Teamwork and the Homophily Trap: Evidence from Open Source Software[*]

Davidson Heath
David Eccles School of Business
University of Utah

Nathan Seegert
David Eccles School of Business
University of Utah

Jeffrey Yang
David Eccles School of Business
University of Utah

February 11, 2024

Abstract

We investigate team diversity and project success in the setting of open source software. We find that team diversity is strikingly low compared to the contributor pool, with the modal team at the lowest level of diversity i.e., monoculture. To identify the causal effects of team diversity, we use teams' exposure to country-by-year variation in Internet access. The evidence is consistent with homophily, the preference to associate with similar people, leading to teams being trapped in low-diversity states. Teams that escape the "homophily trap" go on to add more diversity and increase their productivity and project success relative to teams that do not.

Keywords: Open source, commons-based production, teams, diversity, productivity, homophily

JEL Classification: J16, J17, L17, L86

[*] Email: davidson.heath@eccles.utah.edu. We thank Simona Abis, Diego Garcia, Jill Grennan, David Matsa, Randall Morck, Scott Page, Yihui Pan, Paola Sapienza, Uta Schönberg, Emil Siriwardane, Geoffrey Tate, and seminar participants at Alberta, Boulder, HK PolyU, HKU, LSU and Utah for comments.

1 Introduction

Collaborative teams are increasingly important in many sectors of the economy including science and innovation (Bloom, Jones, Van Reenen, & Webb, 2020; Jones, 2021). However, there is limited evidence on how teams form, specifically with regard to team diversity, and the consequent impact on productivity. While diverse teams offer benefits such as a broader pool of innovative ideas (Hong & Page, 2004; Yang, Tian, Woodruff, Jones, & Uzzi, 2022), they also face higher costs of coordinating (Becker & Murphy, 1992; Hjort, 2014; Cornelissen, Dustmann, & Schönberg, 2017). A fundamental question for organizational design and policy is whether teams form in a way that maximizes the net benefits of diversity.

This paper studies team diversity and its effects on productivity, using a dataset spanning over a decade and involving more than 50,000 teams from open-source software (OSS) projects.[1] We focus on OSS projects for several reasons: The importance of OSS in the modern economy, the prevalence of multi-country collaboration, and the unparalleled transparency and scope of the data.[2] We document that collaboration is widespread and most OSS teams have contributors from multiple countries. Yet on average teams are much less diverse than the contributor pool, and the most common level of team diversity is zero – i.e., monoculture.

We document evidence consistent with homophily, the natural preference of individuals to associate with similar peers (Becker, 1957; Boisjoly, Duncan, Kremer, Levy, & Eccles, 2006; Ductor, Goyal, & Prummer, 2021; Cullen & Perez-Truglia, 2023). In contrast to other costs of team diversity such as disagreement and miscommunication, homophily is a private cost that is decreasing in the team's existing level of diversity. That is, the less diverse a given team is, the more costly it is for an outsider to join it. As a result, teams can find themselves "trapped" in inefficient low-diversity states.

We find that homophily is pervasive across contributors wanting to join projects and has direct effects on team formation. If coders' identities were irrelevant to which project they join, then outsiders (coders whose joining would increase diversity) would mechanically be more likely to join

[1] Prior economic studies of open source software (Lerner & Tirole, 2002, 2005; Athey & Ellison, 2014; Ytsma & Gallus, 2019) have focused on what motivates coders to contribute to OSS such as altruism, career concerns and public recognition.

[2] OSS is the dominant form of software in supercomputing, data science, cloud computing, and machine learning among others (Nagle, 2019).

low-diversity projects. Instead, we find that teams with low levels of diversity are less likely to attract outsiders. This pattern is unlikely to be due to discrimination because the rate at which outsiders' code is accepted is slightly higher, not lower, for low-diversity projects. This pattern is also unlikely to be due to a lack of familiarity because we find the same negative relationship between team diversity and outsider join rates among coders who express prior knowledge of and interest in the project by starring it. This pattern, however, is consistent with homophily because the private costs of homophily are the highest for low-diversity teams consistent with the effects being largest for low-diversity teams.

The existing literature on diversity and team production suggests an alternative explanation. Specifically, OSS teams may be much less diverse than the contributor pool because diversity is irrelevant or a net negative for team performance. For example, in a model with costs and benefits of diversity but without homophily, Lazear (1999) predicts that teams will be lopsided, that is, dominated by one culture. This "efficient teams" hypothesis predicts that increasing team diversity will lead to unchanged or worse project outcomes. By contrast, the homophily-trap hypothesis predicts that teams are inefficient in their low levels of diversity, so increasing team diversity would lead to better project outcomes. The key test to distinguish these hypotheses is how project success changes with team diversity, in particular for teams at low levels of diversity.

We measure project success in three ways: Survival, development activity, and popularity with users. Unconditionally, we find a clear positive relationship between team diversity and all three measures. This finding is present across a variety of specifications and causal estimates. First, in the cross-section, team diversity is positively correlated with all three outcomes. Second, when we control for possible confounds using panel regressions with a variety of project-by-year controls and fixed effects, we again find strong positive associations. Third, we instrument for variation in team diversity using country-by-year variation in broadband access. Our approach is similar to that of Figlio, Giuliano, Marchingiglio, Ozek, & Sapienza (2023), who examine the effects of immigration-induced diversity on student achievement. For example, from 2008 to 2018, the number of broadband subscriptions per 100 individuals in Estonia rose from 22.0 to 33.1, and the number of Estonian coders increased from 9 to 216 during the same period. This relative rise led to a predictable increase in diversity for projects with a minority of Estonian contributors, and a predictable decrease in diversity for projects with a majority of Estonian contributors.

We find that an exogenous increase in team diversity by one standard deviation results in an 11.6 percentage point (pp) increase in the likelihood that a project remains active in the subsequent year. This is a sizeable effect given that more than one-fifth of active projects are abandoned each year. We also find robust benefits of team diversity on project activity and popularity. Crucially, the benefits are highest for teams at a low level of diversity. A one-point increase in diversity raises project survival by 2.6pp for teams with a high level of diversity, while for teams at a low level of diversity, it raises survival by 11.9pp. We see similar relative effects on project activity and popularity. These estimates suggest that teams are "trapped" in a state where their diversity is inefficiently low.

Homophily in team formation has implications for policies that aim to increase diversity. Paradoxically, some policies that would increase diversity in the absence of homophily can decrease diversity when homophily is present. Consider an initiative to raise the diversity of the coder pool by teaching women or underrepresented minorities to code. Such a policy has the unintended effect of making it easier for teams to find similar peers, leading to less diverse teams due to a sorting effect (Tiebout, 1956; Buchanan, 1965). In fact, this is what we see in the data. From 2008 to 2018, the pool of active OSS contributors expanded from 73 countries to 170. Yet over the same period, outsiders were less likely to join low-diversity teams, the average diversity of OSS teams fell, and teams were more likely to be in monoculture in 2018 than they were in 2008.

Conversely, policies such as recruiting outsiders into low-diversity teams can increase both diversity and project success. These policies have both direct effects and indirect trickle-down effects (de Sousa & Niederle, 2022). The direct effect is that adding an outsider to a low-diversity team has a large positive effect on project success. The indirect effect is that adding an outsider to a low-diversity team lowers to cost for other outsiders to join the team, helping lift teams out of the homophily trap. To test for these indirect effects, we construct a matched set of teams that start in monoculture. The first team in each pair adds a single outsider in the focal year, while the matched team adds a single insider. We show that teams that add a single outsider have higher diversity and higher project success in future years and instead of converging, the diversity and project success gaps widen over time.

Our study makes two main contributions. First, we provide the first large-scale causal estimates in the field of the effects of team diversity on project success. We add to a growing literature on team composition and project success (Mas & Moretti, 2009; Chen, 2021; Ewens, 2022). Calder-Wang, Gompers, & Huang (2021) find a negative effect of diversity on team performance in an entrepreneurship competition among MBA students randomly assigned to groups. In contrast, Yang et al. (2022), Azoulay, Jones, Kim, & Miranda (2022) and Bernstein, Diamond, McQuade, Jiranaphawiboon, & Pousada (2022) find positive associations between diversity and citations and business and job creation in large-scale observational data of scientific teams and entrepreneurs. We add to this literature new evidence that diversity in OSS teams leads to higher project success and that it does so in part by broadening the project's appeal.

Second, we document that homophily explains why teams find themselves trapped with low diversity. These findings connect the literature on homophily (Becker, 1957; Boisjoly et al., 2006; Battaglini, Harris, & Patacchini, 2020; Mele, 2022; de Sousa & Niederle, 2022) and club goods (Tiebout, 1956; Buchanan, 1965) with the literature on diversity and productivity (Hong & Page, 2004; Alesina & Ferrara, 2005; Onuchic & Ray, 2023). In light of our results, we conjecture that the fact that all-male and all-female teams are overrepresented relative to mixed-gender teams in scientific and inventive teams (Yang et al., 2022; Subramani, Aneja, & Reshef, 2021) also reflect a homophily trap. In a wide range of contexts, therefore, policies targeted at recruiting outsiders into low-diversity teams could have outsized benefits (de Sousa & Niederle, 2022).

Section 2 discusses homophily in the context of benchmark models of diversity and productivity. Section 3 provides details on open source software, the data, and the distribution and dynamics of team diversity. Section 4 describes our instrumental variable approach and shows our first set of results on team diversity and productivity. Section 5 reports our second set of results on homophily and team formation as well as their policy implications. Section 6 concludes.

2 Homophily and Team Production

Homophily, the tendency of humans to prefer interacting with others who are similar to them, is a robust preference that likely has evolutionary origins (Fu, Nowak, Christakis, & Fowler, 2012). While it is well documented in many economic settings (Becker, 1957; Boisjoly et al., 2006; Battaglini et al., 2020; Mele, 2022), the implications for homophily for team production have not been previously explored.

A workhorse model of team production (Hong & Page, 2004; LiCalzi & Surucu, 2012) presents a team with a creative nonroutine task that has many possible ways to approach it. The team generates ideas on how to approach the task, then works together on the best idea. A more diverse team generates a more diverse set of ideas, so that the best idea will be of higher quality. However, a more diverse team struggles to communicate and coordinate their efforts.

Figure 1 graphs the benefits and costs of diversity in such a model. The benefits exceed the costs at low levels of diversity, so that a more diverse team is more productive. The marginal benefits of adding diversity are large at first, but decreasing, because adding an outsider to an already-diverse team is less of an improvement than adding an outsider to a monoculture. The marginal costs of diversity are small at first, but increasing, because coordination and communication become increasingly more costly with a more diverse team (Becker & Murphy, 1992). Thus, the model predicts an optimum level of diversity in the center of the graph where the net benefits are highest. Low-diversity teams should add diversity and high-diversity teams should reduce diversity to move toward the center, or else fail to thrive.

Figure 1 also graphs a plausible functional form representing homophily, specifically, the utility cost to an outsider of joining the team. We see that in contrast to the standard costs of diversity, the cost of homophily is highest for low-diversity teams, because an outsider pays a higher psychological adjustment cost. By contrast, the cost to an outsider of joining an already-diverse team is lower. Put differently, an outsider who joins a nondiverse team pays a higher private cost than an outsider who joins a more-diverse team.

We now examine two competing hypotheses. If members of the contributor pool have high levels of homophily, then the cost of joining a team in which one would be an outsider is prohibitively high. In this case, we would observe a low level of average team diversity and many teams in monoculture. However, projects also vary in their benefit and cost curves. As a result, without homophily, the distribution of team diversity can take on many different shapes. For example, teams could have low benefits of diversity and high costs of coordination. In this case, we would also observe a low level of average team diversity and many teams in monoculture.

The key difference is that with homophily, teams to the left of the graph are "trapped" at an inefficiently low level from a productivity standpoint. If those trapped teams increased their diversity, we would expect to see better outcomes. By contrast, in the case with low benefits and high costs, teams to the left of the graph are at an efficiently low level of diversity. If those teams increased their diversity, we would expect to see worse outcomes. This is the key distinction that we test.

3 Open Source Software and Data

3.1 Open Source Software projects

Open source software (OSS) is software that is released under a license that allows for the use, inspection, study, modification, and distribution to anyone for any purpose. OSS is a large and growing part of the modern economy. Cloud computing, data science, and e-commerce all directly depend on license-free open-source software. OSS projects include many well-known software products, such as Firefox, LaTeX, R, Python, and Linux, and less well-known but important projects, such as Apache, which is used in 41% of active websites (Netcraft, 2017). Open source projects can start out as hobbies (Python, Linux), private software that is later opened up (Google Tensorflow, Facebook React), or as collaborative OSS from the start (PostgreSQL, AstroPy).

Developers use platforms such as GitHub, GitLab, SourceForge, and Bitbucket to distribute, develop, download, review, and publish their projects. GitHub was created in 2008 and since 2011 has been the most widely used site in the world for coders to collaborate on open-source software.[3] As of 2021 GitHub hosts 118 million open source projects[4] with over 80 million users worldwide[5] and provides an interface for coders to collaborate based on the version control software Git. Git is itself an open-source project created by Linus Torvalds in 2005.

Our data on GitHub activity come from the GHTorrent project (Gousios, 2013) which records comprehensive information about coding activity on GitHub. The GHTorrent database contains detailed data on every public GitHub project since GitHub's inception. This data includes detailed information on the project itself, the set of coders who contribute to the project, and the contributions, also called commits, made to the project.

We focus on large and active software projects, excluding small or non-coding projects. We require that a project has at least one associated coding language, 100 commits, and a team of

[3] https://readwrite.com/github-has-passed-sourceforge/

[4] Code hosted on GitHub is organized in "repositories"; we use "project" interchangeably with "repository".

[5] GitHub, The State of the Octoverse, 2021 https://octoverse.github.com/.

at least ten active contributors when it enters our panel. These filters yield a sample of 148,358 project-years covering 56,696 projects from 2008 to 2018. Those project-years received 83 million commits from 1.8 million coders around the world.

In Table 1A, we report that the average team in a project-year has 22.5 active coders. The median number of coders (team size) is 10, while the 10th and 90th percentile teams have 2 and 42 coders respectively. Our main measure of project development activity is the number of commits made each year. The average number of commits in a project-year is 558.8. The 10th percentile project-year receives 10 commits, while the 90th percentile has 1,267. The average project in our sample is 3.3 years old, and the unconditional survival probability from one year to the next is 79%. That is, each year approximately one-fifth of projects in our sample "die" and cease to be actively developed.

3.2 OSS contributors

We measure the two primary events on GitHub that correspond to proposed and accepted changes to projects: Pull requests and commits. Figure 2 shows a standard workflow. A coder who has identified a feature to add or a bug to fix first "forks" the project, creating a copy that can be modified without affecting the main project (the "master branch"). After they make a change to the code, they issue a pull request which submits the modified code to be merged ("committed") into the master branch. A committer (a coder with commit access to the main project) then reviews the pull request. If the pull request is deemed acceptable, the committer commits the changes into the master branch. For each pull request and commit, we observe the project, timestamp, and submitting coder.

Table 1B contains information on commits per coder per year. The distribution is highly skewed, with a few superstar coders generating substantially more commits than the median coder — a long right tail. Specifically, the 90th percentile coder makes 142 commits per year, more than 20 times more than the 7 commits made by the median coder.[6]

This dispersion in output between coders is striking. The common wisdom is that there is a 10:1 productivity difference between a top coder and an average coder (Oram & Wilson, 2010). The

[6] Not all of the commits necessarily are to the same project. For example, while the average coder contributes seven commits per year, the average number of commits per year from a coder in a given project is 4.2 (Panel A of Table 1).

Google vice president of engineering Alan Eustace has said a top engineer is worth 300 times an average engineer.[7] While there is debate about how to measure productivity differences in software, our measure of output is concrete and consistent with the claims that there are superstar coders.

We also track GitHub "stars", which are similar to a Twitter "follow." After a GitHub user stars a project, the user is informed of any updates or news about the project whenever they log into GitHub. The number of users who star a project is a commonly used measure of the project's popularity.[8] To disentangle users from contributors, we count the stars on each project from users who *never* contribute or merge a pull request to the project. That is, we count stars from Github users who do not ever contribute to the project themselves. The results are similar if we do not exclude contributors from the star count. The average project-year in our sample has 251.6 starring users in total (Table 1A), and collects 72 new stars in a year.

3.3 Measuring Team Diversity

Ideally, we would measure the diversity of a team based on the members' cultural background, knowledge base, or skill set, because those are the sources of a diverse set of ideas. However, these characteristics are less measurable (and also less policy relevant) than verifiable objective characteristics. We make use of objective characteristics that we can observe for the coders in our sample. We use country of location as our primary measure, and (name-based) ethnicity as a secondary measure and robustness check. The advantage of country of location, though fewer coders specify it (22%), is that it is objective and does not need to be imputed. The advantage of name-based ethnicity is that it is available for over 90% of coders in our sample who specify a first and last name in their GitHub profile.

We compute an entropy-based measure of diversity for each project-year. This measure derives from the ecological literature, which centers on the number of *effective species* in a population (Tuomisto, 2010). The measure is based on the fraction of coders contributing to project i in year t from group c, given by $p_{ict} = \frac{N_{ict}}{\sum_c N_{ict}}$, where N_{ict} is the number of active coders. This fraction can be interpreted as the probability of observing a coder from group c in project i in year t if we sampled one coder. We then calculate the generalized Simpson's entropy of order one across the

[7] "Google's Growth Helps Ignite Silicon Valley Hiring Frenzy," Pui-Wing Tam and Kevin J. Delaney, Wall Street Journal, 2005 https://www.wsj.com/articles/SB113271436430704916.

[8] https://www.infracost.io/blog/github-stars-matter-here-is-why/

project. Finally, we monotonically transform it into the index of diversity:

$$
\begin{aligned}
p_{ict} &= \frac{N_{ict}}{\sum_c N_{ict}} \\
X_{it} &= \sum_c p_{ict}(1 - p_{ict}) \\
\text{Effective Groups}_{it} &= \frac{1}{1 - X_{it}} \qquad (1)
\end{aligned}
$$

The measure *Effective Groups* encompasses diversity along different dimensions. For example, when coders are indexed by their country of location, *Effective Groups* is *TeamDiversity*, which is a continuous (noninteger) measure bounded below by 1 and above by the minimum of the number of members and the number of countries. A team with N active coders all from the same country has one effective country (*TeamDiversity* = 1), and a maximally diverse team (each coder from a different country) has N effective countries (*TeamDiversity* = N). Consider projects A and B with coders from the US and China. Project A has 99 coders from the US and one from China, while Project B has 50 from each country. Our diversity index measures these projects as having different levels of diversity: Project A has just over one effective group (*TeamDiversity* = 1.02), while project B has two effective groups (*TeamDiversity* = 2).

The measure nests other standard measures of diversity in the literature (Grabchak, Marcon, Lang, & Zhang, 2017; Yang et al., 2022; Boar & Giannone, 2022). For example, it is a monotone transformation of the racial Herfindahl index (Alesina, Baqir, & Easterly, 1999; Alesina & Ferrara, 2005).

3.4 The Distribution of Diversity

Table 1 shows that the average team in our sample has 22.5 active coders from 6.9 countries with 2.7 effective groups (TeamDiversity=2.7) and from 6.0 ethnic groups with 3.0 effective groups. The difference between the number of countries or ethnicities and their diversity score implies that most projects have an uneven proportion of coders from different countries and ethnicities. The level of diversity varies widely among teams, with the 90th percentile team having TeamDiversity = 5.5 and DivEthnicities = 5.4 groups, while the 10th percentile team has TeamDiversity = 1.0 and DivEthnicities = 1.0, that is, a monoculture.

Figure 3 plots the distribution of team diversity based on country of location. The internet appendix shows that the distribution based on ethnicity is very similar. Theories of the costs and benefits of diversity (e.g. Hong & Page (2004)) predict net benefits of diversity at low levels and net costs at higher levels, leading to an equilibrium distribution with a single peak in the middle. In contrast to this prediction, we see a distribution that is downward-sloping, with teams most likely to be found at the lowest level of diversity, i.e., monoculture.

3.5 The dynamics of diversity

The standard theories of team production, as described, predict a single-peaked distribution of team diversity. Over time, teams at a low level should tend to add diversity because of the benefits, while teams at a high level should tend to lose diversity because of the costs.

Figure 4 plots the transition probabilities between diversity bins from one year to the next. We divide project-years into four bins; monoculture ($TeamDiversity = 1$), low diversity ($TeamDiversity \in (1, 2]$), medium diversity ($TeamDiversity \in (2, 3]$), and high diversity ($TeamDiversity > 3$). Strikingly, more teams transition into monoculture from low-, medium-, and high-diversity teams than monoculture teams transition out to those categories. Moreover, a large portion of teams also transition from low-diversity and medium-diversity bins into the high-diversity teams. The appendix shows that the dynamics based on country are similar. Thus, in contrast to the "central tendency" toward an interior equilibrium, we instead observe a "hollowing out" of the distribution of team diversity, where teams with intermediate levels of diversity tend to become monocultures or high diversity teams. These dynamics lead to a bimodal distribution of diversity, with teams tending toward either monoculture or moderate levels of diversity, consistent with Figure **??**.

To understand why team diversity is evolving in these ways, we next examine the relationship between coders' identity and which coder joins or leaves a team. First, we study the decision to join a team. We define a coder as an outsider relative to the team if, when they *join* the team, it raises the team's level of diversity.

If coders' identity were not a factor in the formation of teams, then outsiders would be more likely to join low-diversity teams. Put differently, the likelihood that a coder with a randomly chosen identity is an outsider to a particular team is naturally decreasing in the team's existing level of diversity. Second, we examine the decision to leave a team. Here, we define a coder as an outsider if when they *leave* (stop contributing to) the team, it lowers the team's diversity. Similarly, if identity were not a factor, outsiders would be more likely to leave high-diversity teams.

We calculate the "no-homophily" null rates of joining and leaving following Jones (2021) by simulating the rates that would occur if coders' identity were irrelevant to these decisions. To do this, we take the set of all coder-joins-project and coder-leaves-project events, and shuffle the coders and projects within each year. This approach preserves the set of projects that were joined and left and the set of coders who joined and left, but it breaks the link between each coder's identity and the identities of the rest of the team.

Figure 5 plots the simulated null rates without homophily, compared with the actual rates at which outsiders join and leave OSS projects. For teams at high levels of diversity, the actual data for outsider join rates are similar to the null simulation. That is, for teams at high levels of diversity, the identities of the coder and the team do not determine which coder joins which project.

By contrast, for teams at low levels of diversity, the actual rate at which outsiders join the project is much lower than the null simulation (Figure 5a). For teams at a medium level of diversity ($TeamDiversity$=3), 62% of new coders who joined increased diversity by joining. Yet for monocultures at a diversity level of 1, only 36% of new coders who joined the team increased diversity by joining. This decrease represents a large gap relative to the null simulation, especially for teams in monoculture, where the predicted fraction of joiners that are outsiders is 80%. This pattern suggests strong homophily in OSS team formation.

Figure 5b shows that at low levels of team diversity, the actual and the null rates at which outsiders leave a project are similar. This pattern is a sharp contrast with the outsider-join rates. It suggests that the effects of homophily are important to outsiders when they join a project, but not when they leave. Once an outsider has joined a low-diversity project, even if the project remains at a low level of diversity, that outsider is not more likely to exit the project. However, for teams at high levels of diversity, we see that outsiders are more likely than the null to leave the project. This is consistent with one of the predictions from the standard model of diversity. Namely, a downside of high team diversity is more disagreement and less retention.

Taken together, the patterns in outsider-join and outsider-leave rates suggest that homophily in coders' decision of which project to join can explain why the level of many OSS teams remain at low levels of diversity. A coder's identity relative to the team's identity has a stark effect on which projects the coder chooses to join. Moreover, this pattern is only observed for low-diversity teams; when joining teams in a high state of diversity, coders' identity is irrelevant. On the other side, coders' decision to leave a project after joining it is unaffected by their identity for teams at low levels of diversity. However, outsiders are more likely to leave high-diversity teams. The appendix shows that the patterns based on country of location are similar. Thus, the low levels of OSS team diversity, and many teams in monoculture, are explained by homophily in coders' decision of which project to join.

3.6 Other explanations

In the internet appendix we examine two alternative explanations for the striking pattern in outsider join rates. First, we test if teams in monoculture are unwelcoming to outsiders. To do this, we examine whether pull requests from outsiders are less likely to be accepted by low diversity teams. We find that low and high diversity teams are equally welcoming to outsiders. Second, we test if outsiders do not join projects simply because they do not know about them. To do this, we reconstruct the outsider join rates as in Figure 5a, only among users who previously starred the project. This ensures that each coder in the analysis knows about the project. The pattern remains: The fraction of joiners who raise team diversity is significantly higher than the no-homophily null distribution, for low-diversity teams only.

4 Team Diversity and Productivity

OSS development involves creative and nonroutine tasks, which benefit from a diverse set of skills and ideas. A diverse team may, however, incur more communication costs and more disagreement because it consists of a decentralized team of coders from around the world without any (formal) hierarchy. Thus, OSS projects are plausibly exposed to significant benefits and costs of team diversity. On the other hand, the patterns in outsider join rates are evidence of strong homophily in OSS coders' decisions of which team to join. On the other hand, Mas & Moretti (2009) find that social pressure helps internalize externalities in teams, so it is possible that OSS teams are not limited by homophily and that teams are efficiently at low levels of diversity.

The key distinction we seek to test is that in a homophily trap, low-diversity OSS teams have large positive marginal effects of diversity. In other words, those teams could increase productivity and popularity by increasing their diversity, but are unable to attract outsiders to join them. By contrast, if homophily is not driving team diversity, an increase in diversity should have a zero or negative marginal effect on productivity and popularity.

We test this hypothesis several ways. Across all our estimates, we find that diversity is beneficial for OSS teams; it increases project success on both the intensive and extensive margins. Moreover, we find that diversity has the largest positive marginal effects for teams at low levels of diversity. These findings are all consistent with the presence of a homophily trap.

4.1 First evidence: Cross-sectional and panel regression

We first examine how the net benefits of team diversity compare for teams at different levels of diversity.

Figure 6 presents binscatter plots across the broad cross section of all teams. Specifically, we sort the entire sample on a measure of team diversity and then divide the sample into bins of approximately equal sizes, roughly 18,000 observations per bin.

First, we investigate how team diversity affects project activity on the extensive margin – the probability that the project continues to be actively developed. One might think that "completed" computer code is static, but in fact, code that is not actively updated and maintained becomes quickly unusable. Even for a small project with few functions, the user base and uses of the code evolve, and bugs continue to be found. For example, the open source cryptographic library OpenSSL, which was created in 1998 and is widely used for secure internet communications, contained a critical bug dubbed "Heartbleed" which was discovered in 2014.[9] Thus, active continued development is a basic indicator of project success.

Our first outcome variable is the indicator $ProjectSurvives_{i,t+1}$, which equals 0 when project i which was active in year t has zero commits in year $t+1$, and 1 otherwise. Cases of a project having a zero-commit year and then being active in a subsequent year are very rare. The average survival rate across all project-years is 79%, meaning that 21% of sample projects become inactive each year. This is a high level of attrition given that we restrict entry into the sample to projects with a

[9] https://blog.torproject.org/openssl-bug-cve-2014-0160/: *"If you need strong anonymity or privacy on the internet, you might want to stay away from the internet entirely for the next few days while things settle."*

healthy codebase and contributor team. Figure 6(a) plots the average likelihood of survival across projects. The sample has been divided into bins, each with approximately 18,000 project-years, on the basis of team diversity.

Second, we examine project activity on the intensive margin – how actively the project is developed. Our second outcome variable is the number of commits (code fixes and additions) that are merged into the project. Figure 6(b) plots the average level of development activity (log of the number of new commits) across bins of team diversity.

Third, we examine a project's popularity with users, measured by the number of users who newly star the project. Figure 6(c) plots the project's popularity with users across bins of team diversity.

The binscatter graphs show strong positive associations between team diversity and project success on the extensive (project survival) and intensive margin (project activity), as well as the project's popularity with users. Moreover, in all cases the marginal benefits of team diversity are larger at low levels of diversity and smaller (flatter) at high levels of diversity. These associations are suggestive and noncausal in nature, since they reflect equilibrium outcomes. For a start, teams with higher diversity are larger on average and those projects likely have a wider appeal, are better run, and are of higher quality in general. To address these potential confounds, we start by estimating panel regressions of the form:

$$Y_{it} = \beta \times Diversity_{it} + \gamma \times (Diversity_{it})^2 + \rho \times \mathbf{X}_{it-1} + \kappa_i + \kappa_t + \epsilon_{it}. \tag{2}$$

These regressions include year fixed effects which sweep out common variation over time, and project fixed effects which sweep out any non-time-varying differences in project quality. They also include project-year controls for team size, size of the codebase (total number of commits since project inception), and project age, all measured as of the prior year.

Table 2 shows that after differencing out the fixed effects and adding controls, team diversity is still strongly positively associated with all three project outcomes. We conclude that both in the overall data, and after controlling for time trends, project-level differences in quality, and confounds such as team size, there is a robust positive relationship between team diversity and project outcomes. Moreover, the coefficient on team diversity squared is robustly negative. This finding supports that the relationship between team diversity and project success is concave – that is, the effects of team diversity are more positive for projects at low existing levels of diversity.

However, it could be that time-varying trends drive changes in both team diversity and project success at the same time. For example, if a particular project is endorsed by a large software company or a prominent coder, then it might attract a burst of popularity, a more diverse set of coders contributing, and better project outcomes at the same time. To address this possibility, we instrument for OSS team diversity using country-by-year changes in broadband internet access.

4.2 Instrumenting for team diversity

We use an instrumental variable (IV) approach to measure the effects of team diversity on OSS project outcomes. Our instrument uses variation across countries in broadband internet access over time. Critically, this instrument changes team diversity and is plausibly unrelated to any given project's outcomes. We find that variation in broadband internet access across countries and time predicts participation in open-source software by that country's residents. This variation in broadband leads to differences in team diversity because coders are more likely to join a project that has coders from the same country. If, in contrast, coders joined projects more or less at random, then on average all projects would have similar team composition. An increase in participation from a particular country would not have much effect on project diversity.

An advantage of using this variation is that an increase in participation in open-source software from a country increases team diversity in some projects and decreases team diversity in other projects. Consider two projects, each with ten active coders. Project A has one coder from Brazil and nine coders from Estonia; Project B has nine coders from Brazil and one coder from Estonia. A higher level of broadband access in Estonia next year predicts more coders from Estonia joining both projects. However, an increase in coders from Estonia raises the diversity of Project A, but lowers the diversity of Project B. Our instrument, therefore, also avoids potential spurious correlation due to changes over time in access to broadband and project success. There is substantial heterogeneity in the effect of our instrument on project diversity, which is necessary to identify causal effects. At the same time because of the three-step procedure described below, the monotonicity assumption is satisfied. Our IV estimates all use $TeamDiversity$ as the measure of diversity, since the instrument varies on a country-by-year level.

We follow the three-step approach recommended by Cameron & Trivedi (2005) and Wooldridge (2010) when using nonlinear instrumental variables:

1. Compute the predicted number of coders from country c contributing to project i in year t.
2. Compute the predicted diversity of each project year based on the estimate from Step 1.
3. Instrument for realized diversity with the predicted diversity from Step 2.

First, we estimate the effect of broadband per capita on the number of coders contributing from country c in year t using the specification,

$$N_{ict} = \beta_1 \times BroadbandperCapita_{ct} + \kappa_i + \kappa_t + \epsilon_{ict} \tag{3}$$

We find that an increase of one broadband subscription per 100 residents predicts a 0.015 log point higher participation rate from that country to a project that has seen contributions from that country before (Table 3, column 3, $z = 21.0$). The results are similar using linear prediction in levels and logs, reported in columns 1 and 2. In levels, we find that one more broadband subscription per 100 residents predicts 0.02 contributors from that country to a given project that has seen contributions from that country before (column 1, $t = 14.6$). In logs, which exclude contributors with zero commits, we find that one more broadband subscription per 100 residents predicts a 0.0066 log point increase in the number of coders (column 2, $t = 51.0$). In all cases, the instrument is statistically strong: The F statistics for the linear models are 212 and 2,603 and the Chi-squared statistic for the Poisson model is 442.

Second, we construct predicted diversity for project i in year t using the predicted values for the number of coders in the project i from country c in year t:

$$PredictedN_{ict} = \beta_1 \times BroadbandperCapita_{ct} + \kappa_i + \kappa_t$$

$$PredictedTeamDiversity_{it} = f(PredictedN_{ict})$$

Third, we use the predicted diversity as an instrument for the realized diversity, as described in Cameron & Trivedi (2005) and Wooldridge (2010):

$$\begin{aligned} TeamDiversity_{it} &= PredictedTeamDiversity_{it} + \lambda_i + \lambda_t + \nu_{it} \\ Y_{it} &= TeamDiversity_{it} + \kappa_i + \kappa_t + \epsilon_{it} \end{aligned} \tag{4}$$

To interpret the IV estimates as causal, two key assumptions are necessary, (1) monotonicity and (2) the exclusion restriction.

4.2.1 Monotonicity The monotonicity requirement in our setting is that for all OSS projects in the sample, higher predicted diversity is correlated with a higher realized diversity. If some OSS projects had a negative relationship between predicted and realized diversity then those "defiers" would bias the IV estimates. As described above, country-by-year broadband access does not have a monotonic relationship with diversity. However, this relationship does not need to be monotonic: Our estimates are consistent as long as monotonicity holds in the relationship between predicted and realized diversity.

The Internet Appendix investigates further and shows that the relationship between predicted and actual diversity is monotonically increasing throughout the sample. We also find that the positive correlation between predicted and realized diversity holds in subsets of the data. Thus, the requirement of monotonicity in the IV relationship is plausibly satisfied.

4.2.2 Exclusion restriction The exclusion restriction for our instrument is that predicted team diversity due to changing broadband access by country-year affects open-source software project outcomes only through its impact on realized team diversity. The exclusion restriction appears likely to hold in our setting because the instrument is a combination of exogenous changes in broadband access across countries, which are not plausibly linked to the outcomes of any OSS project. One concern could be a spurious correlation between project-level outcomes and broadband access in certain countries if both trend similarly over time. Our use of year-fixed effects throughout vitiates this concern. Also, we are able to conduct an additional validation test because (as described above) increased broadband access increases the diversity of some projects and decreases it for others. The Internet Appendix presents estimates showing that our results are similar in cases where increased broadband access *increases* versus *decreases* predicted team diversity.

4.2.3 IV estimates: Project Survival Table 4 Panel A shows estimates of the effects of team diversity on project survival. In the OLS estimates, adding one more effective group of diversity to a team, corresponding to a half a standard deviation increase, is associated with a 3.4 percentage point higher likelihood that the project remains active the next year. Adding the project-by-year control variables (column 2) does not change the positive coefficient of project survival on team diversity.

Columns 3 and 4 present instrumental variables (IV) estimates. We instrument the diversity score of project i in year t with the predicted level of project-year diversity. The conclusion is similar: An increase in team diversity of one effective group, driven only by changes in country-level broadband access, leads to a 5.4 percentage point higher probability that the project remains active the next year. This increase is substantial relative to the baseline that 21% of projects do not survive to the next year.

The IV estimates of the effects of diversity are larger than the OLS estimates. This difference implies that if anything, the OLS estimates are biased downward, yielding estimates of the benefits of diversity that are too small. There are plausible examples of time-varying factors that could account for this finding. For example, projects that experience a burst of popularity for unrelated reasons might attract a more diverse team of contributors, but also be more likely to attract competition.

4.2.4 IV estimates: Project Activity Table 4 Panel B shows OLS and IV estimates of the effects of team diversity on project development activity, i.e., the log number of new commits made to the project. The estimates demonstrate a strong positive relationship between team diversity and the level of development activity in the project. In the OLS estimates, adding one more effective group of diversity to a team is associated with a 25 log point increase in project activity. In the IV estimates, the effect of one more effective group is 56 log points.

As with project survival the IV estimates are larger in magnitude than the OLS estimates for project activity. This difference suggests that omitted variables are biasing the OLS estimates downward; for example, projects with a broader appeal might attract a more diverse team but also have a slower development rate. Again, the conclusion remains across all cases and specifications: Team diversity is positively associated with project development activity.

4.2.5 IV estimates: Project Popularity Table 4 Panel C shows OLS and IV estimates for project popularity, i.e., the log number of users who newly star the project. All the estimates indicate a strong positive relationship between team diversity and project popularity.

In the OLS estimates, raising team diversity by one effective group is associated with an 8 log point increase in project popularity. In the IV estimates, the effect of one more effective group is an increase of 26 log points in users who star the project. The average project-year attracts 72 new stars per year, so the IV estimates correspond to one additional effective group leading to 22 additional users starring the project per year.

The conclusion is the same in all cases and all specifications: Team diversity is a robust positive factor in the success of OSS projects. In the Internet Appendix, we show that the conclusions are the same when we compute diversity based on coders' imputed ethnicity or on coders' imputed gender.

4.3 Other Benefits and Costs

We present further details of the benefits and costs of team diversity in the Internet Appendix and we summarize those findings here.

We find that more diverse teams attract a larger and more diverse user base to the project. This finding suggests a natural "flywheel" that benefits the project, because users of a project are likely to become contributors. Another possibility is that interaction with a more diverse team directly raises coders' productivity, as Corno, La Ferrara, & Burns (2022) find in higher education.

We also find that diversity has downsides. More diverse teams have higher quit rates and are more likely to have a hard fork, which splits the team and signals severe disagreement on the project's future direction. These results are consistent across a variety of estimates.

Thus, team diversity comes with benefits, among which are higher likelihood of project survival, more active project development and a larger and more diverse user base. It also comes with costs, specifically higher quit rates and more disagreement within the team. These findings are consistent with the benefits and costs of diversity that are predicted by models of diversity and team production.

4.4 The Marginal Benefits to Low-Diversity Teams

Our estimates show that the net effect of team diversity on project outcomes (survival, development, and popularity) is robustly positive. These findings are consistent with the hypothesis that diversity is being limited by homophily. However, our estimates reported above include OSS projects at all levels of diversity. The key prediction of the homophily-trap hypothesis is that teams at low levels of diversity, i.e., monoculture, have a positive marginal effect of diversity. That is, those teams *would* benefit from attracting more outsiders, but are unable to do so.

Table 5 shows IV estimates within subsamples split on the lagged level of diversity. Panel A shows estimates of effects on project survival. Columns 1 and 2 select teams that as of the prior year had *TeamDiversity* of 1.5 or less and *TeamDiversity* of 2 or less, respectively. Because of the spike of monocultures in the distribution (Fig 3) these cutoffs are around the 40th and 50th percentile of teams. We see that the net marginal benefits of diversity are larger for low-diversity teams; Column 1 imply that raising a team's *TeamDiversity* from 1.5 to 2.5 raises the project survival rate by 11.9 percentage points. By contrast, Columns 3 and 4 select teams that as of the prior year had *TeamDiversity* of 6 or more and 8 or more respectively. These cutoffs are around the 95th and 98th percentiles of the overall distribution. We see that even for teams already at a high level of diversity, adding to team diversity still has positive effects. Even for the small subset of project years with more than 8 *TeamDiversity*, raising team diversity by one additional point leads to a 2.6 percentage point increase in project survival. However, the effect is much smaller than for low diversity teams. A Wald test between columns 1 and 4 rejects that the benefits of diversity are equal in the two groups ($\chi^2 = 5.81$, $p = 0.016$). Thus, for teams at low levels of diversity, the benefit of increasing team diversity is much larger than for teams at high levels of diversity.

Panels B and C present the same estimates for project development activity and project popularity respectively. Again, we see that the estimated effects of diversity on project outcomes are are highest for low-diversity projects and are still positive but significantly smaller for high-diversity teams.

These results are inconsistent with teams being found *optimally* in a state of low diversity. That is, they suggest that the low average level of diversity (3), the spike of monocultures, and the hollowing-out dynamics of the distribution (4) that we observe are not adequately explained by the project-level costs and benefits of diversity. Instead, these results suggest that many teams are "trapped" in a suboptimal low-diversity state.

5 The Homophily Trap

Our results suggest that team diversity is a positive factor for OSS projects at all levels of team diversity. Put differently, teams appear to be at a low level of diversity relative to their productive potential. The median team has contributors from just 2.7 effective countries, while the set of OSS contributors encompasses 73 countries at the beginning of the sample and 170 by the end.

Why do OSS teams remain at such low levels of diversity when they could benefit from increasing it? As our results in Section 4.2 demonstrate, homophily is a significant factor in OSS team formation. Homophily is a potential explanation for two reasons. First, an outsider joining a group imposes a private cost on the individual, but the benefits of diversity accrue to the team and the users of the project. That is, homophily has private costs and public benefits and in these cases, we expect underutilization (Arrow, 1962). There are documented examples in which homophily may also provide private benefits to individuals. Bell, Chetty, Jaravel, Petkova, & Van Reenen (2019) show that gender homophily can lead to better outcomes for female inventors. This behavior might further drive the homophily trap.

Second, the costs imposed by homophily are highest at low levels of diversity. The first outsider who joins a monoculture pays a high psychic cost. By contrast, an outsider joining an already diverse team faces a lower cost. In other words, homophily imposes private costs on individuals, and is thus likely to be overweighted in their decisions relative to the project-level or social optimum; moreover, those costs are decreasing in the level of existing team diversity. As a result, homophily can cause teams to become trapped at a low level of diversity where the marginal benefits are actually highest. In contrast, other frictions to diversity, such as communication costs, likely increase with diversity and thus would be unable to explain the spike in the distribution at monoculture and other dynamics we find in the data.

5.1 Policy Implications

5.1.1 Expanding the Contributor Pool A key question for policy relating to team diversity is how homophily responds to the changing population of contributors. In particular, the number and diversity of potential OSS team members have expanded dramatically over time. The dramatic "increase in supply" should presumably lead to greater team diversity. Yet in the presence of homophily, a larger and more diverse population gives each coder more similar coders to associate with, and can actually reduce group diversity (Buchanan, 1965; Mele, 2022).

Figure 7(a) shows the graph of outsider join rates split into three-year subperiods. We see that over time, the frequency of outsiders joining low-diversity projects has actually fallen. In 2010-2012, 50% of coders who joined a monoculture were outsiders, i.e., increased the team's diversity by joining. In 2016-2018, only 31% of coders who joined a monoculture were outsiders. Thus, as the pool of OSS contributors has expanded, the homophily trap has become stronger.

The decreasing likelihood over time for outsiders to join low-diversity teams is again consistent with homophily being a first-order impediment to team diversity. As the "supply" of diversity – i.e., the number and diversity of potential contributors – expands, there are two possibilities. If homophily is relatively weak, then the added mixing and greater availability of outsiders will lead to higher team diversity. If homophily is relatively strong, then the larger pool permits coders to increasingly assort only with their own type, increasing the homophily trap. This is what we observe in the data.

Critically, Figure 7(a) also shows that the outsider join rate has been falling over time for low-diversity teams, while for high-diversity teams it has not changed. This observation is again consistent with our mechanism because low-diversity teams are where homophily has bite. By contrast, for high-diversity teams, homophily is less of a factor – their outsider join rates are similar to those under the simulated null distribution, and do not change over time.

What is the net effect on team diversity? Figure 7(b) plots the average team diversity across all teams, and the number countries that are represented by the entire coder population that contributes to our sample projects, year by year. We see that over the sample period, while the diversity of the coder population more than doubled, the diversity of the average OSS team has actually fallen from 3.3 effective groups to 2.7 effective groups. Moreover, the fraction of teams that are in monoculture more than doubled from 15% in 2009 to 34% in 2018.

This striking difference between the trends in the overall coder population compared to team composition is again consistent with strong homophily in OSS teams. This pattern relates closely but not exactly to endogenous public good and club good models such as Tiebout (1956) and Buchanan (1965). In those settings, sorting into increasingly homogeneous groups is unambiguously positive because it better fits all the individuals' preferences. However, in our setting, team diversity is positively related to productivity. As a result, sorting into homogeneous groups is preferable for each individual, but lowers overall productivity.

5.1.2 Escaping the Homophily Trap The previous section suggests that in the presence of strong homophily, policies that broaden the contributor pool (such as encouraging underrepresented minorities to learn to code, or bringing broadband to more countries) can actually reduce the average team diversity.

Our results suggest that for policies to increase diversity effectively, they have to address homophily directly. For example, policies that help recruit outsiders, specifically targeted at low-diversity teams, can have outsized productivity benefits. Such policies have direct positive effects on productivity, per Table 5, and also indirect positive effects because they help the team escape the homophily trap.

To shed light on the possibility that adding diversity helps teams escape the homophily trap, we construct two matched sets of OSS team years, which we denote the blue and red teams. The goal is to directly compare the effects of adding an outsider (blue team) relative to adding an insider (red team) for teams that were monocultures. Specifically, we denote the focal year as year 0. We require that in years -1 and (if available) -2, both teams were monocultures. In year 0, we require that the blue team add exactly one outsider, while the red team add exactly one insider, remaining a monoculture. We also require that the matched pairs have team size and activity (number of commits) within 50% of one another in year 0 and that the two projects are the same age (years since project inception on Github). In the end, we have 142 matched pairs of teams.

Figure 8(a) shows the evolution of teams' diversity in event time. Prior to the event year, both groups were all monocultures as required by our sample selection. In year 0, the red teams added one insider and remained in monoculture while the blue teams added one outsider and escaped to an average of 1.14 effective groups, a difference of 0.14. We see that the gap in diversity between the two groups widens over time. One year later, red teams have an average *TeamDiversity* of 1.05 compared to 1.25 for blue teams, a gap of 0.20 (t=3.4). Two years later, red teams have an average *TeamDiversity* of 1.08 while blue teams have climbed to an average of 1.36, a gap of 0.28 (t=2.0). Thus, the diversity gap widened over time; teams that added a single outsider were more likely to continue adding outsiders and increasing team diversity.

Figure 8(b) shows the evolution of the two groups' development activity (log commits) in event time. By construction, the two groups are identical in their average development activity in pre-event years. After the event year, the groups diverge. The blue teams have more activity on average, with 5.14 log commits in year 1 and 5.56 in year 2, compared to 4.75 and 4.38 for the red teams. Thus, there is a widening gap in development activity of 39 log points (t=1.4) in year 1 and 118 log points (t=3.3) in year 2.

Teams that escape the homophily trap do better on other project outcomes as well. The project survival rates in event-years 1 and 2 are 0.87 and 0.89 for the blue teams compared to 0.83 and 0.77 for the red teams. The log of new starring users are 2.0 and 2.5 for the blue teams compared to 1.4 and 1.6 for the red teams. Importantly, the fraction of contributing coders who stay with the project next year is slightly *higher* for the blue teams in post-event years, and the likelihood of a hard fork is *lower*. These results suggest that the downsides of diversity (disagreement and low retention, discussed in the Internet Appendix) are low or even nonexistent at low levels of diversity, as theory would predict.

We caution that these comparisons are only suggestive because there are endogenous reasons why the blue teams added outsiders while the matched red teams added insiders. Those reasons may be related to future team diversity and future development activity as well. However, the difference in outcomes between the matched pairs is consistent with the idea that recruiting outsiders can help to lift low-diversity teams out of the homophily trap.

Our results again suggest that teams in monoculture have net positive benefits to increasing their diversity – diagnostic of a homophily trap. Adding a single outsider seems to pay off in future years, across all project outcomes that we measure. These findings are also consistent with the evidence of de Sousa & Niederle (2022) who find that a gender quota in French chess led to an increase in the number of female players that was *"...orders of magnitude larger than the number* *of women needed to fulfill quota requirements."*

6 Conclusion

Open source software (OSS) is an increasingly important sector of the modern economy and a fundamental input in the rise of cloud computing, data science, and machine learning. Contributing to OSS is voluntary and unpaid, and the resulting software can be downloaded and used by anyone with an internet connection. Moreover, all code contributions, as well as comments and other aspects of the project and its collaborators, are publicly observable and recorded on GitHub. These features make OSS both a vital public good and a natural laboratory to study what Ostrom (2010) calls commons-based production.

This paper investigates the role of diversity in OSS teams. Across simple associations, panel regressions, and instrumental variables estimates, we find that OSS projects with higher team diversity have higher rates of survival, development and user popularity. Thus, it appears that the voluntary commons-based production of OSS projects benefits from capturing a diverse pool of inputs (Hong & Page, 2004). In this respect, it is similar to other fields such as scientific research (Jones, 2021; Yang et al., 2022) and entrepreneurship (i.e. Hegde & Tumlinson (2014); Azoulay et al. (2022); see Ewens (2022) for a comprehensive review).

One fundamental limit to diversity is homophily – the preference to associate and work with similar peers. Homophily means that diversity imposes a private cost on team members that can lead teams to be too homogeneous. Moreover, this cost is decreasing in the level of team diversity – in other words, the private cost imposed by homophily is highest for the first outsider who joins a monoculture. In the data, we find that the level of diversity among OSS teams is low relative to the pool of contributors, and monocultures are overrepresented among OSS teams. We also find that the rate at which diversity-raising outsiders join a project is sharply increasing in the level of team diversity. Taken together, these findings support homophily as a force that makes it costly for outsiders to join an existing low-diversity team. Homophily causes teams to become trapped in inefficiently low-diversity states.

In the presence of homophily, increasing the pool of diverse candidates can actually worsen the homophily trap and *reduce* team diversity and productivity. We find that as the contributor pool has become larger and more diverse over time, OSS team diversity has actually fallen, and the pattern in outsider join rates has become stronger.

A second implication of our findings is that there is scope for low-diversity projects and communities to increase their productivity by recruiting outsiders. Our findings suggest that such policies can have outsized benefits: First, the direct benefits of diversity to productivity; second, lowering the cost to other outsiders to join and helping the team to escape the homophily trap. In a comparison of matched teams, we find evidence consistent with this hypothesis.

Since OSS development shares key aspects, namely creative nonroutine tasks and homophily among team members, with other settings such as scientific research, innovation, and entrepreneurship, we conjecture that similar conclusions could apply in those settings. Alternatively, the effects of homophily might be stronger in OSS than in other fields precisely because the work is unpaid and voluntary so that preferences do not create profit opportunities (Becker, 1957).

References

Alesina, A., Baqir, R., & Easterly, W. (1999). Public goods and ethnic divisions. *The Quarterly Journal of Economics*, *114*(4), 1243–1284.

Alesina, A., & Ferrara, E. L. (2005). Ethnic diversity and economic performance. *Journal of Economic Literature*, *43*(3), 762–800.

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity: Economic and social factors* (pp. 609–626). Princeton University Press.

Athey, S., & Ellison, G. (2014). Dynamics of open source movements. *Journal of Economics & Management Strategy*, *23*(2), 294–316.

Azoulay, P., Jones, B. F., Kim, J. D., & Miranda, J. (2022). Immigration and entrepreneurship in the United States. *American Economic Review: Insights*, *4*(1), 71–88.

Battaglini, M., Harris, J. M., & Patacchini, E. (2020). *Professional interactions and hiring decisions: Evidence from the federal judiciary* (Tech. Rep.). National Bureau of Economic Research.

Becker, G. S. (1957). *The economics of discrimination.* University of Chicago press.

Becker, G. S., & Murphy, K. M. (1992). The division of labor, coordination costs, and knowledge. *The Quarterly Journal of Economics*, *107*(4), 1137–1160.

Bell, A., Chetty, R., Jaravel, X., Petkova, N., & Van Reenen, J. (2019). Who becomes an inventor in America? The importance of exposure to innovation. *The Quarterly Journal of Economics*, *134*(2), 647–713.

Bernstein, S., Diamond, R., McQuade, T., Jiranaphawiboon, A., & Pousada, B. (2022). The contribution of high-skilled immigrants to innovation in the United States. *NBER Working Paper*.

Bloom, N., Jones, C. I., Van Reenen, J., & Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, *110*(4), 1104–1144.

Boar, C., & Giannone, E. (2022). Consumption segregation. *Working paper*. Retrieved from https://www.minneapolisfed.org/~/media/assets/events/2021/fall-2021-institute-research-conference/elisa-gionnone_consumption-segregation.pdf

Boisjoly, J., Duncan, G. J., Kremer, M., Levy, D. M., & Eccles, J. (2006). Empathy or antipathy? The impact of diversity. *American Economic Review*, *96*(5), 1890–1905.

Buchanan, J. (1965). An economic theory of clubs. *Economica*, *32*(125), 1–14.

Calder-Wang, S., Gompers, P. A., & Huang, K. (2021). Diversity and performance in entrepreneurial teams.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: Methods and applications*. Cambridge university press.

Chen, Y. (2021). Team-specific human capital and team performance: Evidence from doctors. *American economic review*, *111*(12), 3923–3962.

Cornelissen, T., Dustmann, C., & Schönberg, U. (2017). Peer effects in the workplace. *American Economic Review*, *107*(2), 425–456.

Corno, L., La Ferrara, E., & Burns, J. (2022). Interaction, stereotypes, and performance: Evidence from south africa. *American Economic Review*, *112*(12), 3848–3875.

Cullen, Z., & Perez-Truglia, R. (2023). The old boys' club: Schmoozing and the gender gap. *American Economic Review*, *113*(7), 1703–1740.

de Sousa, J., & Niederle, M. (2022). Trickle-down effects of affirmative action: A case study in France. *NBER working paper*. Retrieved from https://www.nber.org/papers/w30367

Ductor, L., Goyal, S., & Prummer, A. (2021). Gender and collaboration. *The Review of Economics and Statistics*, 1–40.

Ewens, M. (2022). Race and gender in entrepreneurship. *Review Article*.

Figlio, D., Giuliano, P., Marchingiglio, R., Ozek, U., & Sapienza, P. (2023). Diversity in schools: Immigrants and the educational performance of us-born students. *Review of Economic Studies*.

Fu, F., Nowak, M. A., Christakis, N. A., & Fowler, J. H. (2012). The evolution of homophily. *Scientific reports*, *2*(1), 845.

Gousios, G. (2013). The ghtorrent dataset and tool suite. In *Proceedings of the 10th working conference on mining software repositories* (pp. 233–236). Piscataway, NJ, USA: IEEE Press. Retrieved from http://dl.acm.org/citation.cfm?id=2487085.2487132

Grabchak, M., Marcon, E., Lang, G., & Zhang, Z. (2017). The generalized simpson's entropy is a measure of biodiversity. *PloS one*, *12*(3), e0173305.

Hegde, D., & Tumlinson, J. (2014). Does social proximity enhance business partnerships? theory and evidence from ethnicity's role in us venture capital. *Management Science*, *60*(9), 2355–2380.

Hjort, J. (2014). Ethnic divisions and production in firms. *The Quarterly Journal of Economics*, *129*(4), 1899–1946.

Hong, L., & Page, S. E. (2004). Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, *101*(46), 16385-16389. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.0403723101 doi: 10.1073/pnas.0403723101

Jones, B. F. (2021). The rise of research teams: Benefits and costs in economics. *Journal of Economic Perspectives*, *35*(2), 191–216.

Lazear, E. P. (1999). Globalisation and the market for team-mates. *The Economic Journal*, *109*(454), 15–40.

Lerner, J., & Tirole, J. (2002). Some simple economics of open source. *The Journal of Industrial Economics*, *50*(2), 197–234.

Lerner, J., & Tirole, J. (2005). The economics of technology sharing: Open source and beyond. *Journal of Economic Perspectives*, *19*(2), 99–120.

LiCalzi, M., & Surucu, O. (2012). The power of diversity over large solution spaces. *Management Science*, *58*(7), 1408-1421.

Mas, A., & Moretti, E. (2009). Peers at work. *American Economic Review*, *99*(1), 112–145.

Mele, A. (2022). A structural model of homophily and clustering in social networks. *Journal of Business & Economic Statistics*, *40*(3), 1377–1389.

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

Onuchic, P., & Ray, D. (2023). Signaling and discrimination in collaborative projects. *American Economic Review*, *113*(1), 210–252.

Oram, A., & Wilson, G. (2010). *Making software: What really works, and why we believe it.* O'Reilly Media, Inc.

Ostrom, E. (2010, June). Beyond markets and states: Polycentric governance of complex economic systems. *American Economic Review*, *100*(3), 641-72. Retrieved from `https://www.aeaweb.org/articles?id=10.1257/aer.100.3.641` doi: 10.1257/aer.100.3.641

Subramani, G., Aneja, A., & Reshef, O. (2021). Persistence and the gender innovation gap.

Tiebout, C. M. (1956). A pure theory of local expenditures. *Journal of political economy*, *64*(5), 416–424.

Tuomisto, H. (2010). A consistent terminology for quantifying species diversity? yes, it does exist. *Oecologia*, *164*(4), 853–860. Retrieved 2023-08-20, from `http://www.jstor.org/stable/40960901`

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data.* MIT press.

Yang, Y., Tian, T. Y., Woodruff, T. K., Jones, B. F., & Uzzi, B. (2022). Gender-diverse teams produce more novel and higher-impact scientific ideas. *Proceedings of the National Academy of Sciences*, *119*(36), e2200841119.

Ytsma, E., & Gallus, J. (2019). The power of public: Recognition and reputation as drivers of open source success. In *2019 meeting papers.*

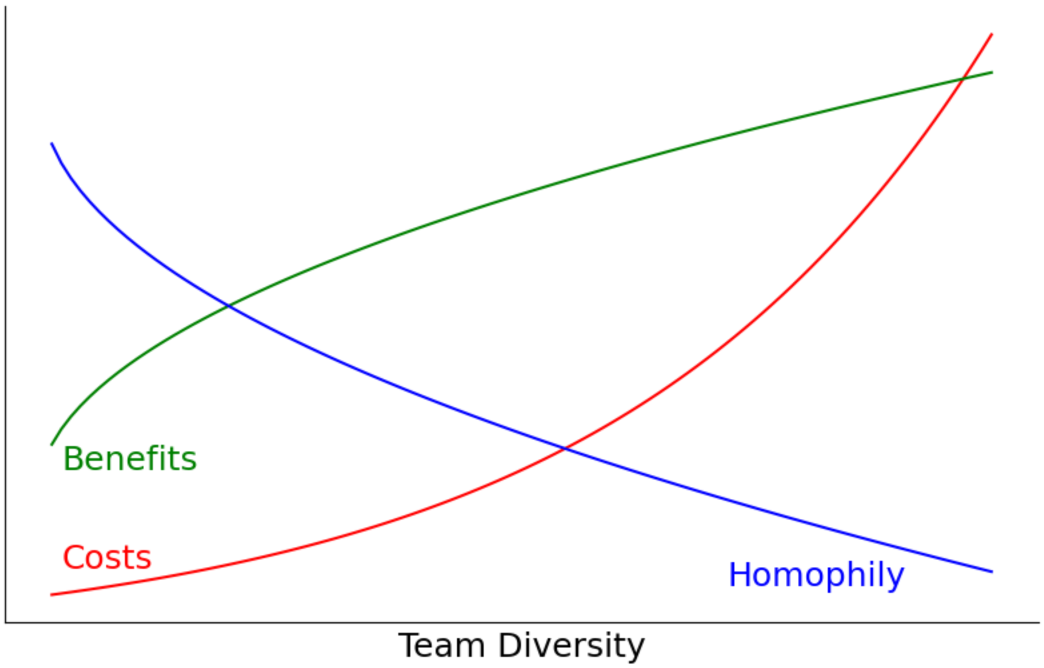


Figure 1: **Homophily and Team Production**. The figure shows the predicted benefits (green) and costs (red) from models of diversity in team production. The figure also shows the disutility of homophily for an outsider who joins the group (blue).

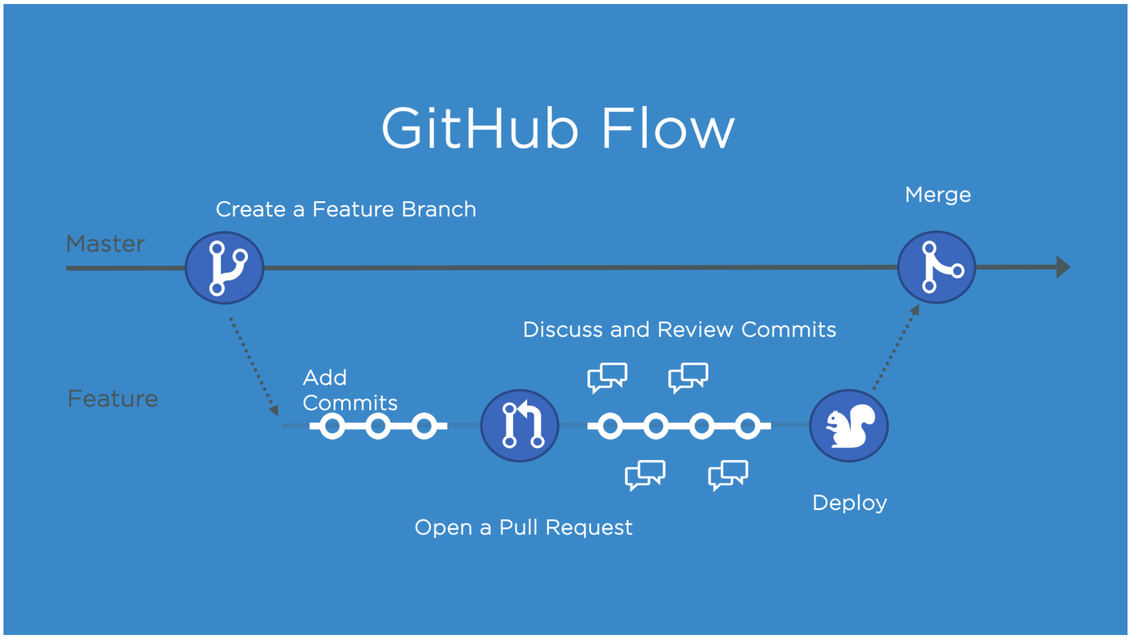


Figure 2: The figure shows a typical GitHub workflow, from https://github.com/SvanBoxel/

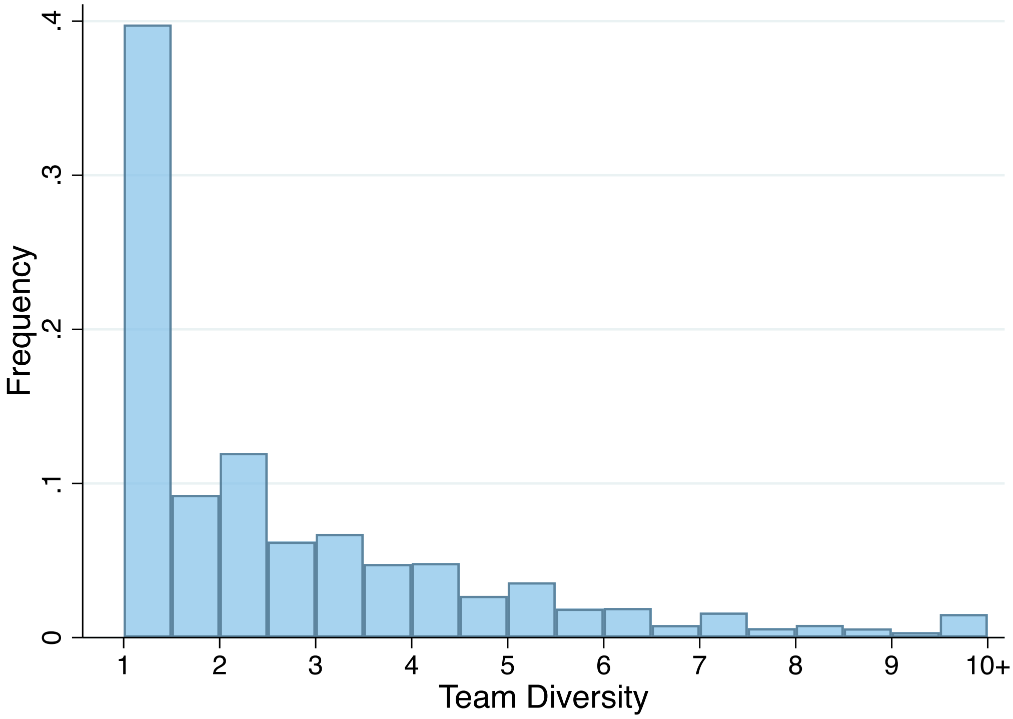


Figure 3: **Distribution of Diversity in OSS Teams**. The figure shows the histogram of team diversity for all project-years in the sample, where diversity is measured on the basis of coders' country of location ($TeamDiversity$).

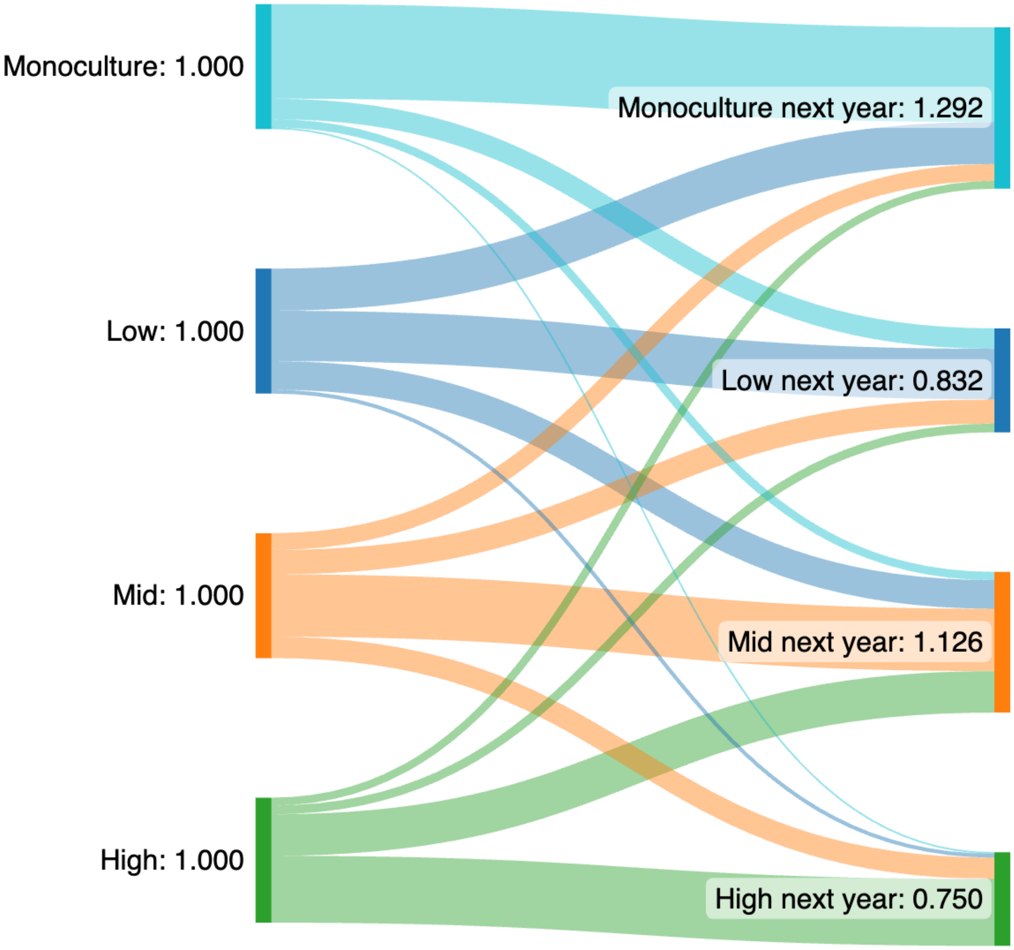
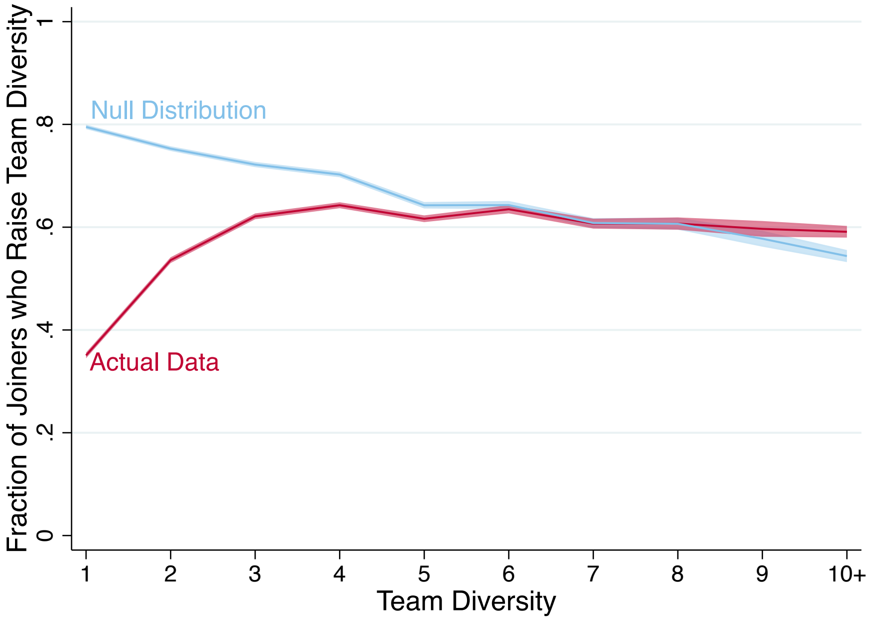
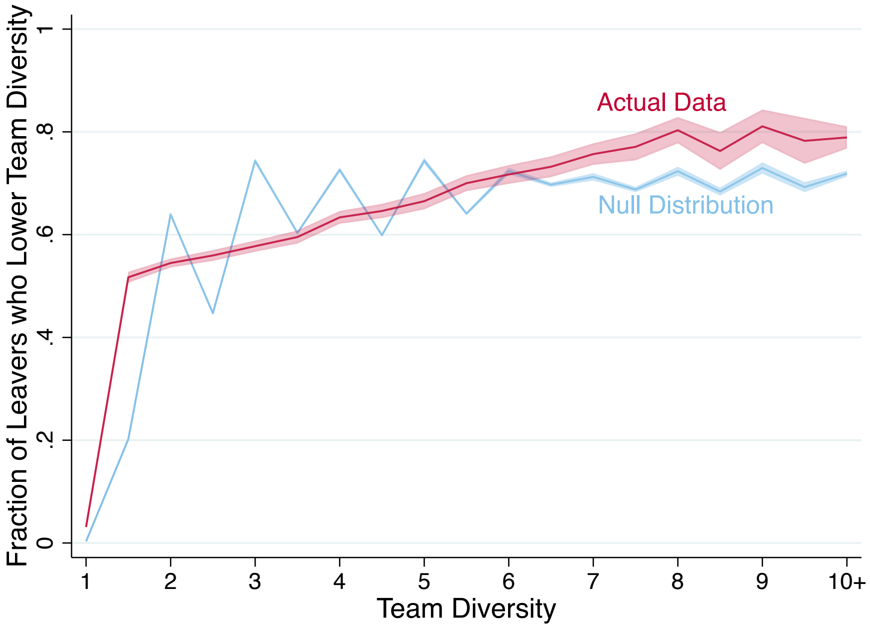


Figure 4: **Dynamics of Team Diversity**. The figure plots the transition probabilities between different levels of team diversity. The left hand side divides teams into bins by their $TeamDiversity$: Monoculture=1, Low∈(1,2], Mid∈(2,3], High∈(3,.). The right hand side plots the measure of teams that, conditional on surviving, are in each bin the following year.

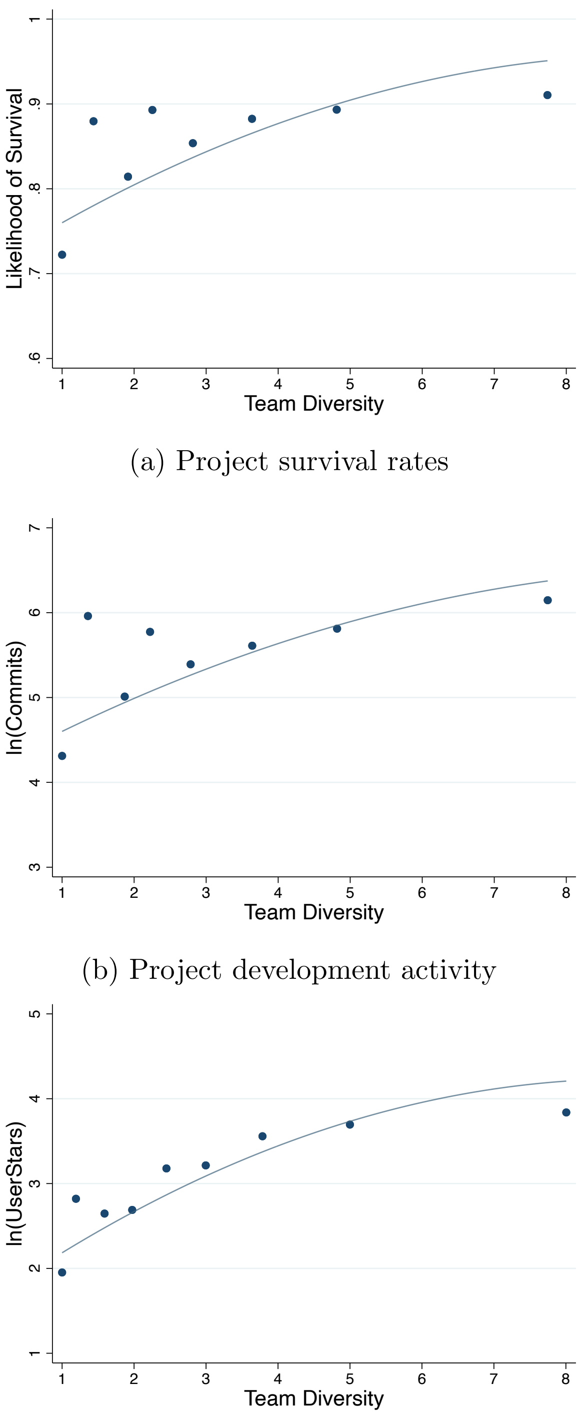


a) Outsiders joining



b) Outsiders leaving

Figure 5: **Homophily in OSS Team Formation**. Figure (a) plots the fraction of new coders who join the project in a given year who are outsiders (raise the team's diversity by joining) sorted by the existing level of team diversity. Figure (b) plots the fraction of coders who leave the project in a given year who are outsiders (lower the team's diversity by leaving), sorted by the existing level of team diversity. Both figures also present simulated null joining rates without homophily. Shaded regions represent 95% confidence intervals.

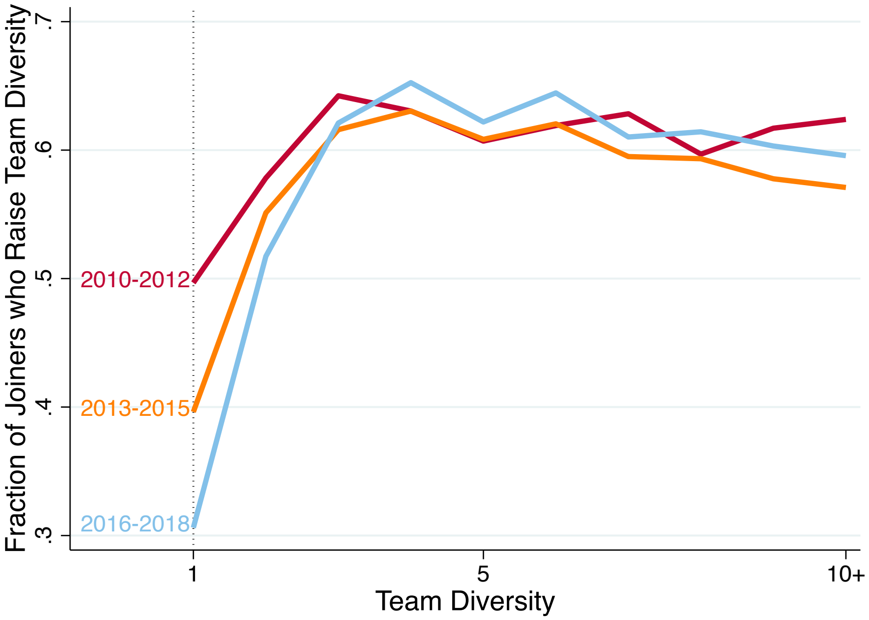


(a) Project survival rates

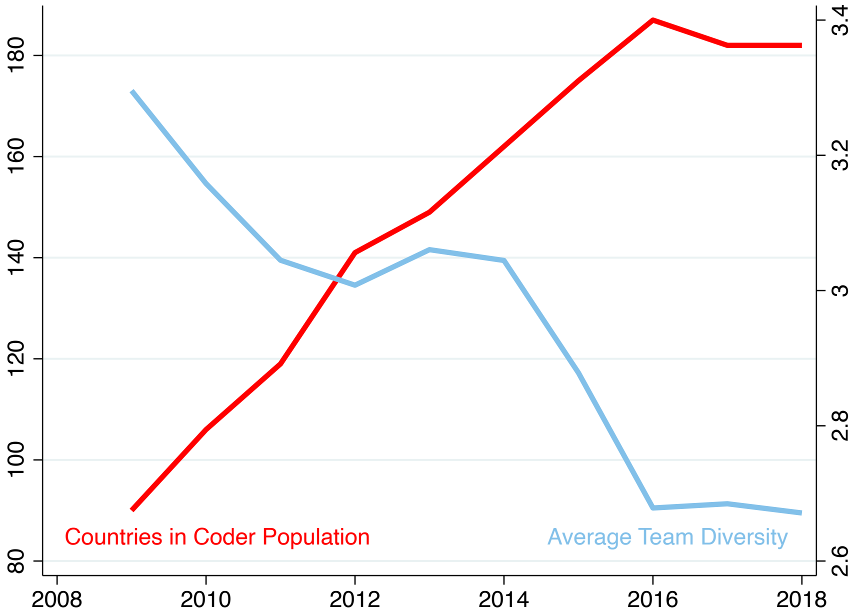
(b) Project development activity

(c) Project popularity

Figure 6: **Diversity and productivity in the cross-section**. Figure (a) plots the overall relationship between team diversity and project survival – the likelihood the project is actively developed the following year. Figure (b) plots project activity — the log number of new commits to the project. Figure (c) plots project popularity — the log number of users who newly star the project. The graphs divide all project-years into bins on the basis of their $TeamDiversity$.

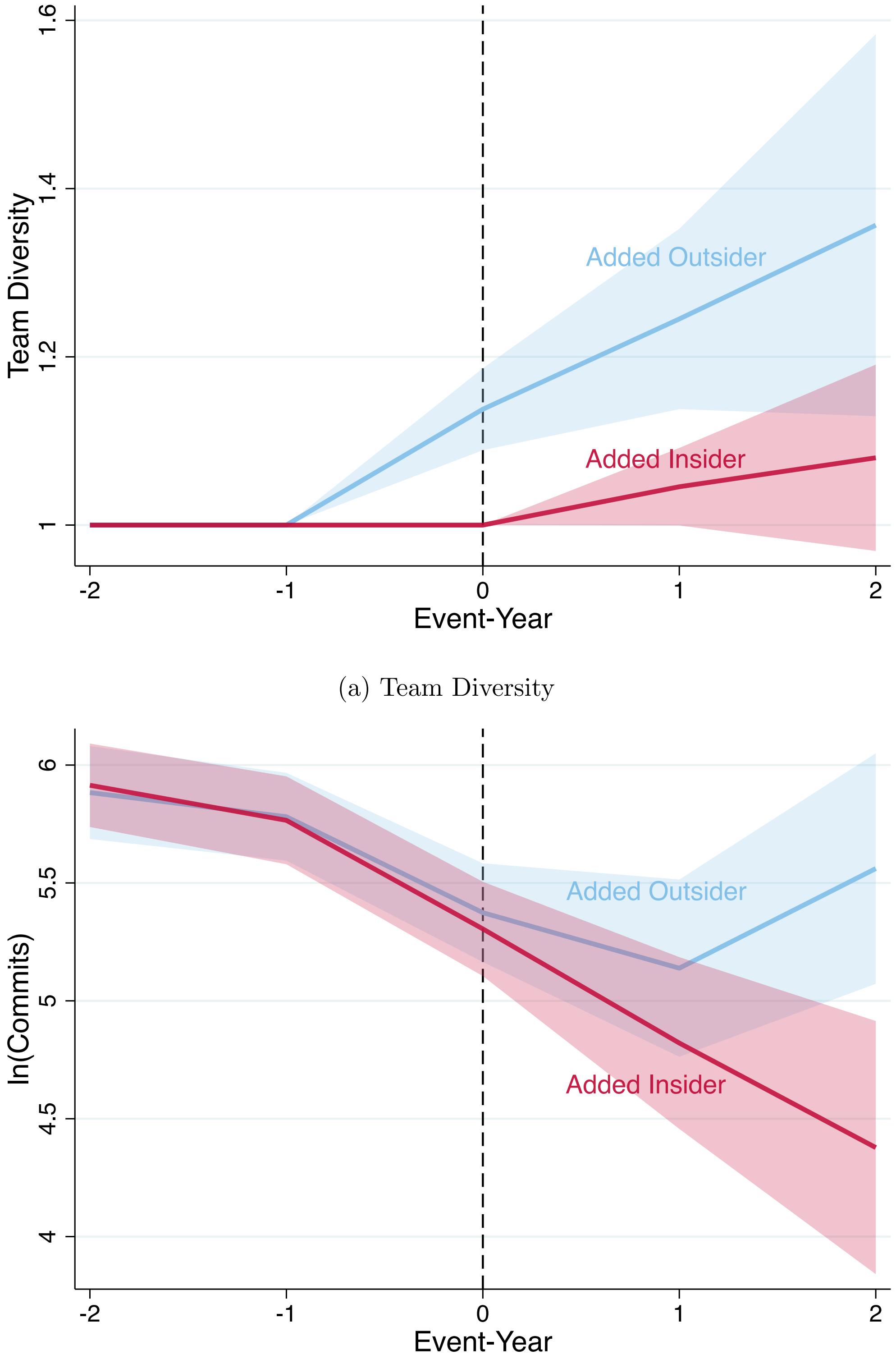


(a) Evolution of outsider join rates over time



(b) Diversity of coder population and teams over time

Figure 7: **Homophily and the Contributor Pool**. Figure (a) shows the fraction of outsiders (new coders who raise the team's diversity by joining) who join the project in a given year, sorted by the existing level of team diversity, dividing the sample into two year subperiods. Figure (b) plots the average diversity of all coder teams and the diversity of the coder population by year.



(a) Team Diversity

(b) Development Activity

Figure 8: **Escaping the Homophily Trap**. The figure plots average diversity and average development activity in event time for two matched sets of OSS teams. The blue teams added one outsider in year 0 while the matched red teams added one insider. The shaded areas show 95% confidence intervals.

Table 1: The table presents summary statistics for both project-year-level and coder-level measures.

A) Project-year measures

	Mean	Std. Dev.	p10	p50	p90	N
Countries	6.93	6.96	1.00	5.00	16.00	127,584
TeamDiversity	2.67	2.18	1.00	2.00	5.54	127,584
Project Age	3.27	1.81	1.00	3.00	6.00	148,358
Coders	22.48	167.74	2.00	10.00	42.00	148,358
Project Survives	0.79	0.41	0.00	1.00	1.00	117,419
Commits	558.78	1792.95	10.00	167.00	1267.00	148,358
Total Commits	2195.78	6212.24	259.00	849.00	4595.00	148,358
User Stars	72.02	392.44	0.00	0.00	118.00	148,358
Total Stars	251.61	1268.74	0.00	2.00	429.00	148,358
Coder Retention	0.42	0.26	0.10	0.39	0.78	93,074
Commits per Coder	33.47	109.14	2.65	14.23	77.12	148,358

B) Coder-year measures

	Mean	Std. Dev.	p10	p50	p90	N
Projects	2.38	10.49	1.00	1.00	4.00	3,016,096
Commits	68.19	1033.20	1.00	7.00	142.00	3,016,096
Total stars	23.21	214.57	0.00	0.00	36.00	3,016,096

Table 2: The table presents estimates of the relationship between team diversity and project outcomes. $ProjectSurvives$ is an indicator variable that equals 1 if the project had at least one commit in future years, and 0 if not. $Commits$ is the number of code fixes and changes committed to the project that year. $UserStars$ is the number of users who newly star the project that year. Project-year controls are each project-year's team size, age since project inception, and lagged total commits. Robust standard errors clustered by project are in parentheses. *: p <0.10, **: p <0.05, ***:p <0.01.

	(1)	(2)	(3)
	$ProjectSurvives_{it+1}$	$ln(Commits)_{it}$	$ln(UserStars)_{it}$
$TeamDiversity$	0.071***	0.500***	0.164***
	(0.002)	(0.009)	(0.007)
$TeamDiversity^2$	-0.0035***	-0.0236***	-0.0076***
	(0.0002)	(0.0008)	(0.0006)
Observations	82,883	108,664	57,333
Adjusted R-squared	0.157	0.675	0.893
Project-year Controls	Yes	Yes	Yes
Project FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes

Table 3: The table presents estimates of the relationship between country-by-year broadband access for open source software participation. N_{ict} is the number of coders in our sample contributing to project i in year t who are from country c. Robust standard errors clustered by project are in parentheses. *: p <0.10, **: p <0.05, ***:p <0.01.

	(1)	(2)	(3)
	N_{ict}	$ln(N_{ict})$	N_{ict}
	OLS	OLS	Poisson
$BroadbandperCapita_{ct}$	0.0188***	0.0066***	0.0148***
	(0.0013)	(0.0001)	(0.0007)
t/z stat	14.6	51.0	21.0
F stat	212.7	2603	
Chi^2 stat			441.7
Adjusted R-squared	0.053	0.119	
Pseudo R-squared			0.193
Observations	1,201,415	734,137	1,201,415
Project FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes

Table 4: **Project success increase with team diversity.** The table presents estimates of the effects of team diversity on open source project outcomes. $ProjectSurvives$ is an indicator variable that equals 1 if the project had at least one commit in future years, and 0 if not. $Commits$ is the number of code fixes and changes committed to the project that year. $UserStars$ is the number of users who newly star the project that year. Project-year controls are each project-year's team size, age since project inception, and lagged total commits. In columns labeled IV, *TeamDiversity* is instrumented with the predicted values from Table 3 column 3. Robust standard errors clustered by the project are in parentheses. *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$.

Panel A: Project Survival				
	(1)	(2)	(3)	(4)
	OLS	OLS	IV	IV
	Dep. Var. = $ProjectSurvives_{it+1}$			
$TeamDiversity_{it}$	0.034***	0.033***	0.054***	0.053***
	(0.001)	(0.001)	(0.003)	(0.003)
Observations	82,883	82,883	82,883	82,883
Adjusted R-squared	0.150	0.150		
Project-year controls	No	Yes	No	Yes
Project FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Panel B: Project Activity				
	(1)	(2)	(3)	(4)
	OLS	OLS	IV	IV
	Dep. Var. = $ln(Commits)_{it}$			
$TeamDiversity_{it}$	0.252***	0.249***	0.568***	0.558***
	(0.004)	(0.004)	(0.013)	(0.013)
Observations	108,664	108,664	108,664	108,664
Adjusted R-squared	0.665	0.665		
Project-year controls	No	Yes	No	Yes
Project FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Panel C: Project Popularity				
	(1)	(2)	(3)	(4)
	OLS	OLS	IV	IV
	Dep. Var. = $ln(UserStars)_{it}$			
$TeamDiversity_{it}$	0.086***	0.083***	0.269***	0.260***
	(0.003)	(0.003)	(0.011)	(0.011)
Observations	57,333	57,333	57,333	57,333
Adjusted R-squared	0.892	0.892		
Project-year controls	No	Yes	No	Yes
Project FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Table 5: **Project success are increasing and concave with team diversity** The table presents estimates of the effects of team diversity on project-level outcomes, in subsamples of teams whose existing level of diversity was low (columns 1-2) versus high (columns 3-4). In all cases, $TeamDiversity_{it}$ is instrumented with the predicted values from Table 3 column 3. $ProjectSurvives$ is an indicator variable that equals 1 if the project had at least one commit in future years, and 0 if not. $Commits$ is the number of code fixes and changes committed to the project. $UserStars$ is the number of users who newly star the project. Robust standard errors clustered by project are in parentheses. *: p <0.10, **: p <0.05, ***:p <0.01.

Panel A: Project Survival				
$TeamDiversity_{i,t-1}$:	<1.5	<=2	>=6	>=8
	(1)	(2)	(3)	(4)
	Dep. Var. = $ProjectSurvives_{i,t+1}$			
$TeamDiversity_{it}$	0.119***	0.091***	0.028***	0.026***
	(0.015)	(0.009)	(0.004)	(0.006)
Project FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	14,532	24,610	9,145	4,259

Panel B: Project Activity				
$TeamDiversity_{i,t-1}$:	<1.5	<=2	>=6	>=8
	(1)	(2)	(3)	(4)
	Dep. Var. = $ln(Commits)_{it}$			
$TeamDiversity_{it}$	0.878***	0.739***	0.368***	0.355***
	(0.065)	(0.045)	(0.016)	(0.022)
Project FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	20,939	34,639	12,307	5,905

Panel C: Project Popularity				
$TeamDiversity_{i,t-1}$:	<1.5	<=2	>=6	>=8
	(1)	(2)	(3)	(4)
	Dep. Var. = $ln(UserStars)_{it}$			
$TeamDiversity_{it}$	0.452***	0.347***	0.165***	0.150***
	(0.073)	(0.037)	(0.015)	(0.019)
Project FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Observations	9,705	16,948	6,145	2,770