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

Abstract

We study diversity and productivity in open-source software teams. Despite a diverse coder pool, many teams persist at low levels of diversity. Using a novel instrument based on initial team composition and country-by-year broadband access, we show that diversity improves project outcomes across multiple measures. We find that homophily – the tendency to associate with similar peers – reconciles these two facts and traps teams in persistently low-diversity states. Teams that escape the "homophily trap" achieve both higher diversity and better project outcomes. Our findings highlight both the benefits of team diversity and the forces that hinder it, offering insights into improving team performance in collaborative environments.

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

JEL Classification: J16, J17, L17, L86

*Email: davidson.heath@eccles.utah.edu. We thank Simona Abis, Colleen Chien, Diego Garcia, Jill Grennan, David Matsa, Randall Morck, Scott Page, Yihui Pan, Paola Sapienza, Uta Schönberg, Emil Siriwardane, Jose de Sousa, Geoffrey Tate, and seminar participants at the AFA, Alberta, CU Boulder, Diversity Pilots Initiative, GEIRC, HK PolyU, HKU, LSU, Michigan, and Utah for comments.

1 Introduction

The growth of the modern economy depends on innovations produced by collaborative teams (Bloom, Jones, Van Reenen and Webb, 2020; Jones, 2021). There is limited evidence on how such teams form, specifically with regard to team diversity and the consequent impact on productivity. Diverse teams offer benefits such as a broader pool of innovative ideas (Hong and Page, 2004; Yang, Tian, Woodruff, Jones and Uzzi, 2022), but they also face higher costs of coordination (Becker and Murphy, 1992; Hjort, 2014; Cornelissen, Dustmann and Schönberg, 2017). A fundamental question for organizational design and policy is how to form collaborative teams in a way that maximizes the net benefits of diversity.

This paper presents evidence that homophily, a private preference, traps teams in suboptimally low levels of diversity. We first document that many teams are at a low level of diversity relative to the pool of coders who contribute. Second, we find that higher team diversity leads to better outcomes across a range of measures. Third, we show that identity matters: Outsiders are less likely to join low diversity teams. Finally, we find that teams that escape the homophily trap exhibit both higher diversity and increased success compared to teams that do not, and that the gap widens over time.

Open source software (OSS) provides an ideal setting to explore how diversity influences team outcomes.[1] By its nature, OSS is produced by collaborative teams who interact globally across national borders. Another advantage of open source software is its size and transparency: We observe granular data on over 50,000 projects across eleven years. Finally, OSS is increasingly important in the modern economy, powering growth in supercomputing, data science, cloud computing, and artificial intelligence, among others (Nagle, 2019).

Despite OSS's global contributor pool, however, we find that team diversity is strikingly low; indeed, monoculture is the most common team composition. The first potential explanation for this pattern is that there are high costs of team diversity and limited benefits, making low levels of diversity efficient. The second potential explanation is that frictions trap teams below their optimal level of diversity. The classic costs of team diversity include communication and coordination costs

[1] Prior economic studies of open source software (Lerner and Tirole, 2002,0; Athey and Ellison, 2014; Ytsma and Gallus, 2019) have focused on the factors that motivate coders to contribute to OSS, such as altruism, career concerns and public recognition.

(Cornell and Welch, 1996), which are lowest for low diversity teams and highest for high diversity teams. Therefore, teams at low levels of diversity are there because the costs are high and benefits are low (Lazear, 1999). Alternatively, homophily is a friction that refers to the preference of individuals to associate with others who are similar to themselves (Becker, 1957; Boisjoly, Duncan, Kremer, Levy and Eccles, 2006; Ductor, Goyal and Prummer, 2021; Cullen and Perez-Truglia, 2023). This friction is *highest* for low diversity teams and *lowest* for high diversity teams. In particular, the less diverse a team is, the more difficult it is for an outsider to join it. As a result, teams can find themselves trapped in an inefficient, low-diversity state. The key test to distinguish between these two explanations is how project success changes with team diversity, particularly for teams at low levels of diversity.

We find positive relationships between team diversity and project success across a variety of research designs and measures. We measure project success in three ways: Survival, development activity, and popularity with users.

First, in the cross-section, team diversity is positively correlated with all three measures of project success. Second, when we control for possible confounds using panel regressions with project-by-year controls and two way fixed effects, we again find strong positive associations. Third, we instrument for variation in team diversity using country-by-year variation in broadband access interacted with team composition.

We find that an instrumented increase in team diversity of one additional effective group (defined below) results in a 5.3 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 positive effects 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, but it raises survival by 11.9pp for teams with a low level of diversity. We see similar relative effects on development activity and popularity. These findings of positive effects from increasing diversity, that are highest for teams at low levels of diversity, are consistent with the homophily-trap hypothesis.

We also find direct evidence of homophily based on contributors' choices of which teams to join. If their identity were irrelevant to which teams they join, then outsiders (coders whose presence

on that team would increase diversity) would naturally be more likely to join low-diversity teams. Instead, we find that outsiders are much *less* likely to join low-diversity teams. Frictions other than homophily do not explain this pattern. We find evidence against discrimination (Evsyukova, Rusche and Mill, 2024) because the rate at which outsiders' code is accepted is higher, not lower, for low-diversity teams. We find evidence against information frictions, such as outsiders being less familiar with low-diversity projects, because we find the same pattern in outsider join rates among coders who have prior knowledge of the project.

We measure team diversity based on either coders' location or their ethnicity. These measures capture different dimensions of diversity, and they vary in coverage. Our baseline estimates use all teams in the GitHub data. The strength of this approach is that it maximizes sample size and generality. The limitation is that some teams have low coverage i.e. we only observe location or ethnicity for a subset of coders. We find similar or stronger results across a variety of robustness tests, including when we require minimum coverage levels from 0% up to 90%. We also find similar results when we measure diversity of core contributors (maintainers) only, regular contributors only, and casual contributors only. That is, the patterns we document are robust features of the data and are not driven by coder characteristics or coverage.

Our instrumental variables research design avoids potential confounders by using variation in existing team composition and broadband access. We predict changes in participation by coders from different countries based on changes in broadband access. These predicted changes have a positive effect on diversity for some projects and a negative effect on diversity for other projects depending on their existing (lagged) composition. 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. Notably, our IV predicts that rising broadband access in Estonia *decreases* diversity for majority-Estonian projects and *increases* diversity for minority-Estonian projects. We find that our estimates are robust and consistent with regard to this variation. In particular, when we restrict to projects whose predicted diversity is *lowered* by higher broadband access, we find that their outcomes are *worse* and not better. Our IV, therefore, avoids any direct effects of broadband access on project success or other confounding factors correlated with the instrument over time.

Homophily in team formation has implications for policies that aim to increase diversity. Some

policies that would increase diversity in the absence of homophily can decrease it in the presence of homophily. Consider an initiative to raise the diversity of the contributor pool by teaching women or underrepresented minorities to code. Such a policy has the unintended effect of making it easier for all coders to find similar peers, and this sorting effect subsequently yields less diverse teams. In fact, this is what we see in the data. From 2008 to 2018, the pool of OSS contributors in our sample expanded and diversified from 73 countries to 170. Yet over the same period, outsiders were less likely to join low-diversity teams, average team diversity 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 produce both direct effects and indirect trickle-down effects (de Sousa and Niederle, 2022). The direct effect is that raising the diversity of a low-diversity team has positive effects on project success. The indirect effect is that it reduces the friction from homophily for other outsiders to join the team, helping the team escape 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, whereas the matched team adds a single insider. We show that teams adding a single outsider have both higher diversity and higher project success in future years and that the gaps in diversity and project success widen over time.

Our study makes two main contributions. First, we provide new large-sample estimates of the effects of team diversity on project success. Doing so adds to a growing literature on team composition and productivity (Mas and Moretti, 2009; Chen, 2021; Ewens, 2022). Calder-Wang, Gompers and Huang (2021) find a negative effect of diversity on team performance in an entrepreneurship competition among MBA students randomly assigned to groups. By contrast, Yang, Tian, Woodruff, Jones and Uzzi (2022), Azoulay, Jones, Kim and Miranda (2022), and Bernstein, Diamond, McQuade, Jiranaphawiboon and Pousada (2022) find positive associations between diversity and citations, business creation, and job creation in large-scale observational data of scientific teams and entrepreneurs. In the context of OSS projects, Daniel, Agarwal and Stewart (2013) find that diversity of reputation improves project outcomes, while diversity of country and language hurts community engagement. Vasilescu, Posnett, Ray, van den Brand, Serebrenik, Devanbu and Filkov (2015) show that gender and tenure diversity improve project output. We add to this literature new evidence on team formation and the underlying mechanisms that limit team diversity.

Second, we document that homophily explains why teams find themselves trapped in low-diversity states. Our findings thus connect the literature on homophily (Becker, 1957; Boisjoly et al., 2006; Battaglini, Harris and Patacchini, 2020; Calder-Wang et al., 2021; Mele, 2022; Evsyukova et al., 2024) and club goods (Tiebout, 1956; Buchanan, 1965) to the literature on diversity and productivity (Hong and Page, 2004; Alesina and Ferrara, 2005; Onuchic and Ray, 2023). In light of our results, we conjecture that a homophily trap explains why both all-male and all-female teams are overrepresented relative to mixed-gender teams in scientific and inventive settings (Subramani, Aneja and Reshef, 2021; Yang et al., 2022). In a wide range of contexts, therefore, policies targeted at recruiting outsiders into low-diversity teams could have outsized benefits (de Sousa and Niederle, 2022).

The paper proceeds as follows. 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 shows our main results on team diversity and productivity, and describes our instrumental variable approach. 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 to interact more with others who are similar, is a robust sociological phenomenon that likely has evolutionary origins (Fu, Nowak, Christakis and Fowler, 2012). Though homophily is well documented in many economic settings (Becker, 1957; Boisjoly et al., 2006; Mele, 2022), exploration of its implications for team formation and production has just begun (Calder-Wang et al., 2021).

A workhorse model of team production (Hong and Page, 2004; LiCalzi and Surucu, 2012) presents a team with a creative nonroutine task approachable in many possible ways. The team generates ideas on how to approach the task and 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 its 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 provides for 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 and 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 both fail to thrive.

Figure 1 also graphs a plausible functional form representing homophily, specifically, the utility cost to an outsider of joining the team. In contrast to the standard costs of diversity, the friction of homophily is highest for low-diversity teams, because an outsider feels least at home and vice versa. By contrast, the friction to an outsider to join an already-diverse team is lower.

We examine two competing hypotheses. If a member of the contributor pool has high levels of homophily, her cost of joining a team is prohibitively high because she would be an outsider. 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.

In the case of homophily, teams to the left of the graph are trapped at an inefficiently low level. If these 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 reside 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 released under a license that allows for the use, inspection, study, modification, and distribution to anyone for any purpose. OSS comprises a large and growing

part of the modern economy. Cloud computing, data science, e-commerce and artificial intelligence 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 it has been the most widely used site in the world through which coders collaborate on open-source software.[2] As of 2021, GitHub hosts 118 million open source projects[3] with over 80 million users worldwide.[4] It 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. These data include 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 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.

Table 1A reports that the average team in a project-year has 22.5 active coders. The median number of coders (team size) is 10, and 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

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

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

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

is, each year just over one in five projects "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 the coder makes 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, submitting coder, and committer.

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.[5]

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 and Wilson, 2010). Google vice president of engineering Alan Eustace has said that a top engineer is worth 300 times an average engineer.[6] Though there is debate about how to precisely measure productivity differences in software, our measure of output is concrete and consistent with the existence of superstar coders.

We also track GitHub "stars," which are similar to Twitter "follows." After a GitHub user stars a project, it becomes bookmarked in the user's profile. The number of users who star a project is a commonly used measure of its popularity.[7] 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

[5] Not all of the commits are necessarily to the same project. For example, while the average coder contributes 68 commits per year, the average number of commits per year from a coder in a given project is 33 (Panel A of Table 1).

[6] "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`.

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

results are similar if we do not exclude contributors from the star count. The average project-year in our sample has 252 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. Hence, We make use of the 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 37% of all coders in the sample. The availability of both measures is higher for more active coders. Among coders with at least two active years and 50 commits in total, country of location and ethnicity are available for 37% and 46% of the sample respectively.

The Internet Appendix shows results across a variety of robustness tests, including when we reestimate on subsamples of the data requiring teams to have minimum coverage levels from 0% (baseline) up to 90%. It also shows the results when we measure the diversity of core contributors (maintainers) only, regular contributors only, and casual contributors only. The conclusions are the same or stronger in all cases, meaning that the patterns we document are robust features of the data and are not driven by coder characteristics or coverage.

We compute an entropy-based measure of diversity for each project-year. This measure comes from the ecological literature where it measures 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 project. Finally, we monotonically transform the order one entropy into the index of diversity:

$$p_{ict} = \frac{N_{ict}}{\sum_c N_{ict}}$$

$$X_{it} = \sum_c p_{ict}(1 - p_{ict})$$

$$\textit{Effective Groups}_{it} = \frac{1}{1 - X_{it}} \tag{1}$$

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 U.S. 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 and Zhang, 2017; Yang et al., 2022; Boar and Giannone, 2022). It is a monotone transformation of the Blau index (Vasilescu et al., 2015) and the racial Herfindahl index (Alesina, Baqir and Easterly, 1999; Alesina and Ferrara, 2005).

The Internet Appendix reruns all our main estimates in subsamples in which we condition the sample down by requiring coverage of up to 90% or higher, for both coder location and name-based ethnicity. It also presents a variety of robustness tests. The conclusions are the same in all cases, which shows that our results are robust to alternative methods and are not explained by variation in reporting across teams.

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 distinct 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. The 90th percentile team has TeamDiversity = 5.5 and DivEthnicities = 5.4 groups, and the 10th percentile team has TeamDiversity = 1.0 and DivEthnicities = 1.0 – in other words, 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 and 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 downward-sloping distribution, 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 high-diversity teams. The appendix shows that the dynamics based on ethnicity are similar. Thus, in contrast to a "central tendency" toward an interior equilibrium, we instead observe a "hollowing out" of the distribution of team diversity; 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 3.

To understand why team diversity is evolving in these ways, we next examine the relationship between coders' identity and which coders join or leave 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, the team's level of diversity increases. If coders' identities 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, the team's diversity decreases. 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 so, we take the set of all coder-joins-project and coder-leaves-project events, and we 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 to 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 in 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, for which the predicted fraction of outsider joiners 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 sharply contrasts 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, we see that, for teams at high levels of diversity, 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 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, a coder's identity is irrelevant. On the other side, a coder's 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 Internet Appendix shows that the patterns based on imputed ethnicity are similar. Thus, the low levels of OSS team diversity are explained by homophily affecting coders' decisions of which project to join.

3.6 Alternative 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 more hostile to outsiders (Evsyukova et al., 2024). 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 same: The fraction of joiners who raise team diversity is significantly lower than in the no-homophily null distribution, for low-diversity teams only.

The Internet Appendix also shows the outsider-join-rate graphs when we filter down to the subset of teams for which our diversity measure has 90% or greater coverage. The patterns are identical – that is, variation in coverage does not explain the patterns that we find in outsider join rates.

4 Team Diversity and Project Outcomes

In this section, we estimate to what extent diversity affects project outcomes and whether homophily is a friction that limits teams from their full potential. OSS development involves non-routine tasks

that benefit from diverse skills and ideas, and is carried out by teams of volunteers from many backgrounds and countries. As a result, OSS projects are plausibly exposed to significant benefits and costs of team diversity.

4.1 First Evidence: Cross-sectional and Panel Regressions

4.1.1 Cross-Sectional Comparisons Figure 6 presents binscatter plots that compare project outcomes at different levels of team diversity. Specifically, we sort all team-years on our measure of team diversity and then divide the sample into bins of equal size, roughly 18,000 observations per bin.

First, we investigate how team diversity is associated with 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 quickly becomes unusable. Even for a small project with few functions, the user base and uses of the code evolve, and bugs develop. For example, the open source cryptographic library OpenSSL, which was created in 1998 and which is widely used for secure internet communications, contained a critical bug dubbed "Heartbleed," which was discovered in 2014.[8] 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 healthy codebase and contributor team. 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. Third, we examine a project's popularity with users, measured by the number of users who newly star the project.

The binscatter graphs show strong positive associations between team diversity and project success on the extensive and intensive margin, as well as the project's popularity with users. Moreover,

[8] 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."*

in all cases, the marginal benefits of team diversity are larger at low levels of diversity and smaller at high levels of diversity. These associations are suggestive and noncausal in nature because they reflect equilibrium outcomes. For example, teams with higher diversity are larger on average, and those projects likely have a wider appeal and are of higher quality in general.

4.1.2 Panel Regressions To address these potential confounds, we start by estimating panel regressions of the form:

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

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 adding the fixed effects and controls, team diversity is positively associated with all three project outcomes. We also see that the relationship between team diversity and project success is concave – that is, the association of team diversity with project success is more positive for projects at low existing levels of diversity.

However, it could be that time-varying trends affect team diversity and project success simultaneously in other ways. For example, if a 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. To address this type of confound, we instrument for team diversity using variation in country-by-year changes in broadband internet access and initial team composition.

4.2 Instrumenting for Team Diversity

We use an instrumental variable (IV) approach to measure the effects of team diversity on OSS project outcomes. Two important facts drive our IV strategy. First, teams vary widely in their country-level composition, even for teams with the same level of diversity. This variation means that a new coder from a given country, e.g., Estonia, can increase diversity in some teams and

decrease it in others.[9] Second, the level of broadband access in a given country strongly predicts participation by coders from that country. Together, these facts suggest that variation in broadband access provides an exogenous shock to the number of coders from a given country, and the same shock increases diversity for some teams and decreases it for others. Our instrumental variable, therefore, isolates exogenous variation in diversity that is created by the interaction of broadband access and pre-existing team composition.

This IV has several appealing features. First, the substantial heterogeneity in the effect of our instrument on project diversity helps to rule out alternative explanations. For example, suppose Estonian coders have unusually high productivity. Adding Estonians raises diversity for some teams and lowers it for others, limiting the extent to which differences in coder ability by country could bias our estimates. We also estimate the IV effects separately for teams that experience (instrumented) increases versus decreases in diversity. We find that teams whose diversity was lowered by increased broadband access have worse outcomes, not better. Thus, differences in coder ability by country cannot explain our findings. Second, we find that instrumented diversity is highly correlated with actual diversity and the relationship is monotonic. Finally, the IV avoids potential spurious correlations due to changes over time in access to broadband and project success by including both project and year fixed effects.

We follow the three-step approach recommended by Cameron and 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 three models of the form,

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

[9]Consider two projects, each with ten active coders, to demonstrate how our instrumental variable isolates the exogenous variation. 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 while lowering the diversity of Project B.

In the Poisson model 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, as 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 relationship 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 Equation 1 and the predicted values for the number of coders in 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 a linear instrument for the realized diversity, as described in Cameron and 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 (step 2) need not be monotonic. Our estimates are consistent as long as monotonicity holds in the step 3 relationship between

predicted and realized diversity.

The Internet Appendix 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 project outcomes only through its impact on team diversity. This restriction is plausible because the instrument is a product of changes in broadband access across countries along with preexisting team composition, and these changes are not plausibly linked to the outcomes and the team composition of any particular project. Also, one might worry about a spurious correlation between project-level outcomes and broadband access if both trend upwards over time. This is not a concern because of our use of preexisting team composition and the inclusion of year fixed effects in all of our estimates.

We are also able to conduct additional validation tests because (as described above) increased broadband access increases predicted diversity for some projects and decreases it for others. The Internet Appendix presents estimates showing that our conclusions are the same in cases where increased broadband access *lowers* predicted team diversity. Put differently, when we restrict to projects whose predicted diversity is lowered by increased broadband access, we find that their outcomes are worse and not better.

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 of 21% of projects not surviving 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 and yield 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 in the case of 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. Once 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.

According to the OLS estimates, increasing 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 19 additional users starring the project per year.

The conclusion is the same in all cases. Team diversity is a robust positive factor in the success of OSS projects.

4.3 Robustness and Validation Checks

4.3.1 Diversity Measurement One concern might be that our findings are specific to our geographic measure of team diversity or due to coders selectively choosing whether to report their location. In the Internet Appendix, we show that our findings hold when we compute team diversity based on coders' imputed ethnicity. This measure is observable for more coders in our sample but must be imputed on the basis of coders' names, increasing coverage but potentially increasing noise as well. We also repeat our estimates while progressively filtering down to subsamples in which at least a minimum fraction of coders, from 10% up to 90%, report their location or name, respectively. In all cases, our findings continue to hold, and the estimated effects are similar in magnitude. This set of robustness checks shows that variation in reporting or coverage does not explain our findings.

4.3.2 Exclusion Restriction Another potential concern is that broadband access might violate the only-through condition by improving project outcomes directly. For example, a rise in broadband access leads to more OSS contributions from the focal country but might also make OSS projects tied to that country more popular. A key advantage of our approach is that, as mentioned, variation in broadband access has a positive effect on diversity for some projects and a negative effect on diversity for other projects. In particular, our IV predicts that rising broadband access *lowers* diversity for some projects.

The Internet Appendix presents estimates when we condition the sample on whether the IV predicts lower or higher diversity than the project's actual realized diversity. We are particularly focused on the case where the IV predicts lower diversity than the reality. We find, within both subsamples, that the instrumented effects of team diversity are strongly positive for all three outcome variables – project survival, activity and popularity. In particular, the positive relationship between team diversity and project success remains in the subsample of project-years for which predicted team diversity was below realized team diversity. For those project-years, changes in broadband access *lowered* team diversity in expectation, yet the relationship between diversity and project success was still positive. This finding cuts against the concern that broadband access could

be affecting project success directly.

4.3.3 Heterogeneity in Coders' Roles Although there is no formal hierarchy in OSS development, coders differ in their roles within the team. First, some coders have "commit access" and are able to approve proposed changes to the project's code. We refer to these coders as "maintainers," as they are the most connected to the project and act somewhat like managers. Second, some coders consistently contribute code but do not have commit access. Finally, some coders are casual or one-off contributors, as anyone can contribute by submitting new code to be merged as in Figure 2.

We find that diversity is important for all roles within OSS. The Internet Appendix shows that the same conclusions apply when we reestimate all our results defining the team as only maintainers; only contributors (i.e. non maintainers); and also splitting on the basis of frequent contributors versus casual contributors to the project. For all of these measurements of team diversity, we find a strong positive effect of diversity on project outcomes. The strong relationship between team diversity and project outcomes across roles in OSS also helps to rule out confounding factors. For example, if we found that the association was positive only for team diversity among casual contributors, we may be more worried about reverse causality (where projects that are expected to succeed attract a more diverse set of causal contributors). This is not what we find.

4.4 Other Benefits and Costs

We present further details of the benefits and costs of team diversity in the Internet Appendix.

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

Overall, 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. Team diversity

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 predicted by models of diversity and team production.

4.5 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 homophily-trap hypothesis. 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, these teams would benefit from attracting more outsiders, but they are unable to do so.

Table 5 shows IV estimates within subsamples split on the lagged level of diversity. Panel A shows estimates of the 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 implies that raising a team's *TeamDiversity* from 1.5 to 2.5 raises the project survival rate by 11.9 percentage points. In 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 the hypothesis that the benefits of diversity are equal in the two groups ($\chi^2 = 5.81$, $p = 0.016$). Thus, the benefit of increasing team diversity is much larger for teams at high levels of diversity than for those at low levels.

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 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 inefficient low diversity states.

5 The Homophily Trap

5.1 Summary

Our results suggest that team diversity is a positive factor for OSS projects at all levels of team diversity. Put differently, many teams appear to reside at a low level of diversity relative to their potential. The median team has contributors from just 2.7 effective countries, whereas 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? Homophily is a plausible 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 for outsiders, and therefore we expect that it may be underprovided (Arrow, 1962). There are documented examples in which homophily can provide private benefits to less-represented individuals. Bell, Chetty, Jaravel, Petkova and Van Reenen (2019) show that gender homophily leads to better outcomes for female inventors. However, ultimately this behavior might exacerbate 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 mental cost. By contrast, an outsider joining an already diverse team faces a lower cost. In other words, homophily imposes private costs on individuals, which are thus likely to be overweighted in their decisions relative to the project-level or social optimum; moreover, those costs decrease with 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, increase with diversity and thus cannot explain the overall low levels of team diversity and the

spike in the distribution at monoculture.

5.2 Policy Implications

5.2.1 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. This 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 additional similar coders with whom to associate – and thus can actually reduce group diversity (Buchanan, 1965; Mele, 2022).

Figure 7(a) graphs 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 acting as 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 possible outcomes. 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. In fact, the latter 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, whereas it has not changed for high-diversity teams. This observation is again consistent with our mechanism because low-diversity teams are where homophily has bite. By contrast, homophily is less of a factor for high-diversity teams – 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 of countries that are represented by the entire coder population that contributes to our sample projects, year by year. We observe that, over the sample period, the diversity of the coder population more than doubled, but the diversity of the average OSS team

fell from 3.3 effective groups to 2.7 effective groups. The fraction of teams 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 as described in Tiebout (1956) and Buchanan (1965). In public good and club good models, 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 it lowers overall productivity.

5.2.2 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 policies must address homophily directly to increase diversity effectively. 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, and 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 consisted of 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 design, 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 is 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 the next year is slightly *higher* for the blue teams in post-event years, and the likelihood of a hard fork is *lower*. These results indicate 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. However, the difference in subsequent outcomes between the matched pairs supports the idea that recruiting outsiders can help 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 and 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 fundamental input into science and innovation. Contributing to OSS is voluntary and unpaid, and the resulting software can be downloaded and used by anyone with an internet connection. Moreover, all OSS code contributions, user stars, and other aspects of each project and its contributors are observable. These features make OSS an important public good and a natural laboratory to study what Ostrom (2010) calls commons-based production.

This paper investigates the role of diversity in team-based production of OSS. Across simple associations, panel regressions, and instrumental variables estimates, we find that 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 benefits from capturing a diverse pool of inputs (Hong and Page, 2004). In this respect, it is similar to other fields such as scientific research (Jones, 2021; Yang et al., 2022) and entrepreneurship (Hegde and 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 friction on team members that leads 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 that monocultures are overrepresented among OSS teams. We also find that the rate at which outsiders join a team is increasing in the level of team diversity. Taken together, these findings support the conclusion that homophily is a friction that makes it costly for outsiders to join a team with low diversity. Thus, it can cause teams to become trapped in inefficient low diversity states.

One implication of these findings is that increasing the diversity of the contributor pool can worsen the homophily trap and *reduce* team diversity and productivity. We find that as the OSS contributor pool has become larger and more diverse over time, team diversity has actually fallen, and the tendency of outsiders to shun low-diversity groups has become stronger. More generally, common practices aimed at increasing the diversity of candidate pools might be ineffective or

counterproductive.

A second implication is that low diversity teams can raise both diversity and productivity by adding outsiders. Policies aimed at this outcome can have outsized benefits. First, they directly enhance productivity. Second, they lower the friction to other outsiders joining and help the team to escape the homophily trap (de Sousa and Niederle, 2022). In a comparison of matched teams, we find that the data is consistent with this hypothesis. More generally, our evidence supports policies that specifically recruit outsiders or lower frictions to outsiders joining low diversity teams.

References

Alesina, Alberto and Eliana La Ferrara, "Ethnic diversity and economic performance," *Journal of Economic Literature*, 2005, *43* (3), 762–800.

Alesina, Alberto, Reza Baqir, and William Easterly, "Public goods and ethnic divisions," *The Quarterly Journal of Economics*, 1999, *114* (4), 1243–1284.

Arrow, Kenneth, "Economic Welfare and the Allocation of Resources for Invention," in "The Rate and Direction of Inventive Activity: Economic and Social Factors," Princeton University Press, 1962, pp. 609–626.

Athey, Susan and Glenn Ellison, "Dynamics of open source movements," *Journal of Economics & Management Strategy*, 2014, *23* (2), 294–316.

Azoulay, Pierre, Benjamin F Jones, J Daniel Kim, and Javier Miranda, "Immigration and entrepreneurship in the United States," *American Economic Review: Insights*, 2022, *4* (1), 71–88.

Battaglini, Marco, Jorgen M Harris, and Eleonora Patacchini, "Professional Interactions and Hiring Decisions: Evidence from the Federal Judiciary," Technical Report, National Bureau of Economic Research 2020.

Becker, Gary S., *The Economics of Discrimination*, University of Chicago press, 1957.

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

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen, "Who becomes an inventor in America? The importance of exposure to innovation," *The Quarterly Journal of Economics*, 2019, *134* (2), 647–713.

Bernstein, Shai, Rebecca Diamond, Timothy McQuade, Abhisit Jiranaphawiboon, and Beatriz Pousada, "The contribution of high-skilled immigrants to innovation in the United States," *NBER Working Paper*, 2022.

Bloom, Nicholas, Charles I Jones, John Van Reenen, and Michael Webb, "Are ideas getting harder to find?," *American Economic Review*, 2020, *110* (4), 1104–1144.

Boar, Corina and Elisa Giannone, "Consumption Segregation," *Working paper*, 2022.

Boisjoly, Johanne, Greg J Duncan, Michael Kremer, Dan M Levy, and Jacque Eccles, "Empathy or antipathy? The impact of diversity," *American Economic Review*, 2006, *96* (5), 1890–1905.

Buchanan, James, "An Economic Theory of Clubs," *Economica*, 1965, *32* (125), 1–14.

Calder-Wang, Sophie, Paul A. Gompers, and Kevin Huang, "Diversity and Performance in Entrepreneurial Teams," 2021.

Cameron, A Colin and Pravin K Trivedi, *Microeconometrics: Methods and Applications*, Cambridge university press, 2005.

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

Cornelissen, Thomas, Christian Dustmann, and Uta Schönberg, "Peer effects in the workplace," *American Economic Review*, 2017, *107* (2), 425–456.

Cornell, Bradford and Ivo Welch, "Culture, Information, and Screening Discrimination," *Journal of political Economy*, 1996, *104* (3), 542–571.

Corno, Lucia, Eliana La Ferrara, and Justine Burns, "Interaction, stereotypes, and performance: Evidence from South Africa," *American Economic Review*, 2022, *112* (12), 3848–3875.

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

Daniel, Sherae, Ritu Agarwal, and Katherine J Stewart, "The effects of diversity in global, distributed collectives: A study of open source project success," *Information Systems Research*, 2013, *24* (2), 312–333.

de Sousa, José and Muriel Niederle, "Trickle-Down Effects of Affirmative Action: A Case Study in France," *NBER working paper*, 2022.

Ductor, Lorenzo, Sanjeev Goyal, and Anja Prummer, "Gender and Collaboration," *The Review of Economics and Statistics*, 2021, pp. 1–40.

Evsyukova, Yulia, Felix Rusche, and Wladislaw Mill, "LinkedOut? A Field Experiment on Discrimination in Job Network Formation*," *The Quarterly Journal of Economics*, 10 2024, p. qjae035.

Ewens, Michael, "Race and gender in entrepreneurship," *Review Article*, 2022.

Fu, Feng, Martin A Nowak, Nicholas A Christakis, and James H Fowler, "The evolution of homophily," *Scientific reports*, 2012, *2* (1), 845.

Gousios, Georgios, "The GHTorrent dataset and tool suite," in "Proceedings of the 10th Working Conference on Mining Software Repositories" MSR '13 IEEE Press Piscataway, NJ, USA 2013, pp. 233–236.

Grabchak, Michael, Eric Marcon, Gabriel Lang, and Zhiyi Zhang, "The Generalized Simpson's Entropy is a Measure of Biodiversity," *PloS one*, 2017, *12* (3), e0173305.

Hegde, Deepak and Justin Tumlinson, "Does social proximity enhance business partnerships? Theory and evidence from ethnicity's role in US venture capital," *Management Science*, 2014, *60* (9), 2355–2380.

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

Hong, Lu and Scott E. Page, "Groups of diverse problem solvers can outperform groups of high-ability problem solvers," *Proceedings of the National Academy of Sciences*, 2004, *101* (46), 16385–16389.

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

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

Lerner, Josh and Jean Tirole, "Some simple economics of open source," *The Journal of Industrial Economics*, 2002, *50* (2), 197–234.

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

LiCalzi, Marco and Oktay Surucu, "The Power of Diversity over Large Solution Spaces," *Management Science*, 2012, *58* (7), 1408–1421.

Mas, Alexandre and Enrico Moretti, "Peers at work," *American Economic Review*, 2009, *99* (1), 112–145.

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

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

Onuchic, Paula and Debraj Ray, "Signaling and discrimination in collaborative projects," *American Economic Review*, 2023, *113* (1), 210–252.

Oram, Andy and Greg Wilson, *Making Software: What Really Works, and Why we Believe it*, O'Reilly Media, Inc., 2010.

Ostrom, Elinor, "Beyond Markets and States: Polycentric Governance of Complex Economic Systems," *American Economic Review*, June 2010, *100* (3), 641–72.

Subramani, Gauri, Abhay Aneja, and Oren Reshef, "Persistence and the Gender Innovation Gap," 2021. Working Paper.

Tiebout, Charles M., "A Pure Theory of Local Expenditures," *Journal of Political Economy*, 1956, *64* (5), 416–424.

Tuomisto, Hanna, "A consistent terminology for quantifying species diversity? Yes, it does exist," *Oecologia*, 2010, *164* (4), 853–860.

Vasilescu, Bogdan, Daryl Posnett, Baishakhi Ray, Mark GJ van den Brand, Alexander Serebrenik, Premkumar Devanbu, and Vladimir Filkov, "Gender and tenure diversity

in GitHub teams," in "Proceedings of the 33rd annual ACM conference on human factors in computing systems" 2015, pp. 3789–3798.

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

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

Ytsma, Erina and Jana Gallus, "The Power of Public: Recognition and Reputation as Drivers of Open Source Success," in "2019 Meeting Papers" number 1141 Society for Economic Dynamics 2019.

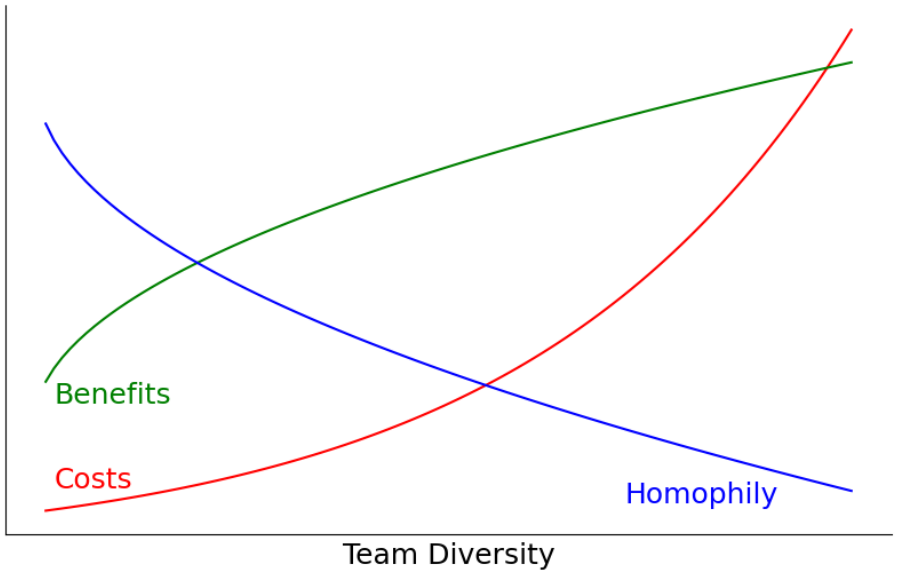


Figure 1: **Homophily and Team Production**. The figure shows the predicted benefits (green) and costs (red) from models of diversity in team production. The figure also shows the disutility of homophily for an outsider who joins the group (blue).

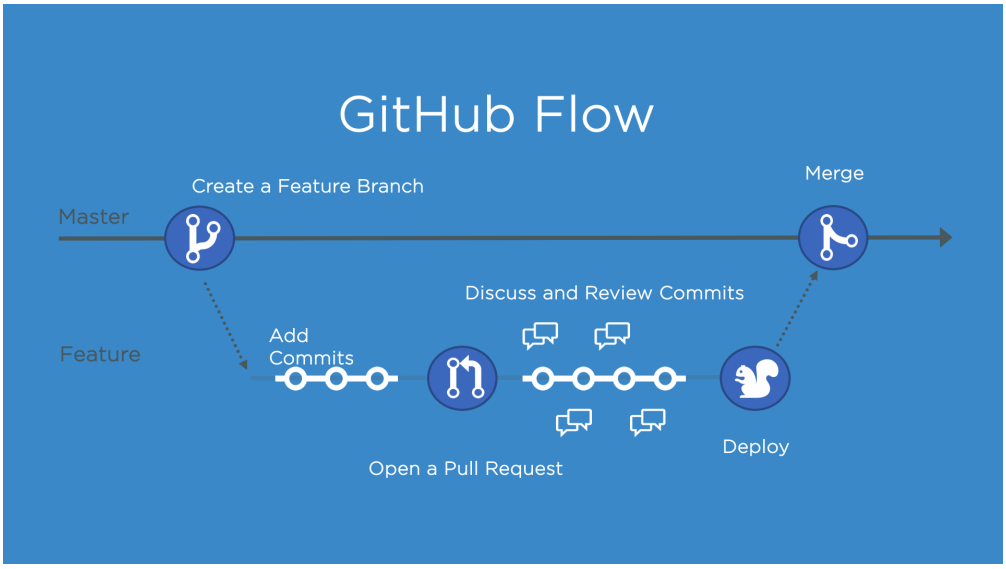


Figure 2: The figure shows a typical GitHub workflow, from https://github.com/SvanBoxel/

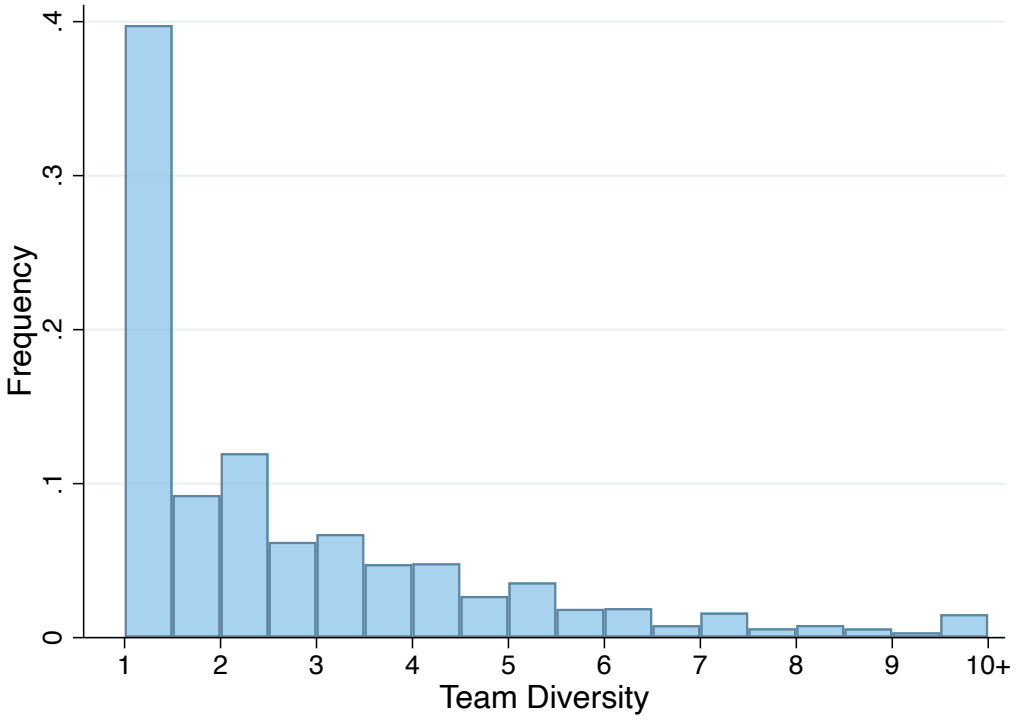


Figure 3: **Distribution of Diversity in OSS Teams**. The figure shows the histogram of team diversity for all project-years in the sample, where diversity is measured on the basis of coders' country of location ($TeamDiversity$).

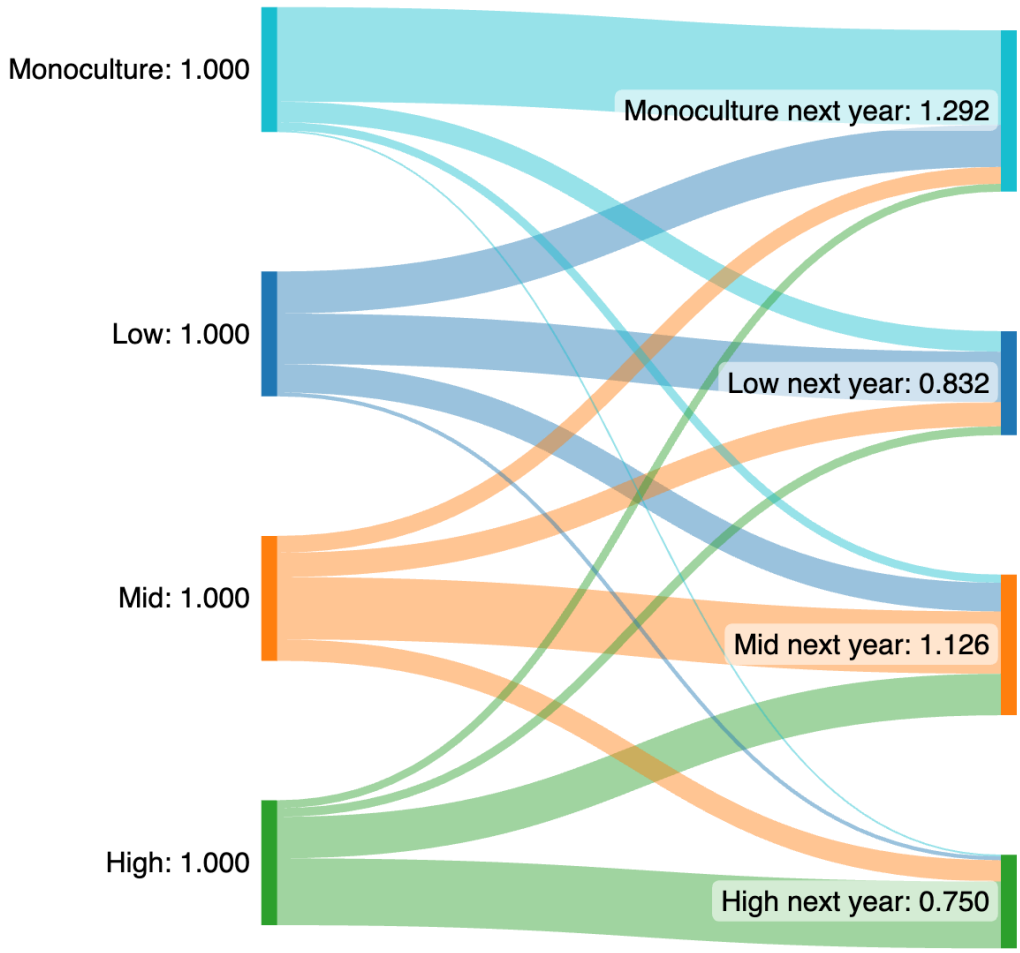
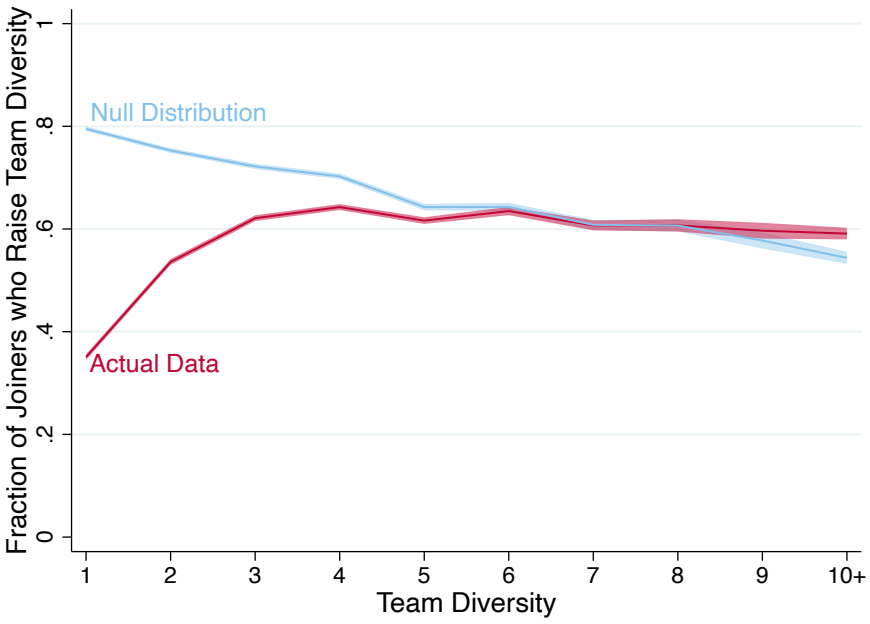
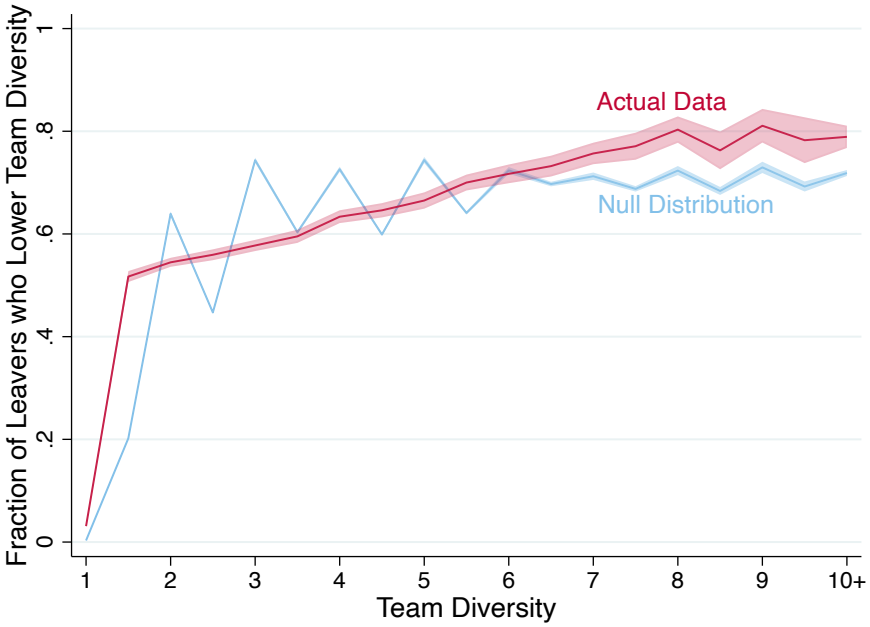


Figure 4: **Dynamics of Team Diversity**. The figure plots the transition probabilities between different levels of team diversity. The left hand side divides teams into bins by their $TeamDiversity$: Monoculture=1, Low$\in$(1,2], Mid$\in$(2,3], High$\in$(3,.). The right hand side plots the measure of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

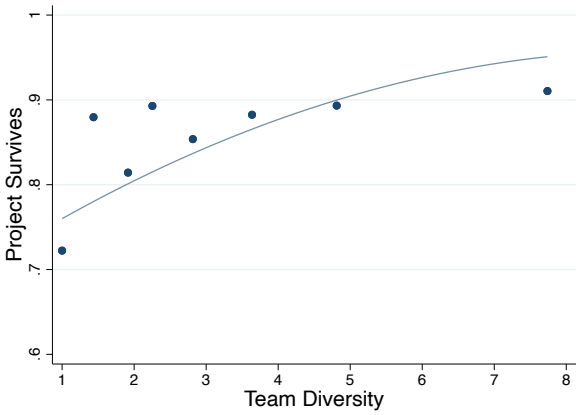


a) Outsiders joining

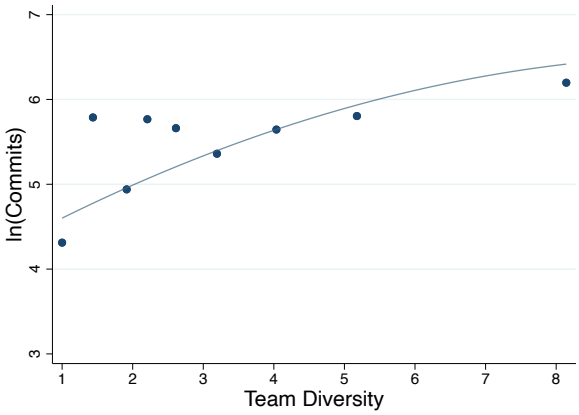


b) Outsiders leaving

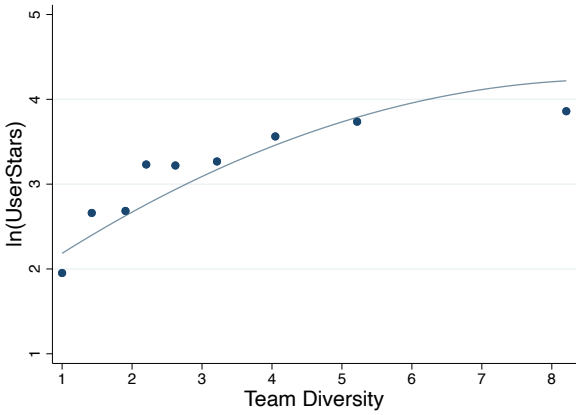
Figure 5: **Homophily in OSS Team Formation**. Figure (a) plots the fraction of new coders who join the project in a given year who are outsiders (raise the team's diversity by joining) sorted by the existing level of team diversity. Figure (b) plots the fraction of coders who leave the project in a 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

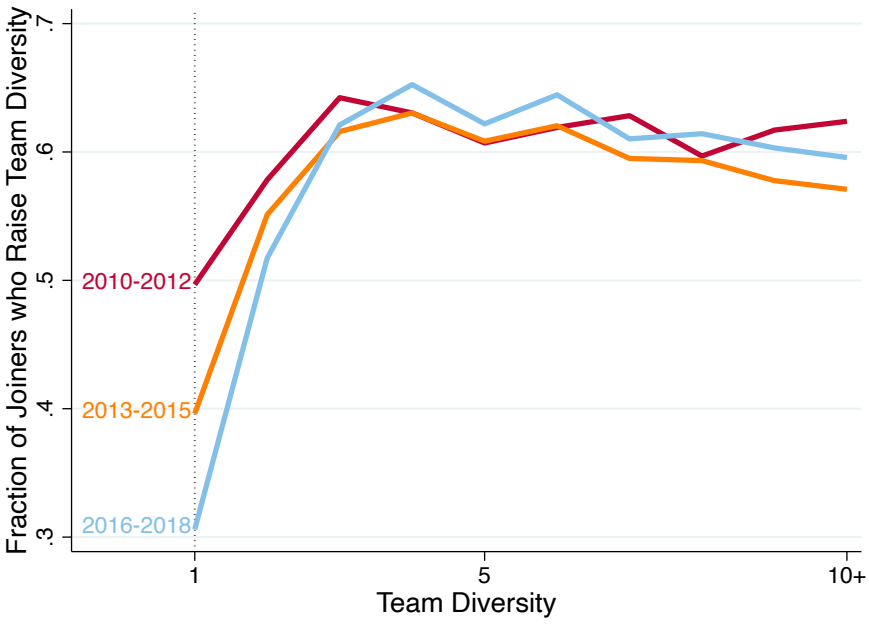


(b) Project development activity

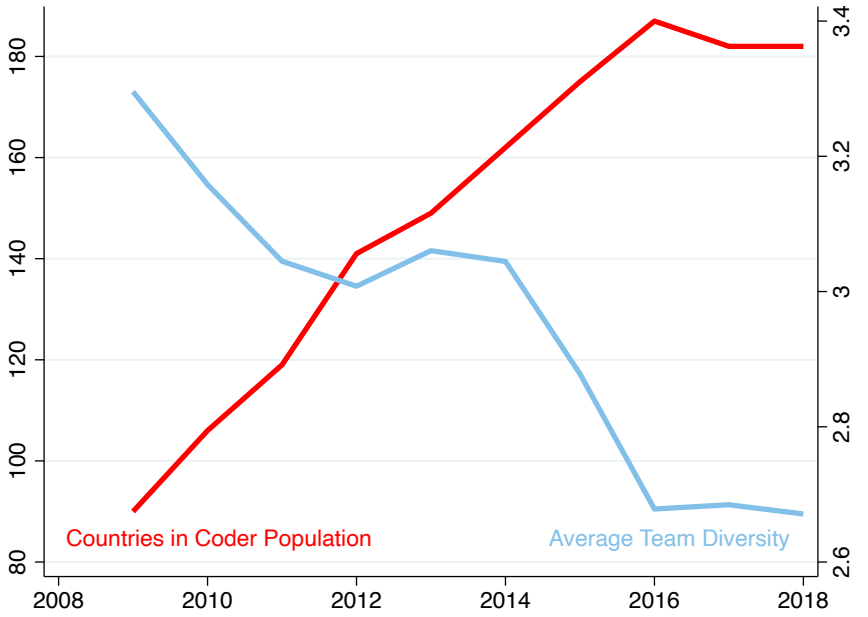


(c) Project popularity

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 $TeamDiversity$.

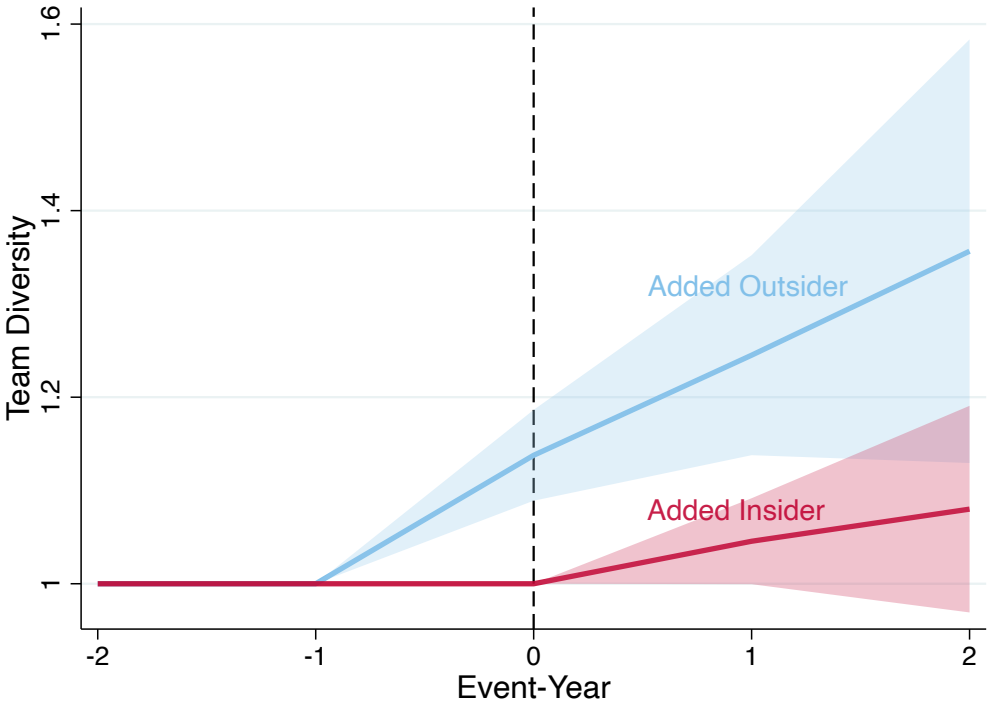


(a) Evolution of outsider join rates over time

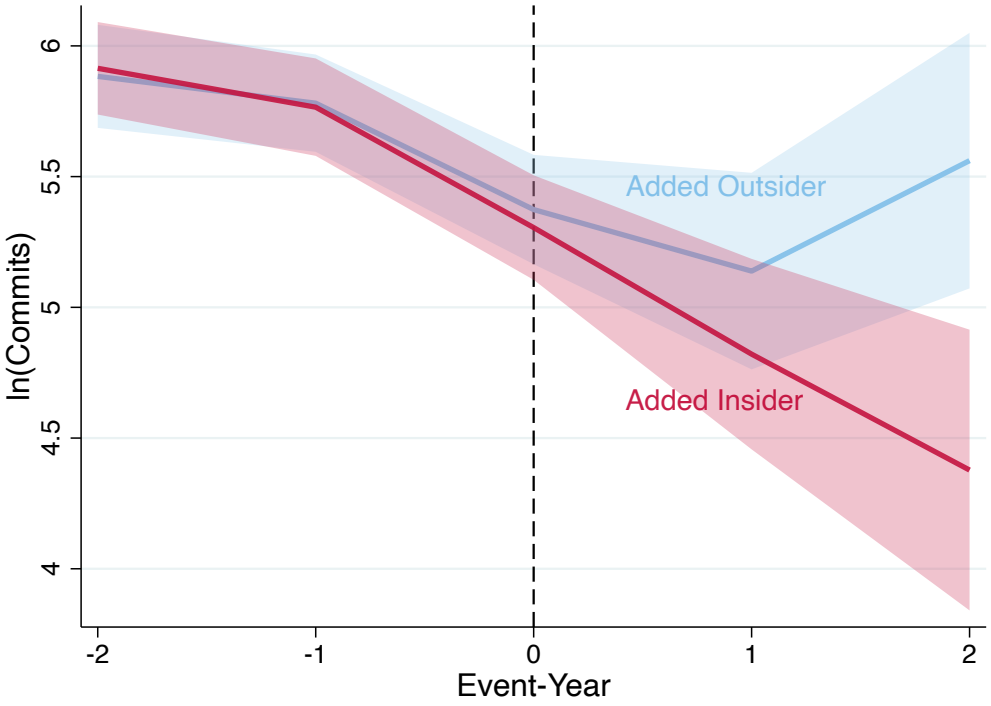


(b) Diversity of coder population and teams over time

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: **Summary Statistics.** 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
Ethnicities	5.96	2.72	3.00	5.00	10.00	143,524
DivEthnicities	3.02	1.64	1.00	2.78	5.40	139,446
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: **Team Diversity is Associated with Project Success** 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) $ProjectSurvives$	(2) $ln(Commits)$	(3) $ln(UserStars)$
$TeamDiversity$	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: **Broadband Access Increases OSS Participation.**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) N_{ict} OLS	(2) $ln(N_{ict})$ OLS	(3) N_{ict} 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 increases 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*			
TeamDiversity	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. = $ln(Commits)$			
TeamDiversity	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. = $ln(UserStars)$			
TeamDiversity	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. = $ProjectSurvives$			
$TeamDiversity_{it}$	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. = $ln(Commits)$			
$TeamDiversity_{it}$	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. = $ln(UserStars)$			
$TeamDiversity_{it}$	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