by about 50% (column 1); conditional, on non-zero collaboration, the number of commits drops by about 40% (column 2). On the extensive margin (column 3), the linear probability results show that if distance doubles, the probability of having a cross-city collaboration drops by about 2.4 percentage points.

Column (4) in Table 1 uses distance bins instead of continuous distance measures to capture non-linear distance effects. The reference category corresponds to collaborations within the same city. The results highlight a 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 or who can collaborate in an offline setting. Beyond the distance of 300km, the effect stays at about the same level.

These effects are economically significant. As Table A1 in the Appendix shows, the

[6]We set a cutoff of at least 400 registered users per city as of March 2021, resulting in 727 cities. Our results are robust to setting a lower cutoff.

distance elasticity for GitHub collaborations is almost two-thirds compared to that for trade, even though the role of trade costs (transportation, legal costs, search costs for partners) should be negligible on the platform. Our results also show, that conditional on distance, state borders reduce virtual collaborations, while a common language mitigates this negative effect. Similar patterns have been observed in the literature on trade, migration and knowledge flows.

Table A2 in the Appendix compares the baseline results with those obtained from running regressions within the same programming language. Hence, we check to what extent the effect of distance is driven by technological differences between cities. While the magnitude of the coefficient decreases (from -0.420 to -0.282), it remains statistically significant suggesting that technological differences cannot fully explain the gravity in online collaborations.

Given the emerging evidence that the COVID-19 pandemic has transformed the organization of work, we are interested in tracing whether the role of proximity for online collaboration has changed between 2012 and 2021. In Figure A2 in the Appendix, we show how the coefficient for distance has 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 a drop in the role of proximity that coincides with the COVID-19 pandemic: in 2020, distance elasticity constituted -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. The pandemic and associated lockdowns likely represented a stronger shock for ex-ante co-located teams compared to already distributed ones, because the former had to adjust their work practices under time pressure. It could also align 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 offline communication.

Table 1 *Gravity model for collaborations on GitHub*

Variables	Contributions (1)	Contributions >0 (2)	Contributions yes/no (3)	Contributions distance dummies (4)
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	0.659	0.520	0.109	0.669

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). For specifications estimated with PPML, pseudo R-squared is calculated.

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

3 The effect of COVID-19: team-level analysis

The analysis in the previous section has shown that international online collaboration patterns have changed during the pandemic. This section moves from the aggregate cross-city perspective to the level of individual projects. How was the activity of existing projects affected by COVID-19 and how did this effect differ by (co-)location of team members?

Our analysis focuses on small projects, which have attracted exactly three members within the first year of existence. We select these small teams for mainly two reasons. First, it allows us to clearly distinguish team compositions into three categories: *Co-located* teams in which all members share the same locations, *mixed* teams in which two members are from the same location and one member is remote, and *(fully) distributed* teams consisting of members from three distinct locations. Second, fixing the number of members ensures that we do not mechanically introduce systematic team-size differences between the categories, since larger teams are more likely to contain at least one non co-located member.[7] Note, however, that we only fix the number of members within the first year and allow additional members and contributions from non-members afterwards. We explicitly investigate this potential mechanism below.

We further choose projects which were started between 2015 and 2018 and thus analyze contributions to pre-existing projects but not to new projects. This way we

[7]Furthermore, our requirement that all team members state their location would less likely be satisfied for larger teams.

exclude effects from systematic differences in team composition of newly formed teams during the pandemic. Using earlier years than 2015 would not add much power, given that only few projects last for more than 5 years. Instead, it might introduce projects that are less comparable if the type of projects publicly developed on GitHub has shifted over time. Including projects that were started after 2018 would limit our ability to observe how these projects performed before the onset of the coronavirus pandemic.[8]

We investigate the effect of the coronavirus pandemic by employing two complementary difference-in-differences (DiD) designs. First, in Section 3.1 we estimate how COVID-19 affected team productivity for the three types of teams individually. For this, we create artificial controls groups using earlier projects. Second, in Section 3.2 we more directly estimate the differential effects of COVID-19 on co-located, mixed and distributed teams. While the first approach allows us to estimate *actual* performance implications of the pandemic, the second approach estimates *relative* performance differences across teams, enabling us to net out general pandemic effects that are independent of team types.

3.1 The effect of COVID-19 on team performance

In our first analysis we use a difference-in-differences design to compare the performance of teams during the coronavirus pandemic to the performance during normal times, separately for the three types of teams.

3.1.1 Empirical setup

Since COVID-19 is a global pandemic affecting all teams at (roughly) the same time, we construct an artificial control group using the following approach: We define projects

[8] In addition, we select projects consisting of three members within the first year. For projects created after February 2019 this could partly include members who joined during COVID-19, leading to undesired selection effects.

which were started in 2018 as the treatment group and compare their performance during the pandemic to the performance of earlier teams at a similar maturity. For these earlier projects we mimic the COVID-19 timing of 2018 projects, i.e. we assume a hypothetical onset of the COVID-19 pandemic in the third calendar year of their existence. Thus, the hypothetical outbreak of the coronavirus is set to March 2019 for projects started in 2017, March 2018 for projects started in 2016 and March 2017 for projects started in 2015.

We restrict the observational period to 24 months, centered around the hypothetical coronavirus outbreak. This ensures that none of the control teams ever actually experience the COVID-19 pandemic, thus avoiding the empirical challenges through staggered treatment (e.g. Callaway and Sant'Anna 2021). Further, we only keep projects which received at least one contribution in the observational period and we remove outliers which received more than 500 contributions within the first two years of existence.

In order to estimate the effect of the coronavirus pandemic we estimate the following equation using a Negative-Binomial regression to account for the fact that the number of commits is count data:

$$c_{it} = \beta_0 + \beta_1 \text{COVID}_t \times \text{Treated}_i + \beta_2 \text{COVID}_t + \beta_3 \text{Treated}_i + X_{it}\delta + \epsilon_{it}, \qquad (1)$$

where c_{it} is the number of code contributions to project i in month t. COVID_t denotes the hypothetical COVID-19 pandemic and is set to one in March of the respective year (see above). Treated_i is set to one for projects created in 2018, those which actually experience the pandemic, and zero for projects started earlier. X_{it} is a vector of control variables and ϵ_{it} is the error term, which is clustered at the project level. The coefficient of interest is β_1 which measures the impact of the (actual) coronavirus pandemic on the activity of treated projects.

We split the sample by team composition and estimate Equation (1) separately for co-located, mixed and fully distributed teams. Because co-located teams had to adjust to new virtual work modes, we expect their performance to decrease during the pandemic (negative β_1). The effect on mixed and distributed teams is ex-ante not clear and therefore we let the data speak. In addition to the split-sample analysis, we estimate an extended version of Equation (1): We add triple interactions between COVID_t, Treated_i and indicators for mixed teams, D(Loc.=2)_i, and fully distributed teams, D(Loc.=3)_i in order to estimate COVID-19 effects on all team types in a single regression.

The key identifying assumption in our setting is that in absence of the coronavirus pandemic the activity of projects started in 2018 would have followed a parallel trend to the activity of those started earlier, conditional on a set of control variables. Note, that this design only compares projects within and not across team types and therefore identification does not hinge on parallel trends across co-located, mixed and distributed teams.

To ensure the robustness of our estimates we add a rich set of control variables. These controls include month fixed effects to capture seasonality in commits; repository age (measured in months, linear and quadratic) to model declining activity as projects mature; the number of commits and the number of watchers within the first year of a project's existence to account for differences in project productivity and success; the country of the project owner[9] to control for differences in COVID-19 responses across countries. Except for project age and month fixed effects, we allow all of these controls to have time-varying effects before and after the onset of the (hypothetical) coronavirus pandemic. In the Appendix we estimate two-way fixed-effects models, dropping month fixed effects but adding time and project fixed effects while keeping interacted controls.

[9] We distinguish between the five most represented countries in our sample (USA, Brazil, Great Britain, India and Canada) and an other category.

Also in the Appendix, we use matching methods to further probe the robustness of our results.

3.1.2 Results

To get a first impression of the effect of the pandemic on team productivity, Figure 3 shows two event-study graphs of quarterly effects on the productivity (number of commits) in different types of teams. Figure 3a splits our sample into co-located teams and distributed teams (including both fully distributed and mixed teams). While the blue line for distributed teams is rather flat throughout, the red line for co-located teams clearly decreases after the start of the pandemic and becomes significant at the 5% level in Q4 of 2020.

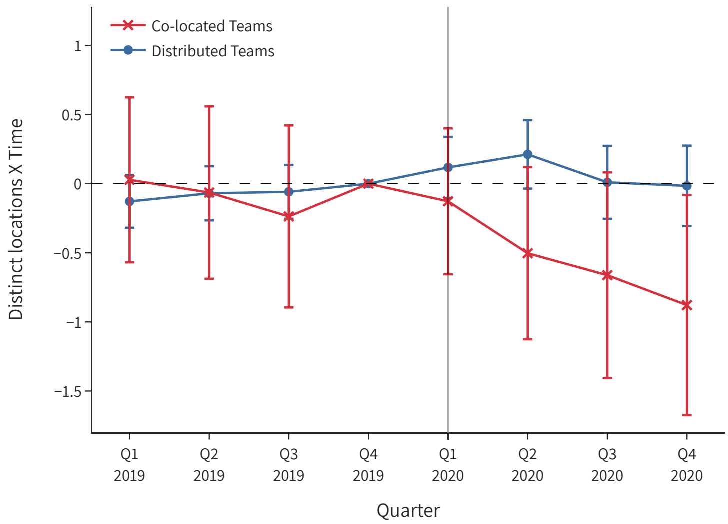
Figure 3b further subdivides distributed teams in those with three locations (fully distributed) and those with two locations (mixed teams). Although differences are not statistically significant, this graph shows an interesting pattern that we will explore further in the analyzes on remote worker productivity in Section 3.3: The yellow line (mixed teams) appears to increase slightly with the start of the pandemic, whereas the blue line (fully distributed teams) is flat throughout 2020. Overall, Figure 3 confirms the intuition that co-located teams, which had to adjust to online collaboration, were significantly negatively affected, while mixed teams show a small increase and fully distributed teams were hardly affected.

Table 2 summarizes the results of estimating Equation (1).[10] In columns (1), (2) and (3) we split the sample into co-located, mixed and fully distributed teams, respectively. For fully distributed teams the interaction effect in column (3) is close to zero and insignificant, confirming that these teams are barely affected by the coronavirus

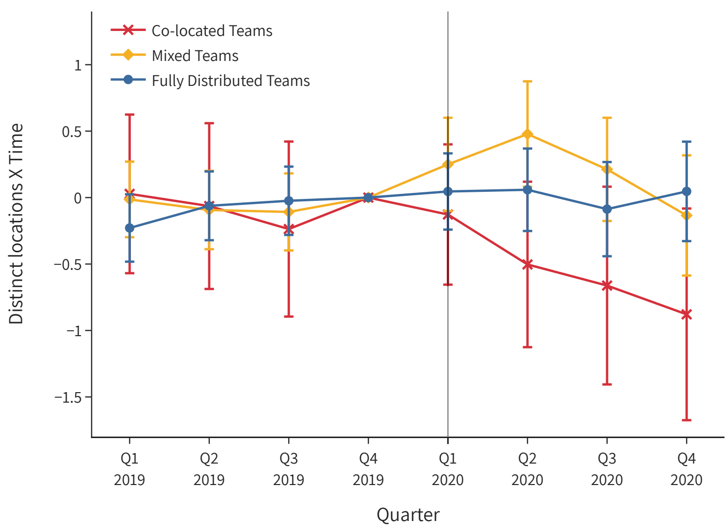
[10] We interact COVID_t with multiple control variables. Therefore, its main effect and the constant are omitted since they cannot be interpreted in a meaningful way.

Figure 3 *Quarterly effects of team distribution (compared to earlier projects)*

(a) *Co-located vs. distributed*



(b) *Number of locations*



Notes: The figures show interaction coefficients and 95% confidence intervals of quarterly time dummies and indicators for projects started in 2018. Each line is from a separate regression with earlier projects as comparison category. The dependent variable is the number of commits and controls include project age (linear & quadratic), number of commits and watchers within a project's first year and country-of-owner fixed effects. Panel (a) splits the sample into co-located teams and distributed teams (two and three locations). Panel (b) splits the sample into co-located, mixed and fully distributed teams. Estimation method: Negative-Binomial ML.

Table 2 *COVID-19 effects of team distribution*

	Dependent: Number of commits			
Team composition	Co-located (1)	Mixed (2)	Distributed (3)	Full sample (4)
Treated	−0.033	0.015	0.162***	−0.067
	(0.115)	(0.069)	(0.058)	(0.119)
Treated × COVID	−0.424**	0.150	0.031	−0.366*
	(0.199)	(0.127)	(0.121)	(0.204)
D(Loc.=2)				0.020
				(0.075)
D(Loc.=3)				0.096
				(0.068)
Treated × D(Loc.=2)				0.064
				(0.140)
Treated × D(Loc.=3)				0.233*
				(0.131)
D(Loc.=2) × COVID				−0.084
				(0.136)
D(Loc.=3) × COVID				0.088
				(0.138)
Treated × D(Loc.=2) × COVID				0.520**
				(0.245)
Treated × D(Loc.=3) × COVID				0.457*
				(0.255)
Controls	X	X	X	X
Pseudo R^2	0.121	0.118	0.094	0.132
N	18,888	68,136	106,968	193,992
Clusters	787	2,839	4,457	8,083

Notes: Columns (1), (2) and (3) split the sample into co-located, mixed and distributed teams, respectively. *Treated* projects are founded in 2018 and are exposed to 12 months of COVID-19. The control group are earlier projects with the hypothetical onset of the pandemic shifted backwards. For example, the hypothetical onset of the pandemic is set to March 2019 for projects started in 2017. Controls include project age (linear & quadratic), number of commits and watchers within a project's first year, country-of-owner fixed effects and month fixed effects. Except for project age and month fixed effects all controls are interacted with the (hypothetical) COVID indicator. The sample covers 24 months centered around the hypothetical onset of the coronavirus pandemic. Estimation method: Negative-Binomial ML. Standard errors are clustered at the project level and shown in parentheses. ***, **, * indicate significance at 1%, 5% and 10%.

pandemic.[11] The coefficient for mixed teams in column (2) is statistically insignificant, as well, but the point estimate is slightly larger at 0.150. The effect on co-located teams in column (1) is negative, the largest in magnitude (-0.424) and significant at the 5% level. The interaction effect suggests that during the pandemic co-located projects were only about 65% ($e^{-0.424} = 0.654$) as productive as one would expect had COVID-19 not happened. In column (4) we estimate an alternative specification, adding triple interactions to measure differential treatment effects for the three team types in a single regression. This yields qualitatively very similar results. The coefficient of $\text{COVID}_t \times \text{Treated}_i$ measures COVID-19 effects for co-located teams (the left-out category) and is negative, slightly smaller than in column (1) and still significant at the 10% level. Both triple interactions for mixed and distributed teams are positive and significant, offsetting the general negative impact of the coronavirus pandemic. Table A5 in the Appendix repeats all four regressions including project fixed and time fixed effects. This not only improves the amount of explained variance (Pseudo R^2), but also leads to even stronger results, especially for co-located teams.[12]

We additionally probe the robustness of our results using two matching methods. First, we use the inverse probability weighting (IPW) approach suggested in Abadie (2005) which keeps all projects but re-weights them to achieve balance between treatment and control group. Second, we use 1:1 propensity score matching applying the optimized algorithm of Hansen and Klopfer (2006). We estimate propensity scores using logistic regressions of Treated_i on the full set of control variables[13] and in two alternative approaches we add pre-pandemic outcome levels or pre-pandemic outcome

[11] The significant effect of Treated in column (3) may indicate that fully distributed teams that were started in 2018 were somewhat more productive than earlier fully distributed teams before the pandemic.

[12] Negative-Binomial models do not converge if both project fixed and time fixed effects are included. Therefore these estimates are obtained using Poisson ML. Negative-Binomial models using project fixed and month fixed effects, instead of time fixed effects, yield very similar results (not shown).

[13] We include project age (at the beginning of the observational period), the number of commits and watchers in the first year of a project's existence as well as country of the project owner.

trends.[14]

Combining the different methods and propensity scores and applying them separately to co-located, mixed and fully distributed teams yields 18 individual regressions, which are summarized in Table A6 in the Appendix.[15] Throughout, the results are closely comparable to those of unmatched regressions in Tables 2 and A5, with two noteworthy (but minor) deviations. First, the estimated negative impact of COVID-19 on co-located teams appears to be even larger when applying 1:1 matching. Second, also when using 1:1 matching, the positive COVID-19 effect on mixed teams is statistically significant at the 10% level in columns (4) and (5). However, Figure A3 suggests that 1:1 matching performs worse than IPW in balancing the covariates across treatment and control projects. Furthermore, we are aware of the general caveats of matching in DiD settings, such as regression to the mean (e.g. Daw and Hatfield 2018), which can increase bias especially when matching on pre-treatment outcomes. Therefore, we refrain from over-interpreting these observations and prefer unmatched results, but conclude that results from the matched regressions further bolster the robustness of our main findings.

3.2 Differential effects of COVID-19 across team types

While the design in the previous subsection allows us to quantify the effect of the coronavirus pandemic separately for co-located, mixed and fully distributed teams, the estimates capture both team-type specific effects and general effects of COVID-19, which are similar for all teams. The combined magnitude and direction of the latter are a

[14]We include quarterly pre-pandemic outcome levels (number of commits), yielding four observations per project, and match on trends by including first-differences of these quarterly levels.

[15]Figure A3 shows standardized mean differences of control variables across treated and control projects. It suggests that, in general, the IPW approach yields better balance than 1:1 matching. Adding pre-pandemic outcome levels or trends slightly improves the balancing of these factors.

priori not clear.[16] Although the resilience of distributed teams supports the notion that the negative impact on co-located teams is driven by the fact that they are co-located, we cannot entirely rule out alternative channels. Therefore, we directly estimate relative performance differences across team types during the coronavirus pandemic in this subsection. This approach better isolates the *differential* impact of the coronavirus pandemic on the three types of teams from general disruptions which affect all teams. One further advantage of this approach is that it does not rely on comparing projects from 2018 to an artificially created control group with hypothetical pandemic periods, but allows us to estimate COVID-19 effects based on projects of all vintages. However, this comes with the caveat that we can only estimate relative performance differences between teams and thus have to rely on our previous results of Section 3.1 to quantify actual performance effects of COVID-19.[17]

3.2.1 Empirical setup

We estimate the following alternative specification, again using Negative-Binomial regression:

$$\begin{aligned} c_{it} = {} & \beta_0 + \beta_1 \text{COVID}_t \\ & + \beta_2 \text{D(Loc.=2)}_i + \beta_3 \text{D(Loc.=3)}_i \\ & + \beta_4 \text{D(Loc.=2)}_i \times \text{COVID}_t + \beta_5 \text{D(Loc.=3)}_i \times \text{COVID}_t \\ & + X_{it}\delta + \epsilon_{it}. \end{aligned} \tag{2}$$

The dependent variable is the number of commits to project i in month t. We

[16] For example, COVID-19 lock-downs could have increased the time available to work on GitHub projects for some users, but could also have limited the available time for parents if they have to supervise kids who are in homeschooling (e.g. Myers et al. 2020).

[17] A relative performance gain of distributed teams over co-located teams can be driven by either distributed teams becoming more productive or co-located teams becoming less productive or a combination of both dynamics.

account for the difference in activity that affected all projects after the (actual) start of the pandemic (coefficient β_1, where the regressor is a dummy that is equal to one starting March 2020 and zero before), as well as for the differences between co-located projects (reference category)[18] and those with two or three different locations in the team. X_{it} is a vector of control variables and ϵ_{it} is the error term, which we cluster at the project level. The coefficients of interest are β_4 and β_5, which estimate the differential effect of the pandemic on projects with two and three different locations, respectively. Based on our intuition and previous results we expect both coefficients to be positive, indicating relative performance gains for mixed and distributed teams compared to those that are co-located.

Equation (2) estimates the causal differential impact of COVID-19, if activity patterns of co-located, mixed and distributed teams would have followed parallel trends in the absence of the pandemic.[19] We probe the robustness of this assumption by adding a rich set of controls, documenting insignificant pre-pandemic trends, using matching methods and performing a placebo analysis. If we further assume that the idiosyncratic COVID-19 shocks of team members are independent of team types we isolate the effect of being geographically co-located or distributed. In addition to adding controls and applying matching methods, we explicitly probe the plausibility of this assumption in Section 3.2.3 which explores potential mechanisms and alternative explanations.

Control variables include month fixed effects, project age (measured in months, linear and quadratic), the number of commits and watchers within the first year, the country of the project owner and projects' starting year to control for potential systematic differences in team composition and coronavirus response across project vintages.

[18] The previous analysis has revealed that co-located teams are affected the most by the coronavirus. Therefore, we choose co-located as reference category since this allows us to compare their performance to both mixed and fully distributed teams.

[19] This assumption differs from the one of Equation (1) as it requires parallel trends *across team types* rather than within team types but *across project vintages*.

Again, we allow all of these controls to have time-varying effects before and after the onset of the coronavirus pandemic, except for project age and month fixed effects. In our most restrictive specification we add project and time fixed effects, which replace the main effects of our controls but not the interaction terms with COVID_t.

3.2.2 Results

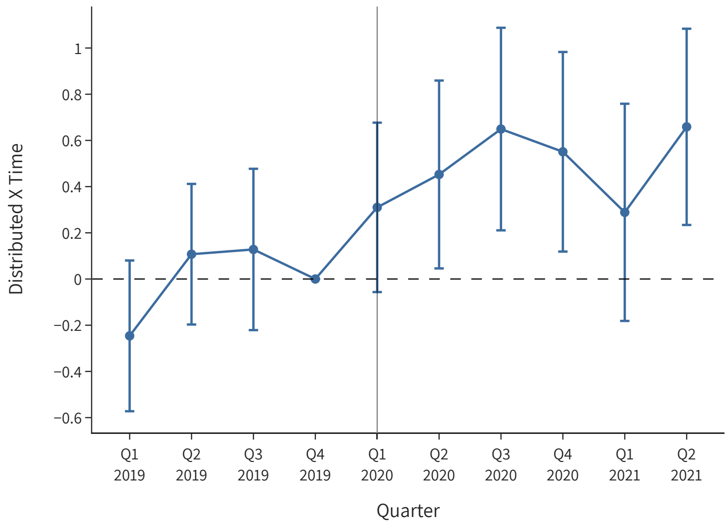
Figure 4 shows the quarterly performance of distributed teams relative to co-located teams in two event study graphs and Figure 4a pools mixed and fully distributed teams together. While pre-pandemic coefficients are close to zero[20] and insignificant, we observe a clear positive shift in relative performance of distributed teams starting with the onset of the pandemic and turning significant from the second quarter of 2020. Figure 4b estimates relative performance of mixed and fully distributed teams separately. The patterns for both types of teams look very similar, but mixed teams seem to have slightly larger relative performance gains than fully distributed teams. However, this difference is not statistically significant and fades in the fourth quarter of 2020. Taken together, Figure 4 confirms our intuition that distributed teams are more resilient to the COVID-19 shock than co-located teams. Based on our results in the previous subsection, these relative performance differences during the pandemic are more likely driven by productivity losses of co-located teams rather than productivity gains of distributed teams.

Table 3 shows detailed regression results. In all specifications, co-located teams are

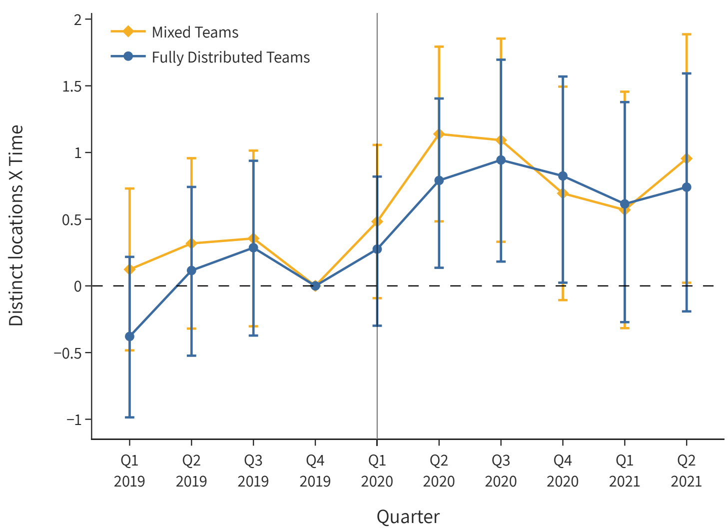
[20] The point estimate of the coefficient for Q1 2019 is a negative insignificant outlier. It is driven by very "young" projects, which seem to be slightly more productive when team members are co-located. Removing projects started either after June 2018 or alternatively after October 2018, changes this estimate to nearly zero (June 2018: 0.007; October 2018: 0.036) while leaving the remaining coefficients largely untouched.

Figure 4 *Quarterly effects of team distribution*

(a) *Co-located vs. distributed*



(b) *Number of locations*



Notes: The figures shows interaction coefficients and 95% confidence intervals of quarterly time dummies and indicators for distributed teams. In all regressions the dependent variable is the number of commits and controls include project age (linear & quadratic) and starting year, number of commits and watchers within a project's first year and country-of-owner fixed effects. Panel (a) compares co-located teams to distributed teams (two and three locations). Panel (b) compares co-located teams to mixed teams and fully distributed teams. Estimation method: Negative-Binomial ML.

Table 3 *Collaborations by team distribution*

	Dependent: Number of commits					
	(1)	(2)	(3)	(4)	(5)	(6)
D(Loc.=2)	−0.027	−0.009	0.022	0.009	−0.004	
	(0.104)	(0.106)	(0.107)	(0.107)	(0.103)	
D(Loc.=3)	0.132	0.157	0.189*	0.177*	0.142	
	(0.095)	(0.097)	(0.099)	(0.099)	(0.098)	
D(Loc.=2) × COVID	0.377**	0.320**	0.296**	0.295**	0.298**	0.356**
	(0.150)	(0.145)	(0.144)	(0.145)	(0.142)	(0.147)
D(Loc.=3) × COVID	0.587***	0.514***	0.492***	0.472***	0.418***	0.486***
	(0.151)	(0.149)	(0.141)	(0.137)	(0.135)	(0.143)
Project age		X	X	X	X	X
Project year FE		X	X	X	X	X
N. Commits (first year)			X	X	X	X
N. Watchers (first year)				X	X	X
Owner country					X	X
Month FE		X	X	X	X	
Project FE						X
Time FE						X
Pseudo R^2	0.008	0.016	0.057	0.059	0.063	0.518
N	102,720	102,720	102,720	102,720	102,720	102,720
Clusters	4,280	4,280	4,280	4,280	4,280	4,280

Notes: Co-located teams are the left-out category. Project age is included in linear and quadratic form. Except for project age and month fixed effects, all controls are interacted with the COVID indicator. Column (6) includes project and time fixed effects, project age (linear & quadratic) and the interacted remaining controls. The sample covers 24 months centered around the onset of the coronavirus pandemic in March 2020. Estimation methods: columns (1) to (5): Negative-Binomial ML; column (6): Poisson ML. Standard errors are clustered at the project level and shown in parentheses. ***, **, * indicate significance at 1%, 5% and 10%.

the reference category and the dependent variable is the number of monthly commits.[21] The coefficients for mixed and fully distributed teams (with two and three distinct locations, respectively) show only small differences in (pre-pandemic) productivity compared to co-located teams. The interaction of COVID_t and the dummies for mixed and fully distributed teams confirm the patterns observed in Figure 4. Teams which are at least partially distributed, did statistically significantly better during the pandemic than co-located teams. Across specifications, fully distributed teams did best, while

[21] Since we interact COVID_t with multiple control variables including project-founding-year fixed effects, we can only interpret relative differences between team types but not the main effect of COVID_t, which we therefore omit in Table 3. In column (1), without any controls, the coefficient of COVID_t captures both the negative effect of the pandemic on co-located teams and the general decline in activity as projects mature.

mixed teams were in between. Note, that this may seem to contradict the order of effects seen in Figure 4b at first. However, the regressions estimate an average effect over the entire post-pandemic period such that the initial relative increase in productivity for mixed teams can be too small to determine the overall effect.

Starting from the simple regression in the first column, the specifications in Table 3 add more and more controls until column (5) which includes the full specification of interacted controls. Column (6) replaces most of their main effects with project fixed effects[22] and includes time fixed effects instead of month fixed effects, which improves the explained variance (Pseudo R^2). Across all specifications our main results are stable, both in magnitude and significance. Our most restrictive specifications in columns (5) and (6) suggest that during the pandemic mixed teams on average received 34% to 43% more code contributions than co-located teams, while fully distributed teams received even 52% to 65% more contributions.[23]

In the Appendix we probe the robustness of our main findings with two additional analyses. First, we re-estimate Equation (2) after applying inverse probability weighting and 1:1 propensity score matching. Analog to the approach in Section 3.1 we use logistic regressions to estimate three propensity scores, including just controls, controls and pre-pandemic outcome levels or controls and pre-pandemic trends.[24] We combine mixed and fully distributed teams into one category ($Distributed_i = 1$). This simplifies the matching process and the differences between these projects are not significant anyway. Figure A4 in the Appendix shows the balance of control variables after applying the individual approaches. All methods perform well and significantly improve the balance between co-located and distributed teams. Table A7 summarizes the regression

[22]The interaction terms of controls and COVID_t are still included.

[23]These percentages are calculated using the following formula: $(e^{\beta} - 1) * 100$.

[24]Controls include project age (at the beginning of the observational period), project starting-year, the number of commits and watchers in the first year as well as country of the project owner. Pre-pandemic outcome levels are measures as quarterly number of commits and pre-pandemic outcome trends are measured as first-differences of quarterly levels.

results and column (1) shows a baseline specification without any matching applied. The interaction coefficient is highly significant and the estimate of 0.372 lies in between the respective coefficients for mixed (0.298) and fully distributed (0.418) teams from Table 3 column (5). Across all matching methods the interaction coefficients stay highly robust and similar in magnitude to the baseline specification in column (1). Applying 1:1 matching on controls and pre-pandemic outcome levels or trends reduces the statistical significance to the 5% level. However, the sample size is also roughly cut in half, which lowers estimation power. Overall, Table A7 shows that our results are robust to matching and suggests a relative performance gain of about 45% for (at least partially) distributed teams over co-located teams during the pandemic.

Second, one might be concerned that distributed teams are systematically more likely to work on long-term projects than co-located teams and that our estimates capture this dynamic rather than exceptional differences caused by COVID-19.[25] If this was the case, we should already observe similar effects in the years before the coronavirus pandemic. We address this concern in Table A8, which summarizes results of placebo tests around alternative pandemic dates. In columns (1) and (2) we consider treatment effects if COVID-19 had happened one year earlier. We remove all projects created after 2017 and set our COVID_t indicator to one from March 2019 on.[26] Otherwise, we re-estimate our main specification without controls in column (1) and with the full set of interacted controls in column (2). Columns (3) and (4) repeat this exercise, shifting the onset of the pandemic backwards an additional year to March 2018. Throughout, the interaction coefficients for mixed and fully distributed teams are insignificant, indicating no systematic differences between co-located and

[25]This concern does not apply to our design in Section 3.1, which estimates COVID-19 effects *within* and not *across* team types (co-located, mixed and fully distributed).

[26]We again consider an observational period of 24 months centered around the hypothetical onset of the coronavirus pandemic and thus exclude the true COVID-19 period by keeping only observations until February 2020.

distributed teams. These results are therefore reassuring that our main specification captures COVID-19 induced differences in productivity.

3.2.3 Mechanisms and additional results

So far we have consistently documented a differential impact of the coronavirus pandemic on co-located, mixed and distributed teams, which is largely driven by a negative productivity shock of co-located teams. We argue that this is largely driven by co-located teams having lost their comparative advantage through offline collaboration, while distributed teams already had a functioning online production process in place before the pandemic. In order to explore underlying mechanisms and dismiss alternative explanations, Table 4 adds additional control variables to the baseline specification in column (1). For each additional variable, an interaction term with COVID_t is added as well, to understand its role during the pandemic.

One possible alternative channel is that experienced developers are more likely to join distributed teams and at the same time are more resilient to pandemic shocks than less experienced developers. We address this in column (2) adding a measure for a team's pre-pandemic experience: the log average number of commits by team members to any project. First, the averages at the bottom of column (2) reveal that co-located teams have the least experienced members, members of mixed teams are more experienced and fully distributed teams' members are the most experienced. Second, experience has a highly significant positive effect on a team's productivity. Furthermore, the pandemic increased the importance of experience a lot, as the coefficient of the interaction with COVID_t indicates, which is more than half the size of the main effect of experience. However, the interaction coefficients of mixed and fully distributed teams are virtually unaffected by the inclusion of experience.

A second alternative channel is that distributed teams, having an established online

Table 4 *Collaborations by team distribution (mechanisms)*

	Dependent: Number of commits					
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.698	−0.085	1.109**	−1.170*	−0.453	
	(0.577)	(0.539)	(0.554)	(0.614)	(0.556)	
COVID	−0.402	−1.085***	−0.252	−1.616***	−1.598***	
	(0.264)	(0.276)	(0.257)	(0.424)	(0.412)	
D(Loc.=2)	−0.004	−0.044	−0.032	0.015	−0.060	
	(0.103)	(0.101)	(0.101)	(0.099)	(0.098)	
D(Loc.=3)	0.142	0.071	0.003	0.102	−0.053	
	(0.098)	(0.096)	(0.097)	(0.093)	(0.094)	
D(Loc.=2) × COVID	0.298**	0.294**	0.250*	0.339**	0.266*	0.333**
	(0.142)	(0.149)	(0.145)	(0.149)	(0.154)	(0.147)
D(Loc.=3) × COVID	0.418***	0.412***	0.337**	0.399***	0.324**	0.408***
	(0.135)	(0.143)	(0.139)	(0.139)	(0.150)	(0.145)
Experience		0.143***			0.114***	
		(0.013)			(0.013)	
Experience × COVID		0.088***			0.087***	0.075***
		(0.016)			(0.017)	(0.017)
N. contributors			0.066***		0.056***	
			(0.006)		(0.005)	
N. contributors × COVID			0.011		0.005	0.007
			(0.007)		(0.006)	(0.004)
Message length				0.583***	0.250***	
				(0.056)	(0.059)	
Message length × COVID				0.319***	0.171**	0.144
				(0.073)	(0.075)	(0.081)
Project FE						X
Time FE						X
Pseudo R^2	0.063	0.095	0.105	0.088	0.138	0.520
N	102,720	102,720	102,720	102,696	102,696	102,696
Clusters	4,280	4,280	4,280	4,279	4,279	4,279
Means		Experience	N. contributors	Message length		
D(Loc.=1)		7.152	4.460	3.774		
D(Loc.=2)		7.300	5.179	3.815		
D(Loc.=3)		7.582	6.594	3.932		

Notes: *Experience* is the log average contributor's experience, measured as the number of commits to any project before the coronavirus pandemic. *N. contributors* indicates the total number of pre-COVID contributors to a given project. *Message length* indicates the log average number of characters of pre-COVID commit messages. Co-located teams are the left-out category. All regression control for project age (linear & quadratic) and starting year, number of commits and watchers within a project's first year, country-of-owner fixed effects and month fixed effects. Except for project age and month fixed effects all controls are interacted with the COVID indicator. Column (6) includes project and time fixed effects, project age (linear & quadratic) and the interacted remaining controls. The sample covers 24 months centered around the onset of the coronavirus pandemic in March 2020. Estimation methods: columns (1) to (5): Negative-Binomial ML; column (6): Poisson ML. Standard errors are clustered at the project level and shown in parentheses. ***, **, * indicate significance at 1%, 5% and 10%.

production process, are better in sourcing remote contributors from the international open-source community. This could give them an additional advantage over co-located teams in dealing with the pandemic. Column (3) includes the total number of pre-pandemic contributors to a given project. Distributed teams indeed receive contributions from a larger number of users than co-located teams and the regression results show that projects with more contributors are more active before the pandemic. However, the insignificant interaction suggests that this relationship was not altered by the pandemic, i.e. the "additional" existing contributors of more distributed teams are not the driving force behind their better performance during the pandemic.

Column (4) includes the length of commit messages (the log number of characters). With each code contribution, Git allows the contributor to write a message explaining the changes. We consider this variable a proxy for written (online) communication and documentation on GitHub, which should be more extensive in distributed teams. Mean values at the bottom of column (4) suggest that commit messages of mixed and distributed teams are 4% and 17% longer than those of co-located teams, respectively.[27] The regression results show that longer messages are associated with higher productivity in general.[28] Further, the coefficient of the interaction term with COVID_t is more than half the size of the main effect of message length. This supports the idea that written communication has become more important for productivity during the pandemic.

Column (5) adds all additional explanatory variables at once. The combined regression shows that their individual contributions remain important. At the same time, the interactions between COVID_t and the dummies for the number of distinct locations

[27] These numbers are calculated as follows: $e^{3.815-3.744} \approx 1.04$ for mixed teams and $e^{3.932-3.744} \approx 1.17$ for fully distributed teams.

[28] Note that if some teams simply had a habit of committing more often (i.e. smaller changes per commit), the relationship may be biased in the negative direction, leading to an underestimate of the importance of communication.

remain statistically significant at (at least) the 10% level, with similar magnitudes to the baseline specification in column (1). Column (6) adds project and time fixed effects, which leaves the interaction coefficients of all three additional variables mostly unchanged, but the effect of message length loses its significance. In contrast, the interaction coefficients for mixed and distributed teams increase in magnitude and become statistically more significant than in column (5).

Overall, Table 4 shows that the additional factors are important determinants of team performance before and after the onset of the coronavirus pandemic. Nevertheless, even combined, they cannot (largely) explain the differential impact of the COVID-19 shock on co-located, mixed and distributed teams. Although we cannot entirely rule out other unobserved systematic differences between team types, our results strongly suggest that the proximity of team members plays a key role for the post-pandemic productivity divergence of co-located and distributed teams.

The results in Table 4 suggest that the number of contributors is a key factor for performance before the pandemic. We turn to the number of post-pandemic contributors next. If during the pandemic co-located teams fail to effectively move their production process into the virtual world, it is likely that some team member become entirely inactive due to the increased communication costs. As this problem is less likely to occur in (at least partially) distributed teams, we may observe a higher number of active contributors in these teams compared to co-located teams. In order to explore this channel, we change the dependent variable to *the number of active contributors* in Table A10.[29]

Column (1) shows results for the full sample without any restrictions. Before the

[29] The observational period still covers 24 months centered around the onset of the pandemic, but we sum the total number of pre-pandemic and post-pandemic contributors, respectively. Therefore, the data contains two observations per project, one before and one after the onset of COVID-19. We include the same configuration of interacted controls as in Equation (2) except for month fixed effects. Project age is measured at the beginning of the observational period.

coronavirus pandemic, there was no size difference between co-located and mixed teams, but distributed teams received contributions from a statistically significantly greater number of users. During the pandemic mixed and distributed teams both were statistically significantly larger than co-located teams, which is consistent with our intuition. In column (2) we re-estimate the regression on a sample in which we only keep projects that were active, i.e. received at least one contribution, during the coronavirus pandemic. Excluding teams which became entirely inactive, which was more likely for co-located teams, these estimates resemble an intensive margin. In this sample the interaction coefficient for mixed teams is much smaller and statistically insignificant. Distributed teams are still statistically significantly larger than co-located teams, but the interaction coefficient is roughly cut in half. The main effect of distributed teams is virtually unchanged compared to column (1).[30] Taken together Table A10 suggests that part of the reason for the better performance of distributed teams is that they were able to maintain a higher number of active contributors.

3.3 The effect of COVID-19 on individual performance

The previous subsection has shown that teams with members in multiple locations were more resilient during the pandemic. This raises the question of why remote teams are better able to stay productive and, in particular, which team members drive this result. We, therefore, focus next on the effects of the pandemic on individual productivity, comparing remote workers to those who shared the same location with at least one other team member.

We analyze code contributions of project members who joined a project within the first year of its existence. In order to focus on active collaborations, we only include

[30]Evaluating the coefficients at the pre-pandemic (3) and post-pandemic (2) mean of co-located teams, suggests that before COVID-19 fully distributed teams received commits from $(e^{0.252}-1)*3 = 0.86$ more contributors and this difference increases to $(e^{0.252+0.235}-1)*2 = 1.25$ during the pandemic.

projects to which at least two members have made contributions during the observational period, i.e. 12 months before and after the start of the pandemic.[31] Further, we consider the consistency of a repository's pre-pandemic activity below for an additional sample split. We measure consistency by the number of months with contributions to a project in the 12 month period preceding the pandemic, i.e. from March 2019 to February 2020. In addition to analyzing the full sample we focus on the most consistently active projects, which we define as those projects that received contributions in at least nine months before the onset of COVID-19.[32]

3.3.1 Empirical setup

To identify the differential effect of the pandemic on remote workers, we estimate the following equation:

$$c_{ijt} = \beta_0 + \beta_1 \text{COVID}_t + \beta_2 \text{Remote}_{ij} + \beta_3 \text{Remote}_{ij} \times \text{COVID}_t + X_{it}\delta + \epsilon_{ijt}. \tag{3}$$

The dependent variable is the number of commits to a project i by user j in month t. Remote_{ij} indicates that the worker is in a different location than the other first-year project members and COVID_t is an indicator that is equal to one from March 2020 on and zero before. The interaction of Remote_{ij} and COVID_t thus estimates the differential impact of the pandemic on remote workers compared to co-located workers, the left-out category. Therefore, β_3 is the coefficient of interest. Note, that the effect identified in

[31]As before, we further exclude projects with more than 500 contributions within their first two years to exclude outliers and, in particular, commits by bots.

[32]The choice of which percentage of the most active projects to include is a trade-off between a high activity level, such that changes are more measurable than for (initially) rather inactive projects, and sample size for statistical power. Requiring 10 or more months of activity would keep 9.2% of repositories in the sample, whereas our choice of requiring at least 9 months keeps 11.9%. The results are qualitatively similar if we further restrict the sample to only include projects that received at least one commit each month prior to the pandemic. By contrast, if we include less active projects by only requiring at least 6 months of activity out of the preceding 12 months, the results are no longer significant and more similar to the unrestricted sample.

this analysis can be seen as a combination of individual and team-level (differential) effects. The regressions include project fixed effects and monthly time fixed effects, and we control for linear and quadratic project age. We further control for projects' starting year, the number of watchers and commits in the first year, as well as their interactions with COVID_t.

3.3.2 Results

We again start by showing quarterly effects in an event-study graph. Figure A5 in the Appendix shows the quarterly coefficients separately for all projects in the sample for this analysis and only for those that were consistently active before the pandemic, as described above. The fourth quarter of 2019 is the reference category.

The blue line, indicating results for all teams, does not reveal a strong effect. The coefficients are mostly positive, but close to zero and insignificant. Furthermore, the coefficients during the pandemic are comparable to the small spike in Q3 of 2019, i.e. similar to the usual variation. The red line, by contrast, shows positive and significant effects for those projects that had consistently received contributions before the pandemic. Again, we observe a small spike in Q3 of 2019, but the remaining pre-pandemic coefficients are very close to zero, strongly suggesting an absence of pre-trends. We cautiously interpret these results to suggest that remote workers became relatively more productive as a result of the pandemic compared to co-located members.

The corresponding Poisson regression results are shown in Table 5, in column (1) for the full sample and in column (3) for projects which were consistently active before the pandemic. Columns (2) and (4) focus on the extensive margin for the full and restricted samples, respectively. We replace the dependent variable with an indicator that is one if there was a contribution by the member to the repository in a given month and zero otherwise, estimating linear probability models (using OLS). In all four columns, the

Table 5 *Individual productivity for co-located and remote workers*

Dependent variables:	Full sample		Active pre-COVID	
	Num. commits (1)	Commits (yes/no) (2)	Num. commits (3)	Commits (yes/no) (4)
Remote	−0.062	−0.014	0.140	0.020
	(0.279)	(0.017)	(0.400)	(0.091)
Remote × COVID	0.076	0.002	0.551***	0.045*
	(0.130)	(0.006)	(0.157)	(0.027)
Controls	X	X	X	X
Project FE	X	X	X	X
Time FE	X	X	X	X
Model	Poisson	OLS	Poisson	OLS
(Pseudo) R^2	0.349	0.409	0.205	0.110
N	92,016	108,312	10,872	10,872

Notes: The dependent variable in columns (1) and (3) is the number of commits. In columns (2) and (4), the dependent variable is an indicator whether there was at least one commit. The first two columns include the full sample, while the last two columns restrict the sample to consistently active projects (which had commits in at least 9 out of the 12 months preceding the pandemic). Remote indicates that the worker is in a different location from the two other project members and COVID is an indicator that is equal to one in March 2020 and later. The sample includes 24 months centered around the start of the pandemic. The regressions include project fixed and time fixed effects. All regression additionally control for project age (linear & quadratic) and starting year, number of commits and watchers within a project's first year. Except for project age all controls are interacted with the COVID indicator. Standard errors are clustered at the project level and shown in parentheses. ***, **, * indicate significance at 1%, 5% and 10%.

coefficient of the remote indicator is insignificant. While the sign is negative for the full sample and positive for the restricted sample, overall remote workers seem to be similarly active compared to co-located members before the coronavirus pandemic.

The coefficient of interest on the interaction of Remote_i and COVID_t is positive in columns (3) and (4). The effect in the full sample is insignificant, with a magnitude of slightly less than 8%. Column (3), by contrast, shows a sizable and significant increase of more than 70% in the productivity of remote workers for the projects that were among the most consistently active projects before the pandemic.[33] In this restricted sample, the effect is thus somewhat larger than the general effect on the productivity of distributed teams estimated in Section 3.2. Column (2) shows no significant effect on the intensive margin for the full sample. Again, the corresponding effect for the most active projects in column (4) is larger, implying that remote members were 4.5 percentage points more likely to make at least one contribution to their projects in a given month in the 12 months after the start of the pandemic. The effect in this linear probability model is smaller than the one in column (3) and significant at the 10% level.

Overall, the results show no significant change for the full sample, but a positive effect on remote worker productivity among the most active projects, positively affecting both intensive and extensive margins. Due to the limited sample size, we interpret the results of this section cautiously as suggestive evidence that remote workers may be important drivers of the higher productivity of (partly) distributed teams. This is in line with the idea that communication plays a central role in all of our results and the forced move online may have led to better integration of remote workers into teams' communication.

4 Conclusion

The COVID-19 pandemic has led to disruptive changes in the labor market by forcing a large share of the workforce to work from home. This is expected to entail a persistent increase in remote work of various forms, such as hybrid work or fully distributed teams. We show that even if digitally skilled knowledge workers have a virtual production environment (GitHub) already at their disposal, geographic distance still matters.

[33] This percentage change is calculated from the estimation coefficient β as in the previous section: $(e^{\beta} - 1) * 100 = (e^{0.551} - 1) * 100 \approx 73$ percent.

Hence, when organizations allow for work from home or tap the global talent pool by hiring remote workers, it is crucial to understand what makes distributed teams productive to exploit the full potential of remote work models. We use the exogenous shift in exposure to remote work modes that was caused by the coronavirus pandemic to investigate the determinants of successfully organizing team work across different locations.

Analyzing small teams of open-source programmers, we find that the productivity of co-located teams suffered as a result of the pandemic, while fully distributed teams' performance was remarkably stable. Mixed teams may even have become slightly more productive. Directly comparing post-pandemic productivity patterns of teams with different geographic distributions, suggests that remote work experience made distributed teams more resilient to the pandemic than co-located teams. These results are robust to including a rich set of controls, matching and placebo tests. Exploring mechanisms, we find that while experience and talent sourcing are important for outcomes and systematically vary among (unmatched) types of teams, they do not drive our main results. Digital communication, proxied by commit messages on the platform, is crucial to team productivity and is likely one channel of our main effects. Nevertheless, the entirety of our results suggests that the performance differences between co-located and distributed teams are not (largely) driven by pre-existing differences or selection. Finally, focusing on individual productivity, we find suggestive evidence that remote workers have become relatively more productive when everyone had to work online. This is in line with the idea that this work mode helped them become better integrated into communication, instead of missing out on water cooler talks and other informal discussions offline. Nevertheless, our suggestive results can only provide a starting point and it remains for future research to identify the best ways to ensure that these remote workers are not excluded from communication.

Taking advantage of fine-grained, real-time collaboration data, our study sheds light on which aspects matter for successful remote collaboration in teams. While co-located teams suffered during the pandemic, an optimistic view of our results implies that learning from teams with remote work experience and setting up effective online communication can help mitigate negative effects of the shift to remote work. Future research should investigate further tools and organizational practices to enhance the productivity of remote teams. A better understanding of how to keep both co-located and remote employees productive will help organizations strive in a future of increasingly flexible work modes.

References

Abadie, Alberto (Jan. 2005). "Semiparametric Difference-in-Differences Estimators". *The Review of Economic Studies* 72(1), pp. 1–19.

Aksoy, Cevat Giray, Jose Maria Barrero, Nicholas Bloom, Steven J Davis, Mathias Dolls, and Pablo Zarate (2022). *Working from home around the world*. Tech. rep. National Bureau of Economic Research.

Bahar, Dany, Prithwiraj Choudhury, Do Yoon Kim, and Wesley W Koo (2022). "Innovation on Wings: Nonstop Flights and Firm Innovation in the Global Context". *Forthcoming in Management Science*.

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.

Bloom, Nicholas, Ruobing Han, and James Liang (2022). *How hybrid working from home works out*. Tech. rep. National Bureau of Economic Research.

Bojinov, Iavor, Prithwiraj Choudhury, and Jacqueline N Lane (2021). "Virtual watercoolers: A field experiment on virtual synchronous interactions and performance of organizational newcomers". *Harvard Business School Technology & Operations Mgt. Unit Working Paper* (21-125).

Callaway, Brantly and Pedro H.C. Sant'Anna (2021). "Difference-in-Differences with multiple time periods". *Journal of Econometrics* 225(2). Themed Issue: Treatment Effect 1, pp. 200–230.

Daw, Jamie R. and Laura A. Hatfield (2018). “Matching and Regression to the Mean in Difference-in-Differences Analysis”. *Health Services Research* 53(6), pp. 4138–4156.

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.

Emanuel, N, E Harrington, and A Pallais (2022). *The Power of Proximity: Office Interactions Affect Online Feedback and Quits, Especially for Women and Young Worker*. Tech. rep. Mimeo.

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.

Hansen, Ben B. and Stephanie Olsen Klopfer (2006). “Optimal Full Matching and Related Designs via Network Flows”. *Journal of Computational and Graphical Statistics* 15(3), pp. 609–627.

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 (June 2015). “Social preferences in the online laboratory: a randomized experiment”. *Experimental Economics* 18(2), pp. 251–283.

Horton, John, William R Kerr, and Christopher Stanton (2017). “Digital labor markets and global talent flows”. *High-skilled migration to the United States and its economic consequences*. University of Chicago Press, pp. 71–108.

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

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 (Aug. 1993). “Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations”. *The Quarterly Journal of Economics* 108(3), pp. 577–98.

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

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

Lu, Xuan, Wei Ai, Yixin Wang, and Qiaozhu Mei (2023). “Team Resilience under Shock: An Empirical Analysis of GitHub Repositories during Early COVID-19 Pandemic”. *arXiv preprint arXiv:2301.12326*.

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*.

Myers, Kyle R, Wei Yang Tham, Yian Yin, Nina Cohodes, Jerry G Thursby, Marie C Thursby, Peter Schiffer, Joseph T Walsh, Karim R Lakhani, and Dashun Wang (2020). "Unequal effects of the COVID-19 pandemic on scientists". *Nature human behaviour* 4(9), pp. 880–883.

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

Thompson, Peter and Melanie Fox-Kean (Mar. 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".

Yang, Longqi, David Holtz, Sonia Jaffe, Siddharth Suri, Shilpi Sinha, Jeffrey Weston, Connor Joyce, Neha Shah, Kevin Sherman, Brent Hecht, et al. (2022). "The effects of remote work on collaboration among information workers". *Nature human behaviour* 6(1), pp. 43–54.

Appendix

A Combining GitHub Torrent and GitHub Archive

We use the latest (March 2021) snapshot from GitHub Torrent (GHT) (Gousios 2013) and the GitHub Archive (GHA) dataset. Both GHT and GHA provide a mirror of the GitHub public event stream from 2012 on. Our main data source is GHT, from which we derive information about projects (repositories) and project members, including their location data. GHT also features data on code contributions (commits), however, commit data is missing for the period between June 2019 and December 2019 as well as after March 2021 (see Figure A1). Therefore, we utilize GHA to get commit data with full coverage for the period used in our analysis.

Matching project data is straightforward and can be achieved using unique project-owner-name combinations. Matching user data and their contributions is more arduous. A straightforward way to match contributions is to match based on push events.[34] However, there are two major caveats with this method. First, push events can contain multiple commits, potentially from multiple users, and thus commits could be attributed to the wrong user.[35] Second, a single commit can appear in multiple push events, e.g. if code is merged from one project or forked to another.[36] We want to ensure that we count each commit only once and associate it with the user and project it originates from. Consequently, we prefer matching at the commit level rather than the push level. This, however, is more complicated because GHT associates code commits with GitHub user accounts while GHA associates contributions with Git usernames,

[34] A commit records changes to the (local) code repository. It is made using the Git version control system. A push event then sends one or several commits to the remote repository on GitHub.

[35] We do not consider this a problem in Section 2.2 because we want to allow for commits to be ”exported” to multiple destinations. In Section 3, however, we measure teams’ (team members’) productivity, where it is more crucial to attribute code contributions to their original author correctly.

[36] Forks are new repositories that share the code of an original “upstream” repository. They are created by contributors who do not have direct access to (i.e. are not members of) the original project.

which are distinct from each other.[37]

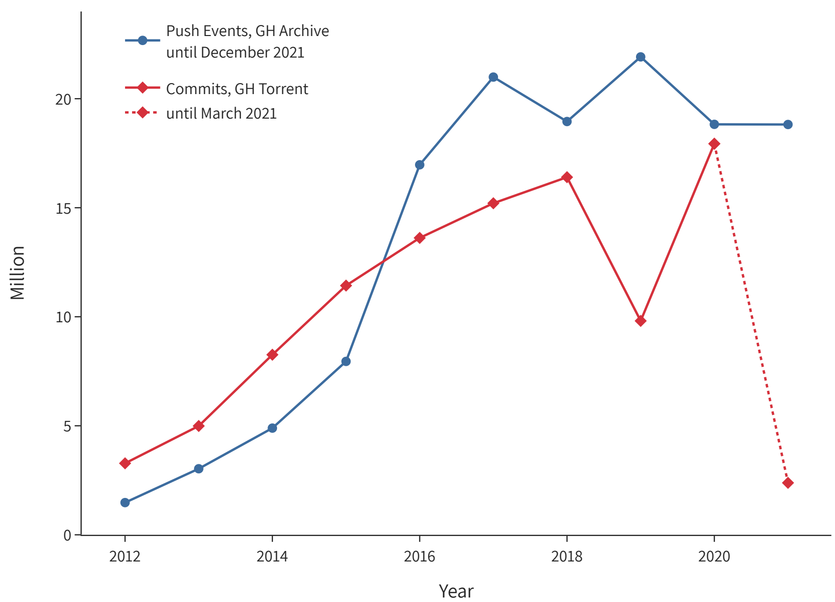
In order to identify individual code contributions Git assigns a unique hash value to every commit. These hash values are featured in both datasets and thus allow us to directly match commits in the overlapping period from January 2015 to May 2019 and January 2020 to March 2021. Using this method we are able to obtain Git-GitHub user pairs. If a GitHub account is matched to multiple Git usernames we attribute all contributions from these usernames to the GitHub account. If a Git username is matched with multiple GitHub accounts, we choose the most frequent combination and attribute all contributions to the respective GitHub account. Using these matches, we replace missing GitHub account information of commits recorded in the second half of 2019. To further improve our data coverage, we fill in the remaining missing values based on push events as described above.[38] Overall, we are able to replace missing GHT user data for 97% of the GHA commits between June 2019 and Dec 2019.[39]

[37]Git is a locally installed version control system that helps manage source code and keep track of code history. It can be used independently of GitHub. GitHub is a cloud-based platform that is used to host, manage and collaborate on Git repositories.

[38]This improves data coverage by nine percentage points.

[39]The remaining 3% of commits are from users who only contributed in the second half of 2019, but neither before nor afterward. Therefore, these users are of little relevance to our analysis.

Figure A1 *Number of contributions on GitHub*



Notes: This figure shows commits from users with reported locations and to repositories owned by users with reported locations. The drop in the number of commits in the GHT dataset in 2019 is due to missing data from June 2019 until December 2019.

B Estimation of the gravity equation with GitHub Data

Our baseline specification for the gravity model is the following:

$$c_{ijt} = \beta_0 D_{ij}^{\beta_1} \{X\}^{\{\gamma\}} \tau_{c_i t} \tau_{c_j t} \epsilon_{ijt} \tag{4}$$

where

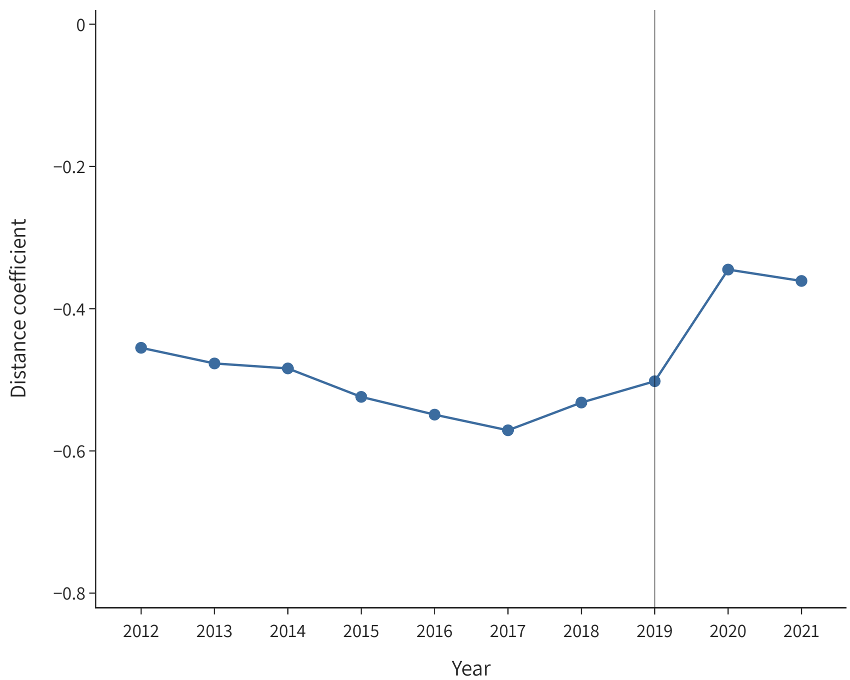
c_{ijt} — number of collaborations between a city pair ij in a year t. It is measured by the number of contributions by users from a city i submitted to a project owned by users from a city j. In our setting, direction matters: collaborations between city pairs ij and ji are treated as two different observations. To make an analogy to the trade and migration literature, we think of the city of a committer as the origin (e.g. origin of a service provider — an exporter) and the city of a project owner as the destination (e.g. destination of services — an importer). Another analogy to make is a citation of a patent or a scientific article, where a cited contribution could be considered as an origin of knowledge and a citing patent (article) — a destination of the knowledge transfer.

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 (in case of cross-border collaborations). Conditional on geographic distance, these dummies capture the effects of state borders and language barriers. In line with the trade literature, we also control for 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

Figure A2 *Distance elasticity of collaborations on GitHub in 2012–21*



Notes: This figure shows the yearly distance elasticity of code contributions. The coefficients are obtained from estimating regressions equivalent to that of Table 1 column (1), separately for each year.

fixed effects.[40] We conservatively cluster standard errors at the country-pair level.

For most regressions, we use Poisson pseudo-maximum likelihood estimation with multi-way fixed effects, which is consistent for models with count data in the presence of many zero outcomes. We estimate the effects first jointly and then separately at the intensive and extensive margins. For the latter, we use a linear probability model.

To benchmark the effect of geographic barriers on GitHub to their role in shaping modern trade flows, we use 2012–19 cross-country export data from the CEPII Gravity dataset, Version 202102, and aggregate GitHub data at the country-pair level for the same period.

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

Table A1 *Country-level gravity regressions: benchmark to trade*

Variables	Tradeflows (1)	Tradeflows > 0 (2)	Contributions (3)	Contributions > 0 (4)
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	37,419	31,922	30,797	5283

Notes: The dependent variables are trade flows (exports from origin to destination) in columns (1) and (2) as well as GitHub contributions (from origin to destination) in columns (3) and (4) between a given country pair. Columns (2) and (4) limit the sample to pairs with non-zero trade flows/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.

Table A2 *City-level gravity regressions accounting for programming language*

Variables	Contributions (1)	Contributions >0 (2)	Contributions yes/no (3)	Contributions distance dummies (4)
Distance	−0.420***		−0.282***	
	(0.037)		(0.027)	
1–50km		−0.267		−0.009
		(0.357)		(0.270)
50–100km		−1.963***		−1.325***
		(0.334)		(0.279)
100–300km		−2.399***		−1.645***
		(0.252)		(0.191)
300–700km		−2.655***		−1.776***
		(0.284)		(0.186)
>700km		−3.038***		−2.035***
		(0.310)		(0.227)
Foreign country	−1.353***	−1.692***	−0.864***	−1.069***
	(0.233)	(0.170)	(0.157)	(0.108)
Common language	0.322***	0.280***	0.176**	0.127*
	(0.102)	(0.076)	(0.080)	(0.070)
Users, owner	0.161***	0.176***	0.093**	0.107***
	(0.031)	(0.031)	(0.043)	(0.040)
Users, committer	0.733***	0.744***	0.551***	0.548***
	(0.030)	(0.031)	(0.031)	(0.031)
Remoteness, committer	−0.011	−0.083	0.023**	0.023**
	(0.126)	(0.115)	(0.011)	(0.010)
Remoteness, owner	2.666***	2.622***	0.393***	0.378***
	(0.156)	(0.179)	(0.072)	(0.077)
Observations	307,718	307,718	610,522	610,522
Clusters	4487	4487	4075	4075
Same prog. language	No	No	Yes	Yes

Notes: The dependent variable is contributions between a city pair (the sample includes observations with non-zero contributions). All specifications include country of committers and country of project owners time-specific fixed effects. The specification in columns (3) and (4) adds programming language time-specific effects. Standard errors are clustered at a country pair level. Estimation method: PPML.

C Additional figures and graphs for the team-level analysis

Table A3 *Team-level data: summary statistics*

	Control			Treatment		
	Co-located (1)	Mixed (2)	Distributed (3)	Co-located (4)	Mixed (5)	Distributed (6)
Avg. mtly. commits						
pre-pandemic (hyp.)	2.52	2.44	2.59	2.25	2.28	3.02
	(4.10)	(4.06)	(4.34)	(3.71)	(3.81)	(4.68)
post-pandemic (hyp.)	1.18	1.08	1.39	0.74	1.21	1.82
	(3.41)	(3.54)	(5.47)	(2.36)	(3.92)	(7.68)
Num. commits (first year)	90.36	80.59	78.75	85.46	73.83	82.03
	(91.00)	(87.89)	(86.51)	(84.27)	(82.71)	(84.44)
Num. watchers (frist year)	8.32	24.50	42.21	7.95	23.55	37.15
	(52.28)	(189.49)	(374.22)	(69.07)	(272.72)	(269.65)
Project age	7.42	7.82	7.73	7.28	7.68	8.14
	(3.45)	(3.46)	(3.45)	(3.42)	(3.56)	(3.55)
Project year						
2015	0.29	0.29	0.32	0.00	0.00	0.00
2016	0.33	0.38	0.37	0.00	0.00	0.00
2017	0.38	0.33	0.31	0.00	0.00	0.00
2018	0.00	0.00	0.00	1.00	1.00	1.00
Owner country						
USA	0.23	0.30	0.44	0.21	0.31	0.42
Brazil	0.03	0.04	0.04	0.03	0.05	0.04
India	0.05	0.04	0.03	0.07	0.04	0.03
Great Britain	0.08	0.08	0.06	0.07	0.07	0.07
Canada	0.02	0.04	0.04	0.04	0.03	0.03
Other	0.60	0.51	0.40	0.58	0.50	0.41
Num. projects	533	2212	3695	254	627	762

Notes: This table shows mean values and standard errors in parentheses of control and dependent variables. Treatment group projects are those started in 2018. The control group are earlier projects with the hypothetical onset of the pandemic shifted backwards. For example, the hypothetical onset of the pandemic is set to March 2019 for projects started in 2017. Columns (1) to (3) show control group summary statistics for co-located, mixed and distributed teams, respectively. Columns (4) to (6) show treatment group summary statistics for co-located, mixed and distributed teams, respectively. The number of commits is first averaged within teams and then across teams. Project age is measured at the start of the observational period, i.e. 12 months before the (hypothetical) onset of the coronavirus pandemic. The sample includes projects which received at least on contribution in the observational period of 24 months centered around the (hypothetical) onset of the coronavirus pandemic.

Table A4 *Team-level data: summary statistics*

	Co-located (1)	Mixed (2)	Distributed (3)
Avg. mtly. commits			
pre-pandemic	2.30	2.24	2.63
	(4.15)	(5.39)	(5.63)
post-pandemic	1.11	1.57	2.27
	(3.41)	(5.53)	(10.42)
Num. contributors			
pre-pandemic	3.01	3.16	3.48
	(4.00)	(6.36)	(8.66)
post-pandemic	0.18	0.28	0.54
	(2.28)	(3.93)	(8.14)
Num. commits (first year)	88.29	80.35	83.18
	(86.56)	(87.84)	(86.80)
Num. watchers (frist year)	11.50	32.88	45.32
	(74.77)	(260.46)	(260.47)
Log experience (pre-pandemic)	7.15	7.30	7.58
	(2.47)	(2.49)	(2.30)
Log message length (pre-pandemic)	3.75	3.79	3.91
	(0.60)	(0.58)	(0.60)
Project age	17.06	20.47	22.68
	(12.90)	(13.25)	(13.10)
Project year			
2015	0.09	0.11	0.15
2016	0.16	0.22	0.24
2017	0.21	0.25	0.28
2018	0.54	0.42	0.33
Owner country			
USA	0.22	0.32	0.44
Brazil	0.02	0.04	0.03
India	0.06	0.03	0.02
Great Britain	0.07	0.08	0.06
Canada	0.03	0.03	0.04
Other	0.60	0.50	0.41
Num. projects	467	1499	2313

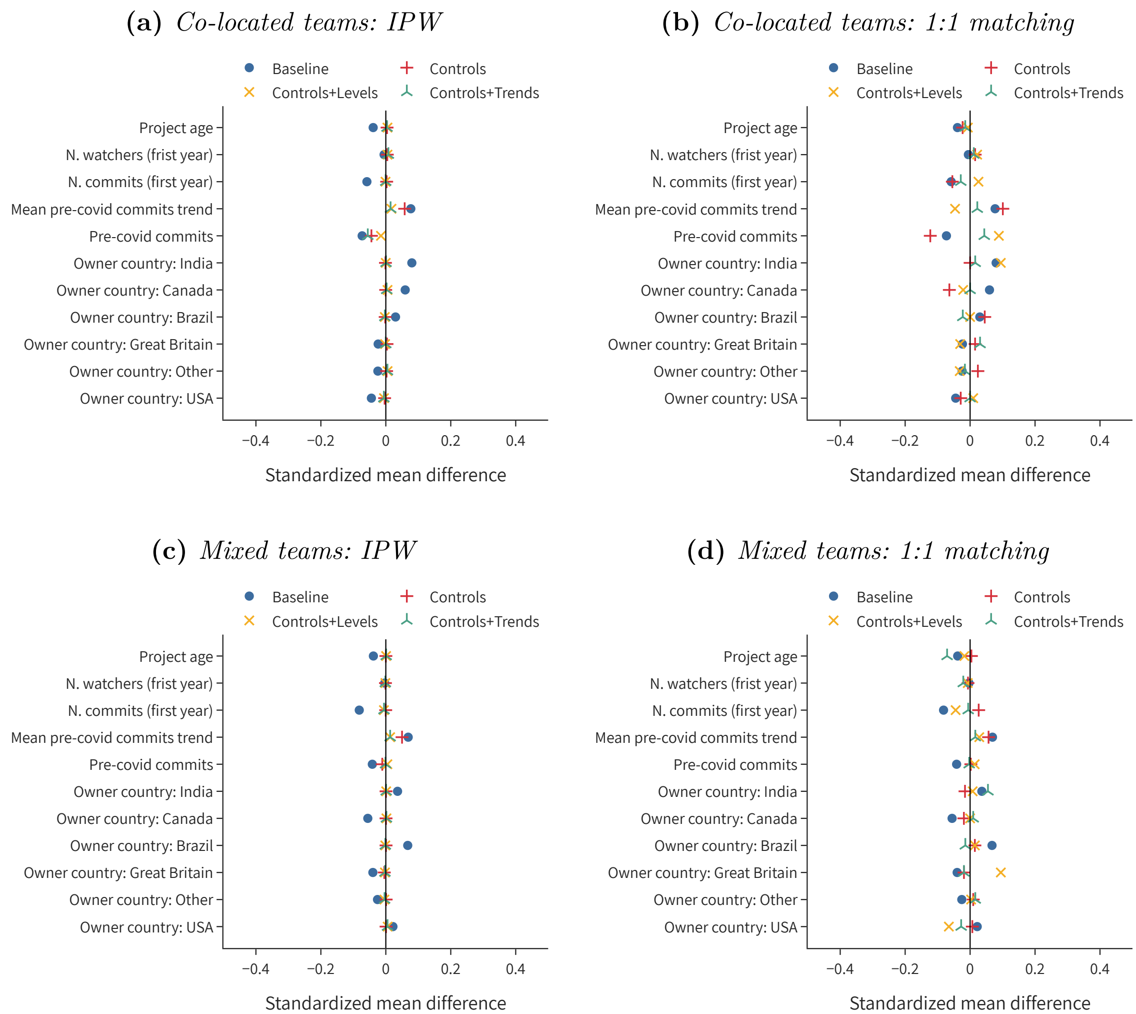
Notes: This table shows mean values and standard errors in parentheses of control and dependent variables. Columns (1), (2) and (3) show summary statistics for co-located, mixed and distributed teams, respectively. The number of commits is first averaged within teams and then across teams. Project age is measured at the start of the observational period, i.e. 12 months before the onset of the coronavirus pandemic. The sample includes projects which received at least on contribution in the observational period of 24 months centered around the onset of the coronavirus pandemic in March 2020.

Table A5 *COVID effects of team distribution (Project×Month FE)*

Team composition	Dependent: Number of commits			
	Co-located	Mixed	Distributed	Full sample
	(1)	(2)	(3)	(4)
COVID × Treated	−0.440**	0.175	0.108	−0.426**
	(0.215)	(0.132)	(0.155)	(0.213)
COVID × D(Loc.=2)				−0.082
				(0.135)
COVID × D(Loc.=3)				0.092
				(0.138)
COVID × Treated × D(Loc.=2)				0.585**
				(0.251)
COVID × Treated × D(Loc.=3)				0.531**
				(0.267)
Controls	X	X	X	X
Project FE	X	X	X	X
Time FE	X	X	X	X
Pseudo R^2	0.434	0.440	0.448	0.443
N	18,888	68,136	106,968	193,992
Clusters	787	2,839	4,457	8,083

Notes: Columns (1), (2) and (3) split the sample into co-located, mixed and distributed teams, respectively. *Treated* projects are founded in 2018 and are exposed to 12 months of COVID-19. The control group are earlier projects with the hypothetical onset of the pandemic shifted backwards. For example, the hypothetical onset of the pandemic is set to March 2019 for projects started in 2017. All regressions include project and time fixed effects. Further controls include project age (linear & quadratic) as well as interactions of the number of commits and watchers within a project's first year and country-of-owner fixed effects with the (hypothetical) COVID indicator. The sample covers 24 months centered around the hypothetical onset of the coronavirus pandemic. Estimation method: Poisson ML. Standard errors are clustered at the project level and shown in parentheses. ***, **, * indicate significance at 1%, 5% and 10%.

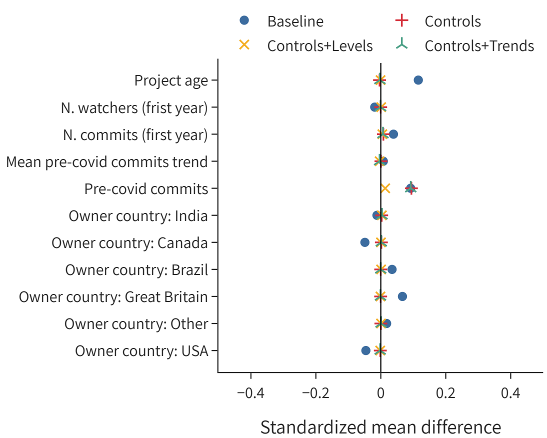
Figure A3 *Artificial control teams: balance of covariates*



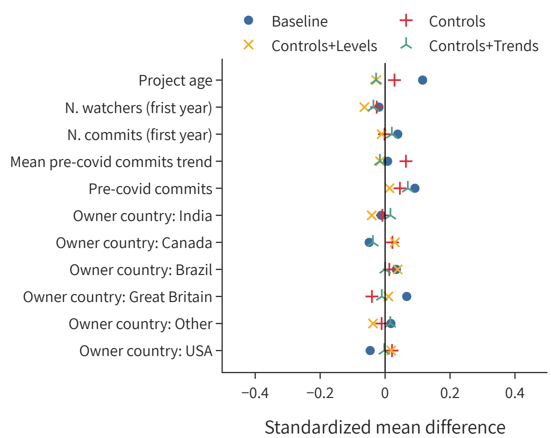
Notes: The figures shows standardized mean differences of control variables between projects started in 2018 (treatment) and earlier projects (control). The differences are calculated by subtracting the means of treated teams from those of control teams and dividing by the standard deviation of treated teams. In panels (a) and (c) teams are balanced using the inverse probability weighting approach suggested in Abadie (2005) and in panels (b) and (d) we apply 1:1 propensity score matching (Hansen and Klopfer 2006). Propensity scores are estimated using logistic regressions of a treatment indicator on either just controls, quarterly pre-pandemic commit levels and controls or quarterly pre-pandemic commit trends and controls. Controls include project age, number of commits and watchers within a project's first year and project-owner-country fixed effects. Pre-covid commits is the sum of commits a project has received in the 12 months before the onset of the (hypothetical) coronavirus pandemic. Mean pre-covid commits trend measures the project-mean of quarterly first-differences in the number of commits before the onset of the (hypothetical) coronavirus pandemic.

Figure A3 *Artificial control teams: balance of covariates (cont'd)*

(e) *Fully distributed teams: IPW*



(f) *Fully distributed teams: 1:1 matching*



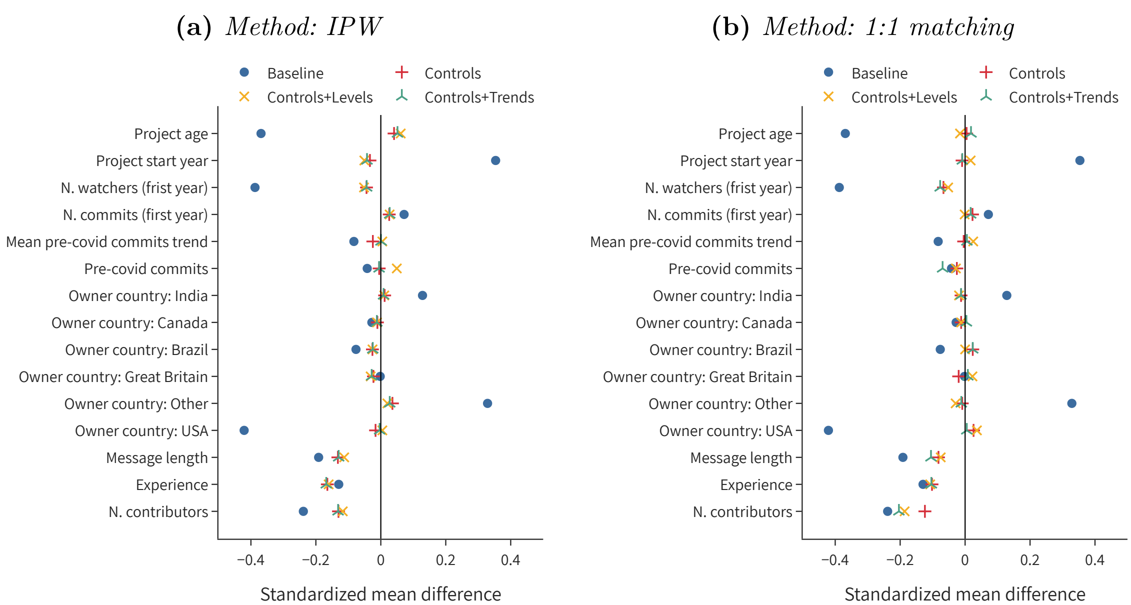
Notes: The figures shows standardized mean differences of control variables between projects started in 2018 (treatment) and earlier projects (control). The differences are calculated by subtracting the means of treated teams from those of control teams and dividing by the standard deviation of treated teams. In panel (e) teams are balanced using the inverse probability weighting approach suggested in Abadie (2005) and in panel (f) we apply 1:1 propensity score matching (Hansen and Klopfer 2006). Propensity scores are estimated using logistic regressions of a treatment indicator on either just controls, quarterly pre-pandemic commit levels and controls or quarterly pre-pandemic commit trends and controls. Controls include project age, number of commits and watchers within a project's first year and project-owner-country fixed effects. Pre-covid commits is the sum of commits a project has received in the 12 months before the onset of the (hypothetical) coronavirus pandemic. Mean pre-covid commits trend measures the project-mean of quarterly first-differences in the number of commits before the onset of the (hypothetical) coronavirus pandemic.

Table A6 *COVID effects of team distribution (matched)*

Method	Dependent: Number of commits					
	Inverse Probability Weighting			1:1 Matching		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A *Co-located teams*						
Treated	−0.049	−0.018	−0.080	−0.065	0.066	0.073
	(0.114)	(0.113)	(0.112)	(0.122)	(0.135)	(0.127)
Treated × COVID	−0.440**	−0.448**	−0.457**	−0.548**	−0.606***	−0.486**
	(0.197)	(0.198)	(0.197)	(0.229)	(0.220)	(0.248)
Pseudo R2	0.125	0.138	0.123	0.111	0.124	0.125
N	18,888	12,192	18,888	12,192	18,888	12,192
Clusters	787	508	787	508	787	508
Panel B *Mixed teams*						
Treated	0.011	0.017	0.008	−0.014	0.053	0.034
	(0.069)	(0.079)	(0.078)	(0.086)	(0.085)	(0.087)
Treated × COVID	0.143	0.060	0.063	0.265*	0.263*	0.085
	(0.127)	(0.123)	(0.123)	(0.151)	(0.148)	(0.149)
Pseudo R2	0.113	0.117	0.117	0.117	0.116	0.114
N	68,136	68,136	68,136	30,096	30,096	30,096
Clusters	2,839	2,839	2,839	1,254	1,254	1,254
Panel C *Fully distributed teams*						
Treated	0.162***	0.033	0.162***	0.119	0.070	0.057
	(0.058)	(0.054)	(0.060)	(0.081)	(0.075)	(0.079)
Treated × COVID	0.006	0.022	0.012	0.170	0.163	0.113
	(0.122)	(0.121)	(0.122)	(0.132)	(0.131)	(0.130)
Pseudo R2	0.097	0.095	0.097	0.105	0.110	0.088
N	106,968	106,968	106,968	36,576	36,576	36,576
Clusters	4,457	4,457	4,457	1,524	1,524	1,524
Matching variables						
Controls	X	X	X	X	X	X
Pre-treatment outcome levels		X			X	
Pre-treatment outcome trends			X			X

Notes: Columns (1) to (3) estimate weighted regressions using propensity scores and the inverse probability weighting method of Abadie (2005). Columns (4) to (6) estimate regressions on 1:1 matched samples using propensity score matching (Hansen and Klopfer 2006). Pre-treatment outcome levels are measured as quarterly number of commits. Pre-treatment outcome trends are measured as first-differences of quarterly number of commits. Controls include project age (at the beginning of the observational period), number of commits and watchers within a project's first year and country-of-owner fixed effects. Propensity scores are estimated using logistic regression. All regression control for project age (linear & quadratic), month fixed effects and the full set of interacted remaining controls. The sample covers 24 months centered around the hypothetical onset of the coronavirus pandemic. Estimation method: Negative-Binomial ML. Standard errors are clustered at the project level and shown in parentheses. ***, **, * indicate significance at 1%, 5% and 10%.

Figure A4 *Co-located and distributed teams: balance of covariates*



Notes: The figures show standardized mean differences of control variables between co-located and distributed teams. The differences are calculated by subtracting the means of distributed teams from those of co-located teams and dividing by the standard deviation of co-located teams. In panel (a) teams are balanced using the inverse probability weighting approach suggested in Abadie (2005) and in panel (b) we apply 1:1 propensity score matching using the optimized approach of Hansen and Klopfer (2006). Propensity scores are estimated using logistic regressions of an indicator for co-location on project age and starting year, number of commits and watchers within a project's first year and project-owner-country fixed effects. Pre-COVID commits is the sum of commits a project has received in the 12 months before March 2020. Mean pre-COVID commits trend measures the project-mean of quarterly first-differences in the number of commits.

Table A7 *Collaborations by team distribution (matched)*

Method	Dependent: Number of commits						
		Inverse probability weighting			1:1 Matching		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Distributed	0.084	0.037	−0.055	0.045	−0.001	0.015	0.109
	(0.093)	(0.109)	(0.116)	(0.110)	(0.097)	(0.095)	(0.095)
Distributed × COVID	0.372***	0.381***	0.393***	0.369***	0.358***	0.319**	0.354**
	(0.129)	(0.129)	(0.136)	(0.134)	(0.134)	(0.135)	(0.139)
Pseudo R^2	0.061	0.065	0.065	0.065	0.079	0.083	0.083
N	102,696	102,696	102,696	102,696	44,832	44,832	44,832
Clusters	4,279	4,279	4,279	4,279	1,868	1,868	1,868
Matching variables							
Controls		X	X	X	X	X	X
Pre-COVID outcome levels			X			X	
Pre-COVID outcome trends				X			X

Notes: Columns (1) shows a baseline regression without any matching applied. Columns (2) to (4) estimate weighted regressions using propensity scores and the inverse probability weighting method of Abadie (2005). Columns (5) to (7) estimate regressions on 1:1 matched samples using propensity score matching (Hansen and Klopfer 2006). Distributed is a dummy variable that is 1 for mixed and fully distributed teams and 0 for co-located teams. Pre-treatment outcome levels are measured as quarterly number of commits. Pre-treatment outcome trends are measured as first-differences of quarterly number of commits. Controls include project starting year and age (at the beginning of the observational period), number of commits and watchers within a project's first year and country-of-owner fixed effects. Propensity scores are estimated using logistic regression. All regression control for project age (linear & quadratic), month fixed effects and the full set of interacted remaining controls. The sample covers 24 months centered around the onset of the coronavirus pandemic in March 2020. Estimation method: Negative-Binomial ML. Standard errors are clustered at the project level and shown in parentheses. ***, **, * indicate significance at 1%, 5% and 10%.

Table A8 *Collaborations by team distribution (placebo)*

	Dependent: Number of commits			
	(1)	(2)	(3)	(4)
D(Locations=2)	−0.054	−0.034	−0.002	−0.001
	(0.113)	(0.119)	(0.119)	(0.107)
D(Locations=3)	0.072	0.126	0.052	0.120
	(0.109)	(0.120)	(0.122)	(0.123)
D(Locations=2) × COVID	0.015	−0.047	−0.131	−0.160
	(0.155)	(0.144)	(0.206)	(0.205)
D(Locations=3) × COVID	0.033	−0.038	−0.021	−0.150
	(0.137)	(0.135)	(0.199)	(0.202)
Placebo COVID year	2019	2019	2018	2018
Controls		X		X
Pseudo R^2	0.008	0.064	0.021	0.086
N	99,312	99,312	83,376	83,376
Clusters	4,138	4,138	3,474	3,474

Notes: The sample in columns (1) and (2) includes projects started in 2017 or earlier and the placebo onset of the coronavirus pandemic is set to March 2019. The sample in columns (1) and (2) includes projects started in 2016 or earlier and the placebo onset of the coronavirus pandemic is set to March 2018. Co-located teams are the left-out category. Controls include project age (linear & quadratic) and starting year, number of commits and watchers within a project's first year, country-of-owner fixed effects and month fixed effects. Except for project age and month fixed effects all controls are interacted with the placebo COVID indicator. The sample covers 24 months centered around the placebo onset of the coronavirus pandemic. Estimation method: Negative-Binomial ML. Standard errors are clustered at the project level and shown in parentheses. ***, **, * indicate significance at 1%, 5% and 10%.

Table A9 *Collaborations by team distribution: no time restriction*

	Dependent: Number of commits					
	(1)	(2)	(3)	(4)	(5)	(6)
D(Locations=2)	−0.088***	0.049	0.091**	0.085**	0.096**	
	(0.027)	(0.045)	(0.041)	(0.041)	(0.041)	
D(Locations=3)	−0.067**	0.225***	0.257***	0.248***	0.263***	
	(0.027)	(0.046)	(0.044)	(0.044)	(0.046)	
D(Locations=2) × COVID	0.490***	0.433***	0.421**	0.388**	0.381**	0.481**
	(0.170)	(0.166)	(0.167)	(0.168)	(0.168)	(0.174)
D(Locations=3) × COVID	1.019***	0.867***	0.843***	0.786***	0.733***	0.992***
	(0.176)	(0.179)	(0.168)	(0.159)	(0.157)	(0.176)
Project age		X	X	X	X	X
Project year FE		X	X	X	X	X
N. Commits (first year)			X	X	X	X
N. Watchers (first year)				X	X	X
Owner country					X	X
Month FE		X	X	X	X	
Project FE						X
Time FE						X
Pseudo R^2	0.095	0.308	0.395	0.398	0.400	0.535
N	1,668,117	1,668,117	1,668,117	1,668,117	1,668,117	1,668,117
Clusters	27,601	27,601	27,601	27,601	27,601	27,601

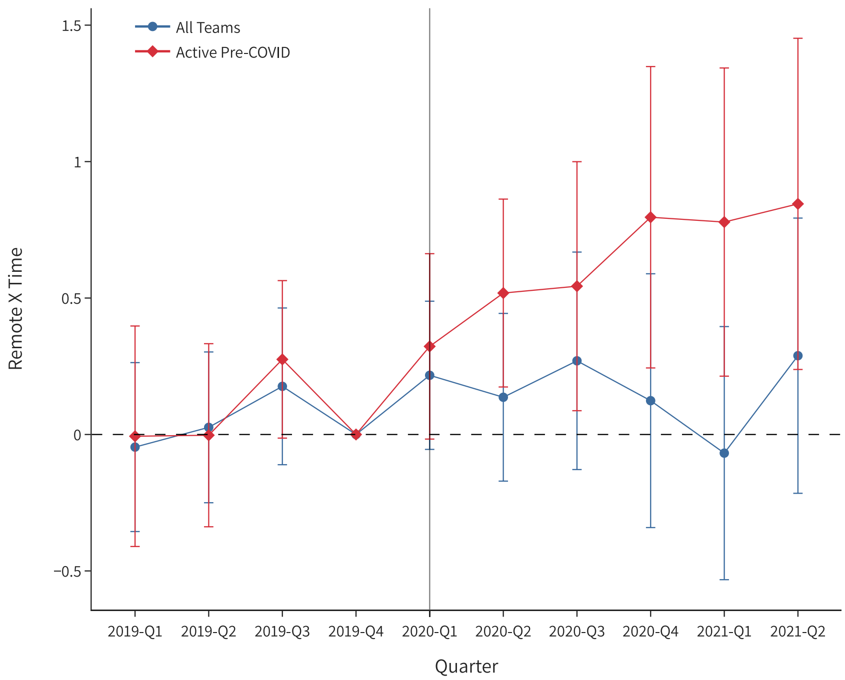
Notes: Co-located teams are the left-out category. All regression control for project age (linear & quadratic) and starting year, number of commits and watchers within a project's first year, country-of-owner fixed effects and month fixed effects. Except for project age and month fixed effects all controls are interacted with the (hypothetical) COVID indicator. Column (6) includes project and time fixed effects, project age (linear & quadratic) and the interacted remaining controls. The sample covers a period from January 2015 to December 2021. Estimation methods: columns (1) to (5): Negative-Binomial ML; column (6): Poisson ML. Standard errors are clustered at the project level and shown in parentheses. ***, **, * indicate significance at 1%, 5% and 10%.

Table A10 *Number of contributors by team distribution*

	Dependent: Number of contributors	
	(1)	(2)
D(Loc.=2)	0.007	0.062
	(0.060)	(0.071)
D(Loc.=3)	0.244***	0.252***
	(0.059)	(0.069)
D(Loc.=2) × COVID	0.301***	0.115
	(0.107)	(0.106)
D(Loc.=3) × COVID	0.409***	0.235**
	(0.109)	(0.106)
Sample	Full	Active during COVID-19
Pseudo R^2	0.034	0.020
N	8,558	3,718

Notes: The dependent variable is the total number of contributors in the 12 months before and the 12 months after the onset of the coronavirus pandemic in March 2020. Column (1) includes all projects without any further restriction. Column (2) only keeps projects which were active during the coronavirus pandemic, i.e. received at least one contribution. Co-located teams are the left-out category. All regression control for project age (at the beginning of the observational period) and starting year, number of commits and watchers within a project's first year and country-of-owner fixed effects. Except for project age all controls are interacted with the COVID indicator. Estimation method: Negative-Binomial ML. Standard errors are clustered at the project level and shown in parentheses. ***, **, * indicate significance at 1%, 5% and 10%.

Figure A5 *Remote workers: quarterly effects*



Notes: This figure shows interaction coefficients and 95% confidence intervals of quarterly time dummies and indicators for remote workers. The dependent variable is the number of commits and controls include project age (linear & quadratic) and starting year (interacted with remote). The blue line includes the full sample. The red line limits the sample in the regression to repositories which had commits in at least 9 out of the 12 months preceding the pandemic, approximately the 10% most consistently active projects. Estimation method: Negative-Binomial ML.