Teamwork and the Homophily Trap: Evidence from Open Source Software

Davidson Heath, Nathan Seegert, Jeffrey Yang

University of Utah

Introduction

Team Production

- Team production happens everywhere
- How to organize teams to maximize productivity?
- Left alone, do teams get to that optimum point?

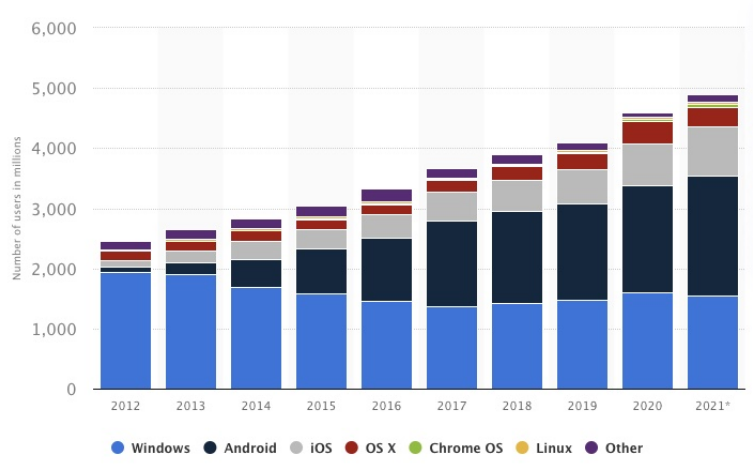
Diversity and Team Production

- Hong and Page (2004): A team faces a non-routine task
- The team pools ideas, then picks the best one
- A more diverse team generates a better best idea
    - ...benefits
- However, a more diverse team has higher communication and coordination costs
    - ...costs

This Paper

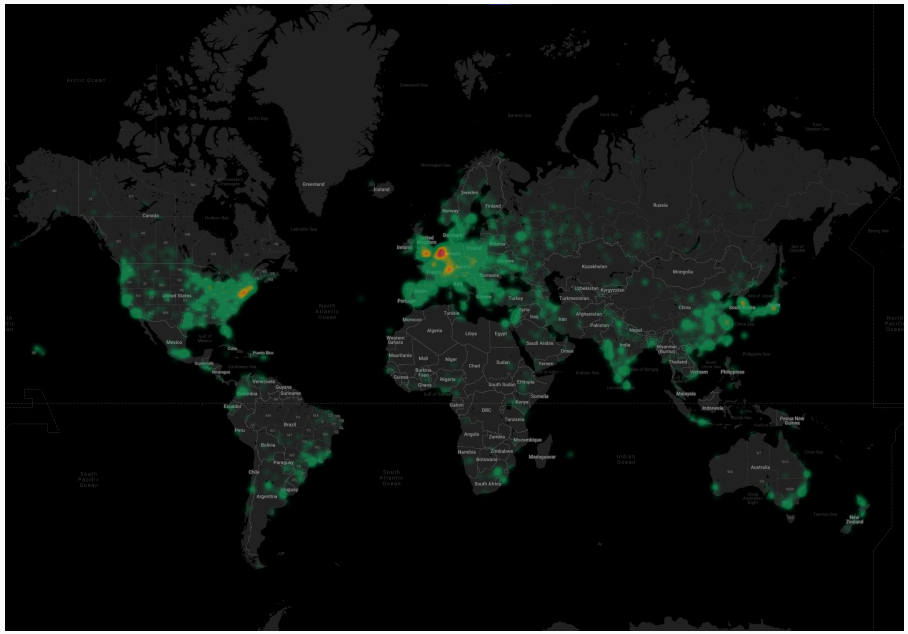
- Data are inconsistent with Hong+Page predictions
- Why?
- We argue **homophily** is a first-order behavioral phenomenon that's not in the Hong + Page model and its descendants
- Homophily has first-order consequences for team organization and policy

Open Source Software



- What is Open Source Software?
- Software whose source code is made available for anyone to copy & edit
- There are many successful OSS projects, which coexist with & even outcompete commercial software
    - Linux, Apache, Hadoop, Spark, R, LaTeX, Python

OSS



- Github provides amazingly granular data on OSS production
- 2008: 69,000 coders from 73 countries
- 2018: 1.1 million coders from 170 countries

Project Outcomes

Q: How to measure project outcomes?

- A1: Project survival
  - $= 0$ if the project has zero commits in this & subsequent years
- A2: Coding activity $=$ # of new commits
- A3: Popularity $=$ # of users who star the project

Team Diversity

- We construct a continuous measure of diversity:

$$TeamDiversity_{it} = \frac{1}{1 - X_{it}} \tag{1}$$

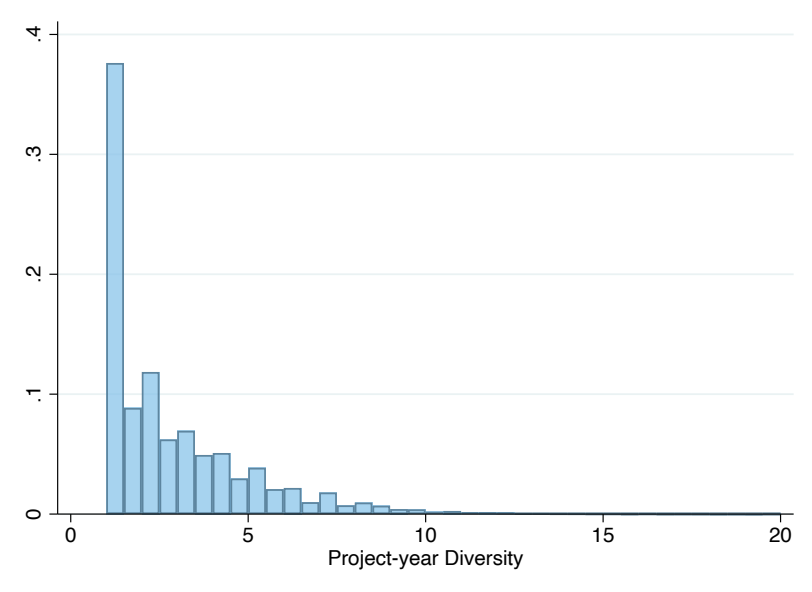
$$X_{it} = \sum_{c} p_{ict}(1 - p_{ict})$$

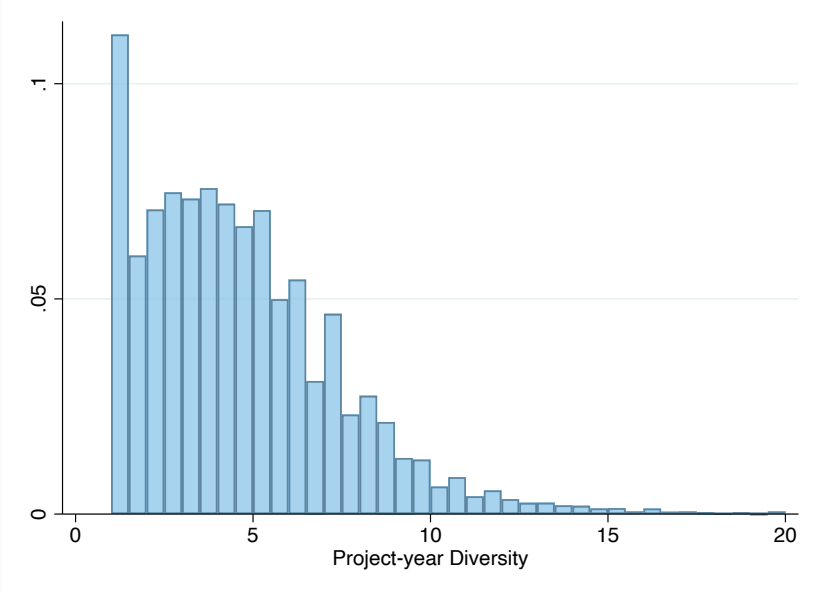
$$p_{ict} = \frac{N_{ict}}{\sum_{c} N_{ict}}$$

- *TeamDiversity* is a real-valued number between 1 and N
- Monotone transformation of racial HHI, Blau index

Stylized Facts

Distribution of Team Diversity



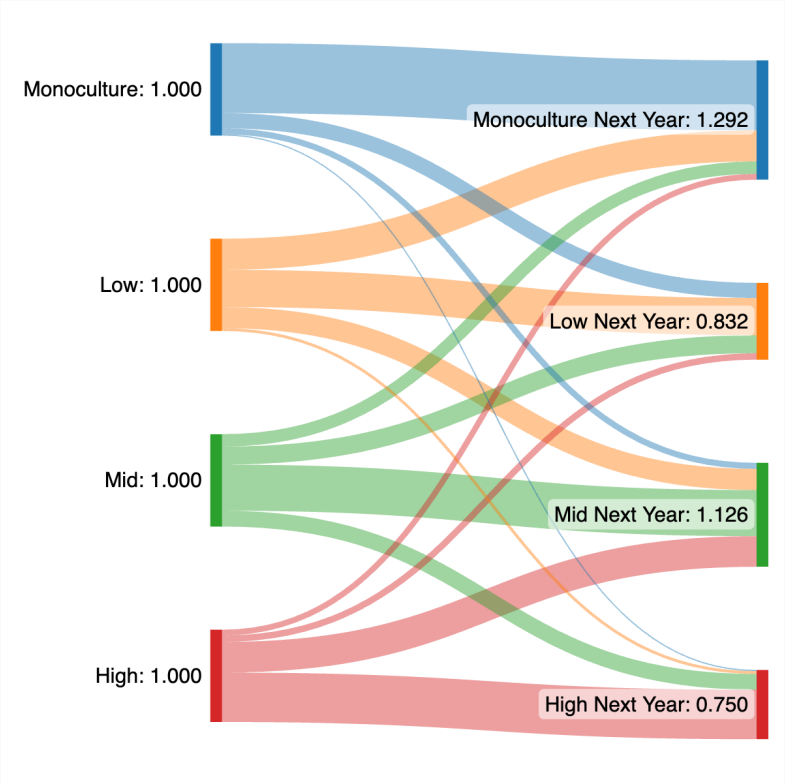


- Left is all teams; Right is all teams with $\geq$20 coders

Distribution of Team Diversity

- Team diversity is low relative to the population of coders (73 countries in 2008, 170 countries in 2018)
- Team diversity has a bimodal distribution with a "spike" at monoculture, and a "gap" above

Dynamics of Team Diversity



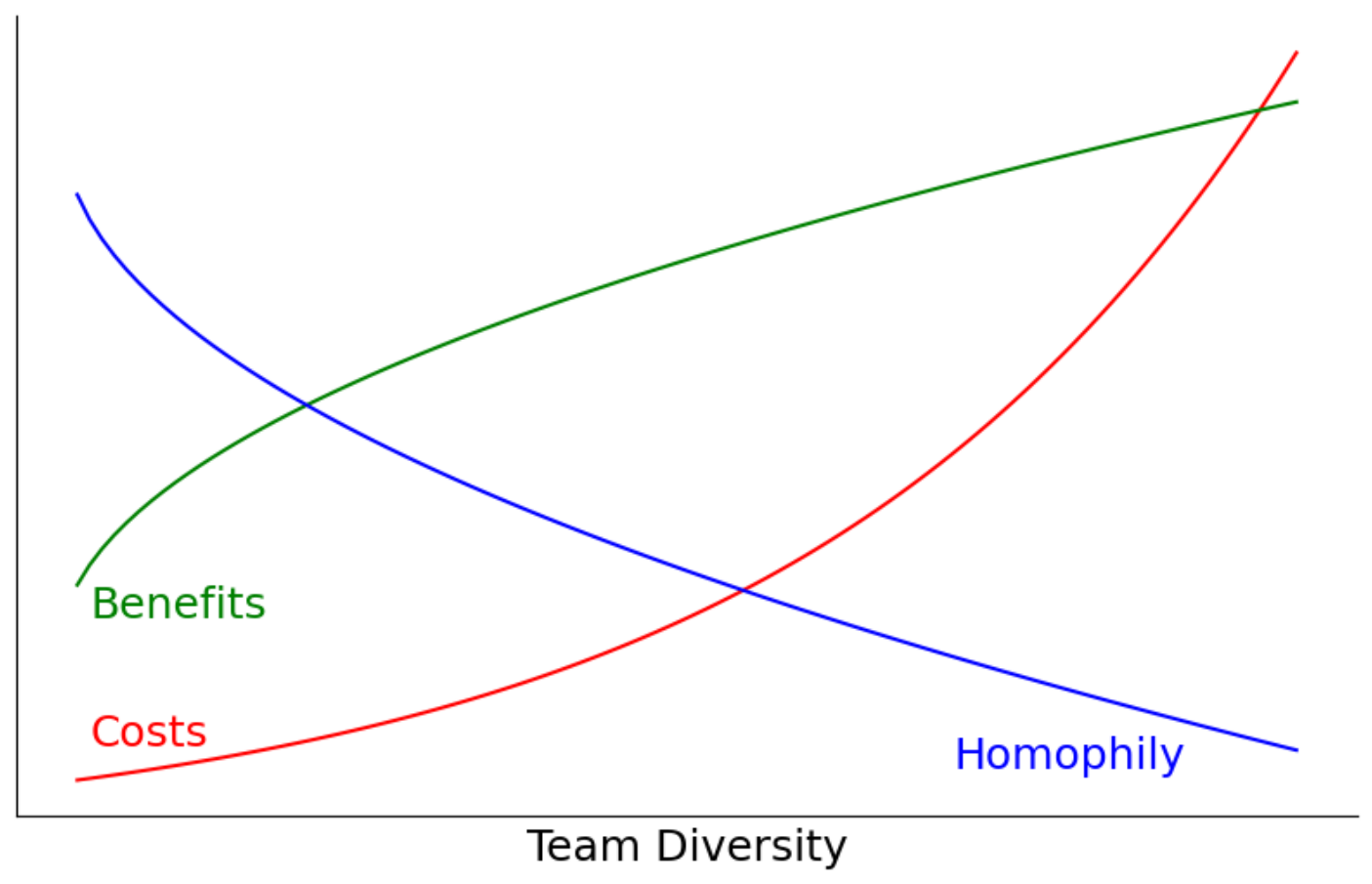
- Teams in the middle bucket either move up, or else down to monoculture

Homophily

A: Homophily

- Preference to join a team with other coders from the same country (ethnicity, language, gender)

A: Homophily



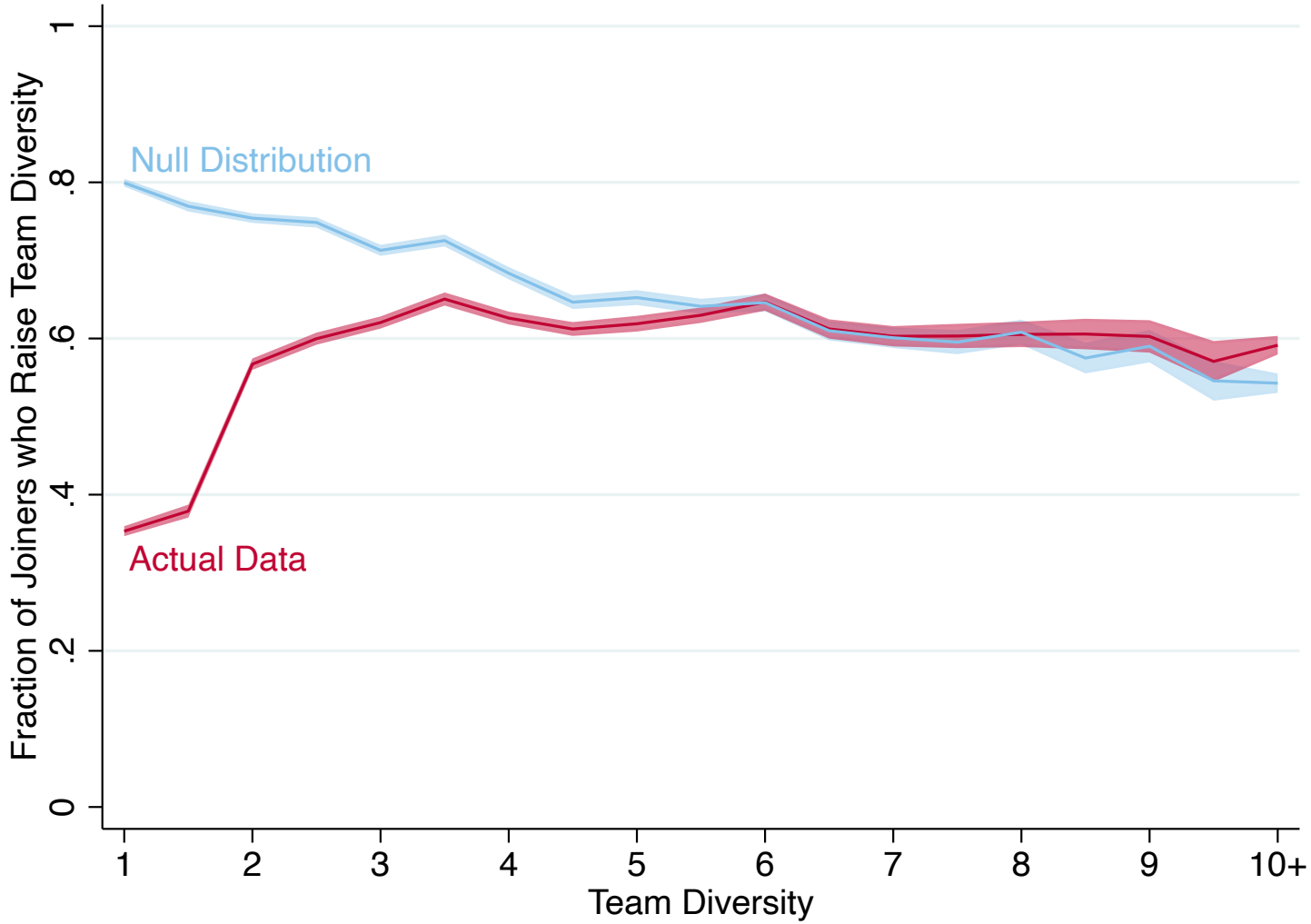
Two Key Features of Homophily

1. Homophily is strongest at *low* levels of diversity
    - The first outsider to join a monoculture pays a high cost
2. Homophily is a *private* preference
    - It has negative consequences on productivity which teams are not able to "internalize"

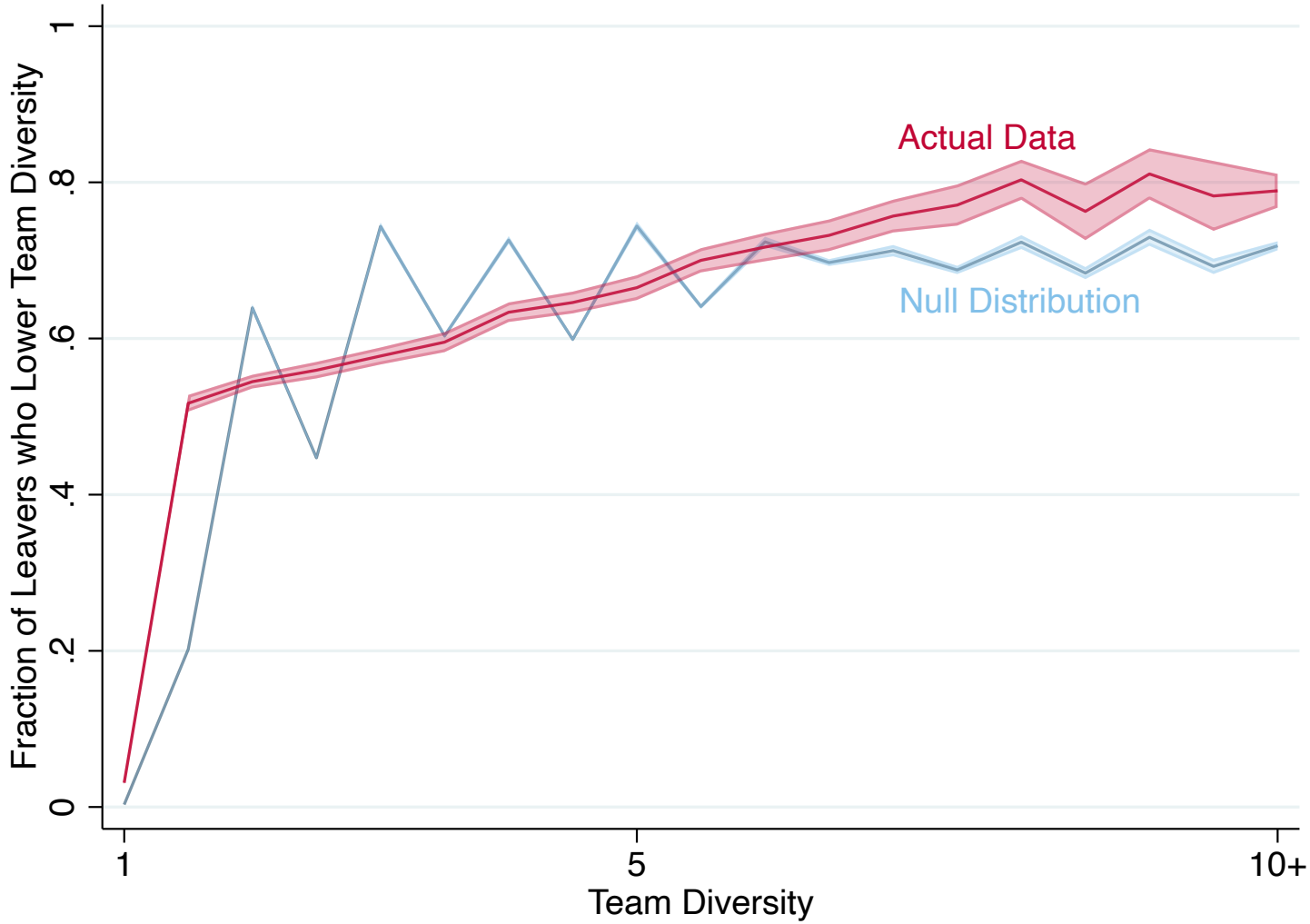
Homophily

- Can we observe homophily in team dynamics?
- We code an "outsider" as a team member who joins (leaves) the team who raises (lowers) diversity by joining (leaving)
- We simulate null (no-homophily) distributions of outsider joining rates and outsider leaving rates, by shuffling join- and leave-events within each year.

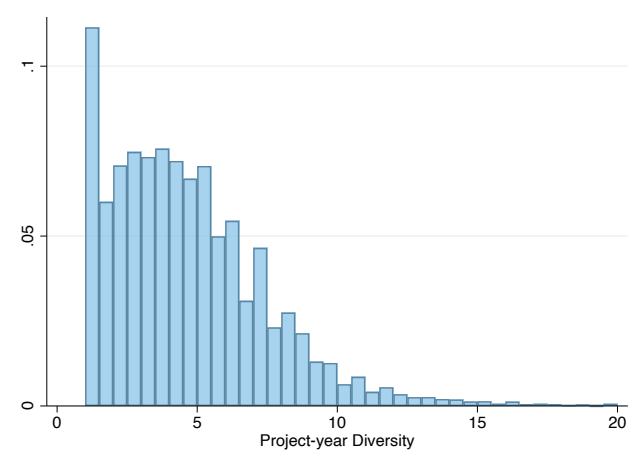
Coders who Join

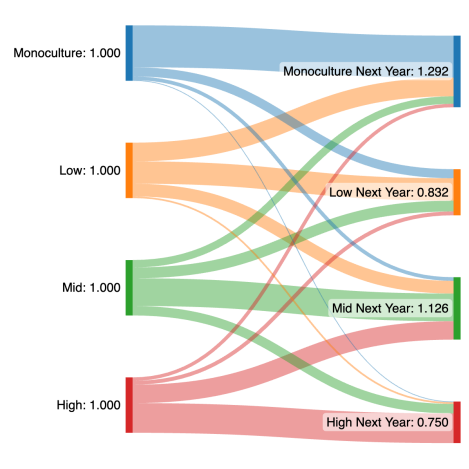


Coders who Leave



The Homophily Trap





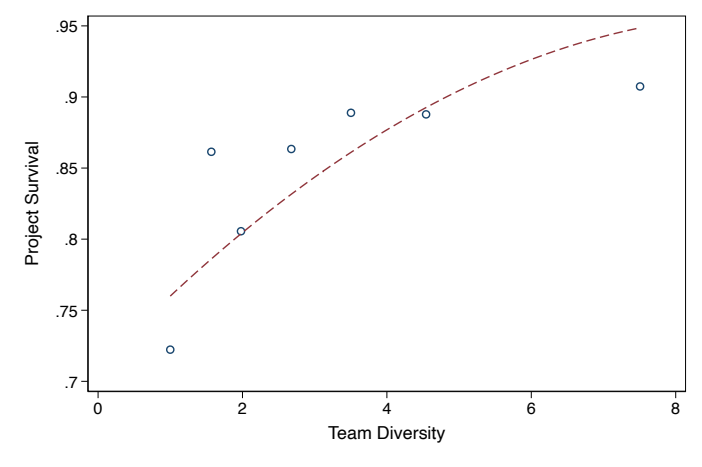
- Reality: Outsiders are *less* likely to join a low-diversity team
  - H0: The other way around!
- Suggests multiple equilibria and a "homophily trap"

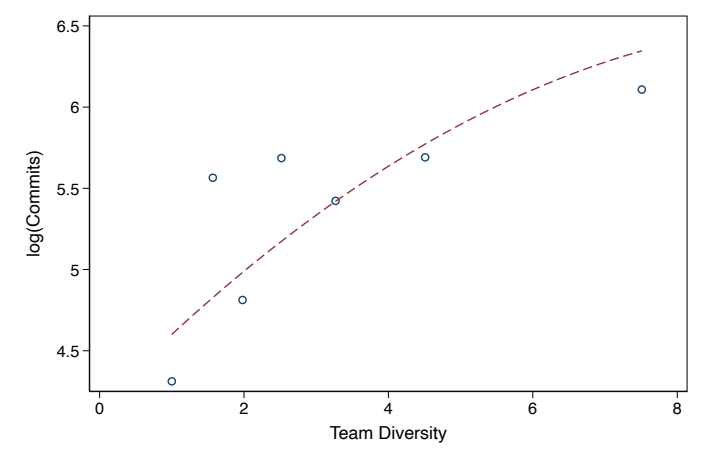
The Homophily Trap

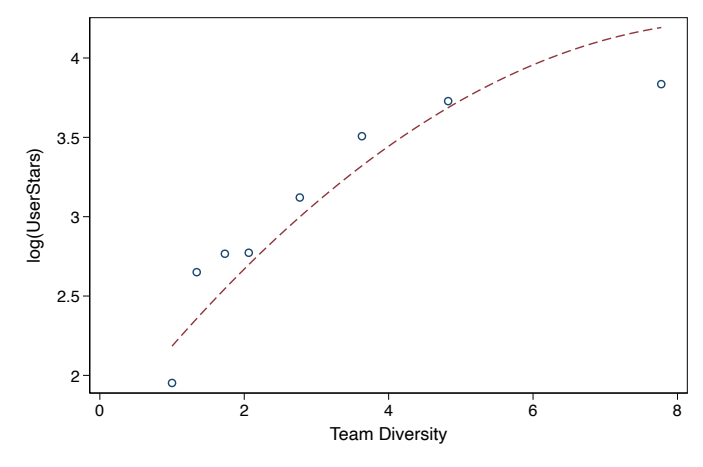
Testing for a Homophily Trap

- H0: Observed levels of diversity are efficient, most teams should have zero or negative marginal benefits
- The diagnostic for a homophily trap: Marginal benefits of team diversity are positive...
- ... & largest at **low** levels of team diversity

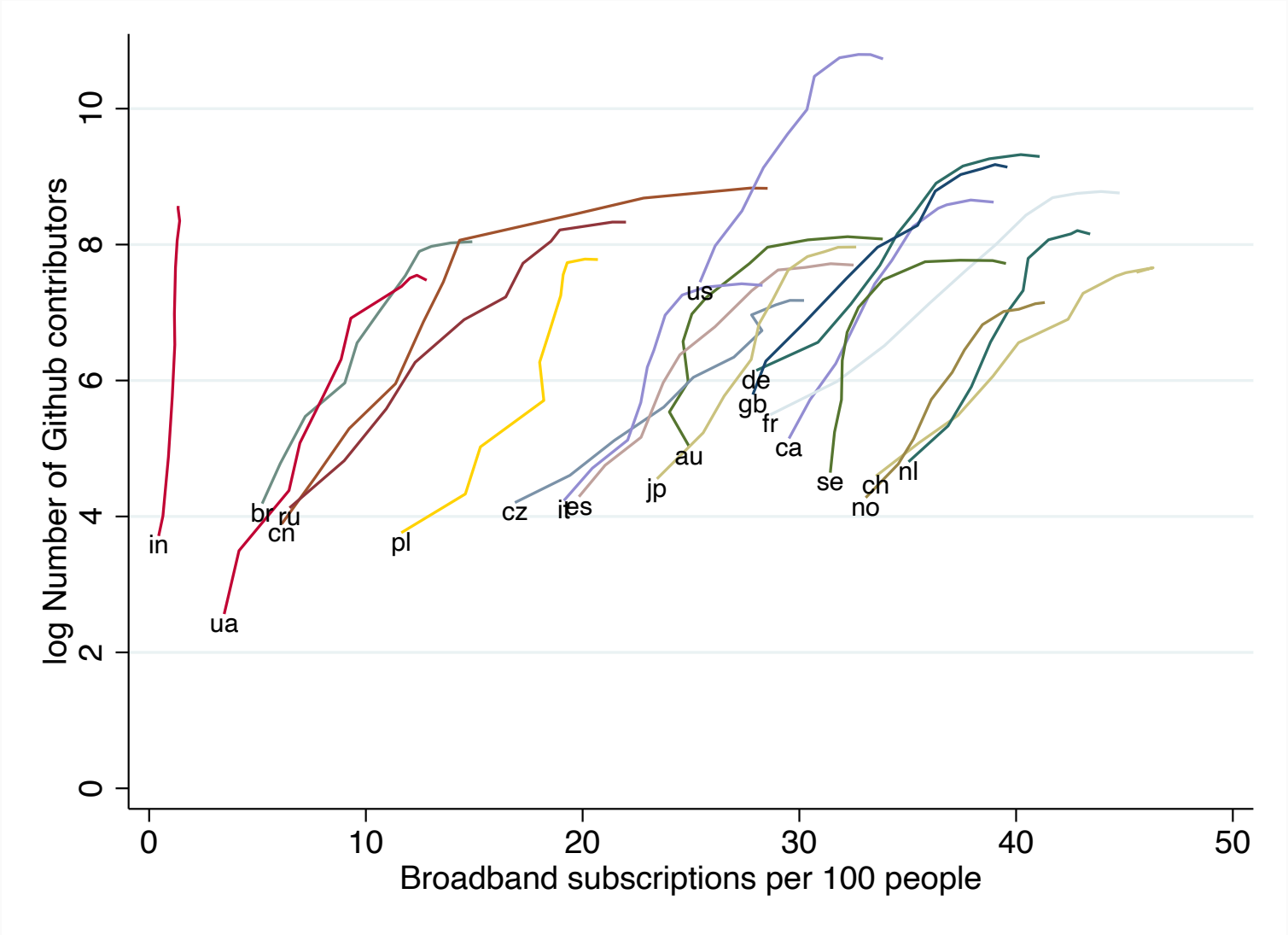
Outcomes are Concave in Team Diversity







IV Design


log Number of Github contributors
0
2
4
6
8
10
0
10
20
30
40
50
Broadband subscriptions per 100 people
in
ua
br
ru
cn
pl
cz
it
es
us
jp
au
de
gb
fr
ca
se
ch
no
nl

Marginal Benefits Higher at Low-Diversity Teams

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

Marginal Benefits Higher at Low-Diversity Teams

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

Marginal Benefits Higher at Low-Diversity Teams

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

Marginal Benefits Higher at Low-Diversity Teams

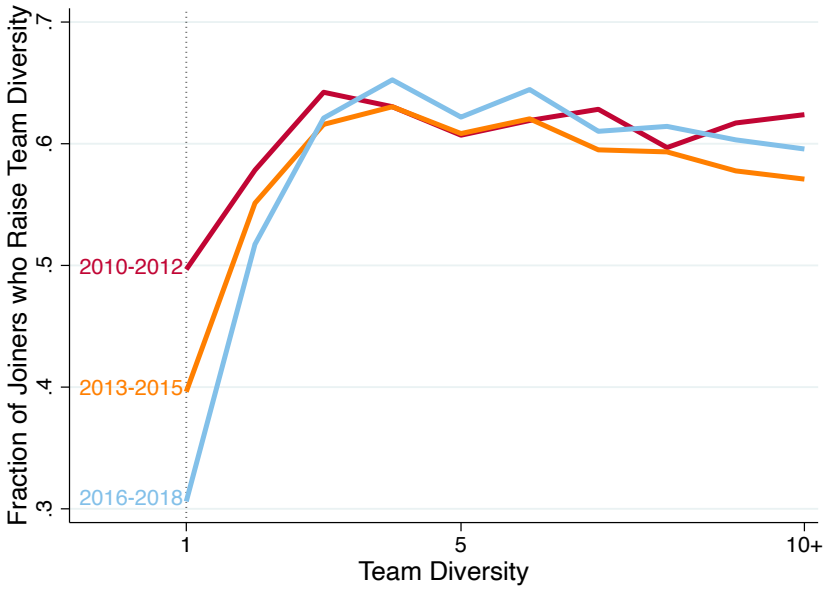
- Marginal benefits of diversity are robustly positive, for teams at high and low levels of diversity
- Highest for teams in monoculture
- We argue this is *diagnostic* of a homophily trap
- $\Rightarrow$ Low observed levels of diversity are suboptimal

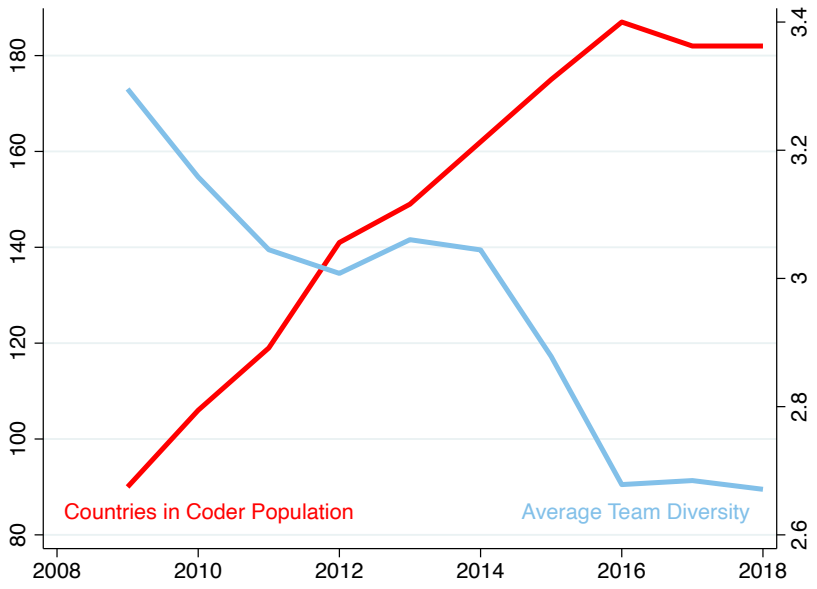
Policy

Policy 1: Sorting Effect

- In the presence of homophily, broadening the contributor pool gives you more similar peers to sort with (Tiebout, Buchanan)
- Over the sample period 2008-2018, the OSS contributor pool became much larger and more diverse

Policy 1: Sorting Effect



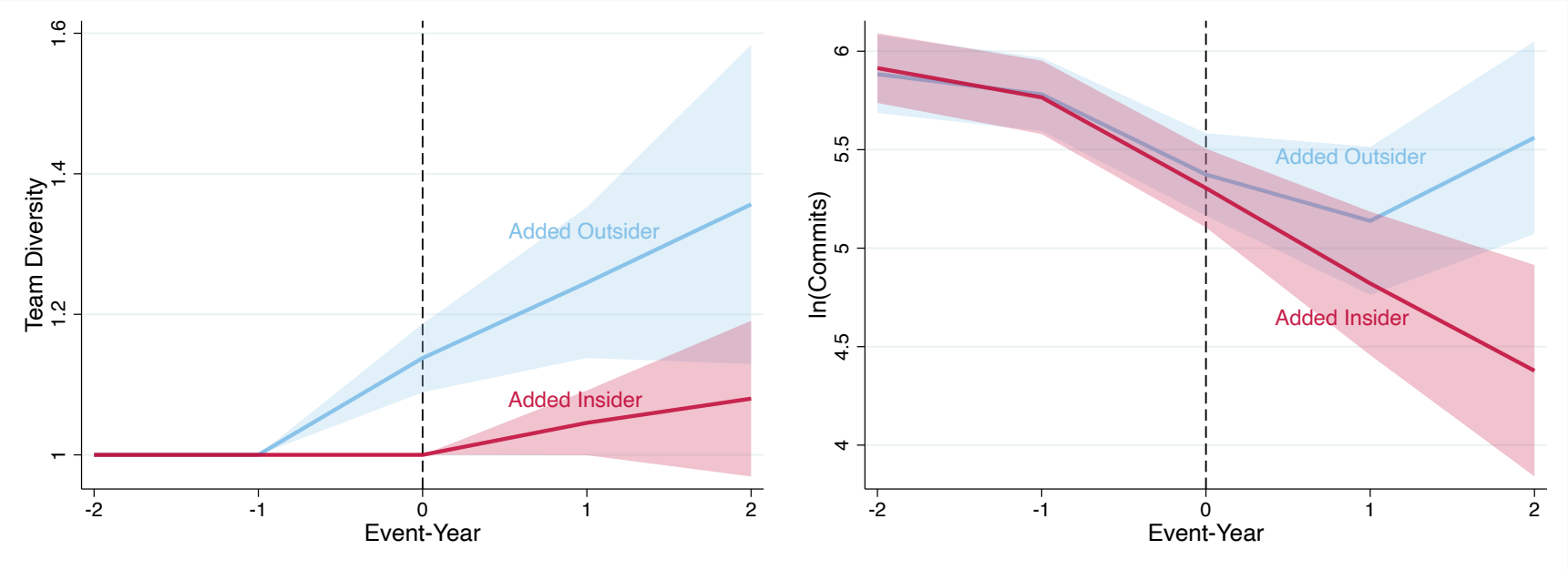


- With homophily, increasing diversity of the participant pool actually **lowers** team diversity
- Good from a preference standpoint; bad from a welfare standpoint

Policy 2: Trickle-Down Effect

- Homophily is a private preference which limits diversity, outsiders don't want to join
- Teams do not fully internalize this preference and so end up in a suboptimally low-diversity state
- Suggests that policies to recruit outsiders into low-div teams can pay off (de Sousa & Niederle, WP)
- We **match** teams in monoculture on lagged observables
- Blue team added one outsider; Red team added one insider

Policy 2: Trickle-Down Effect



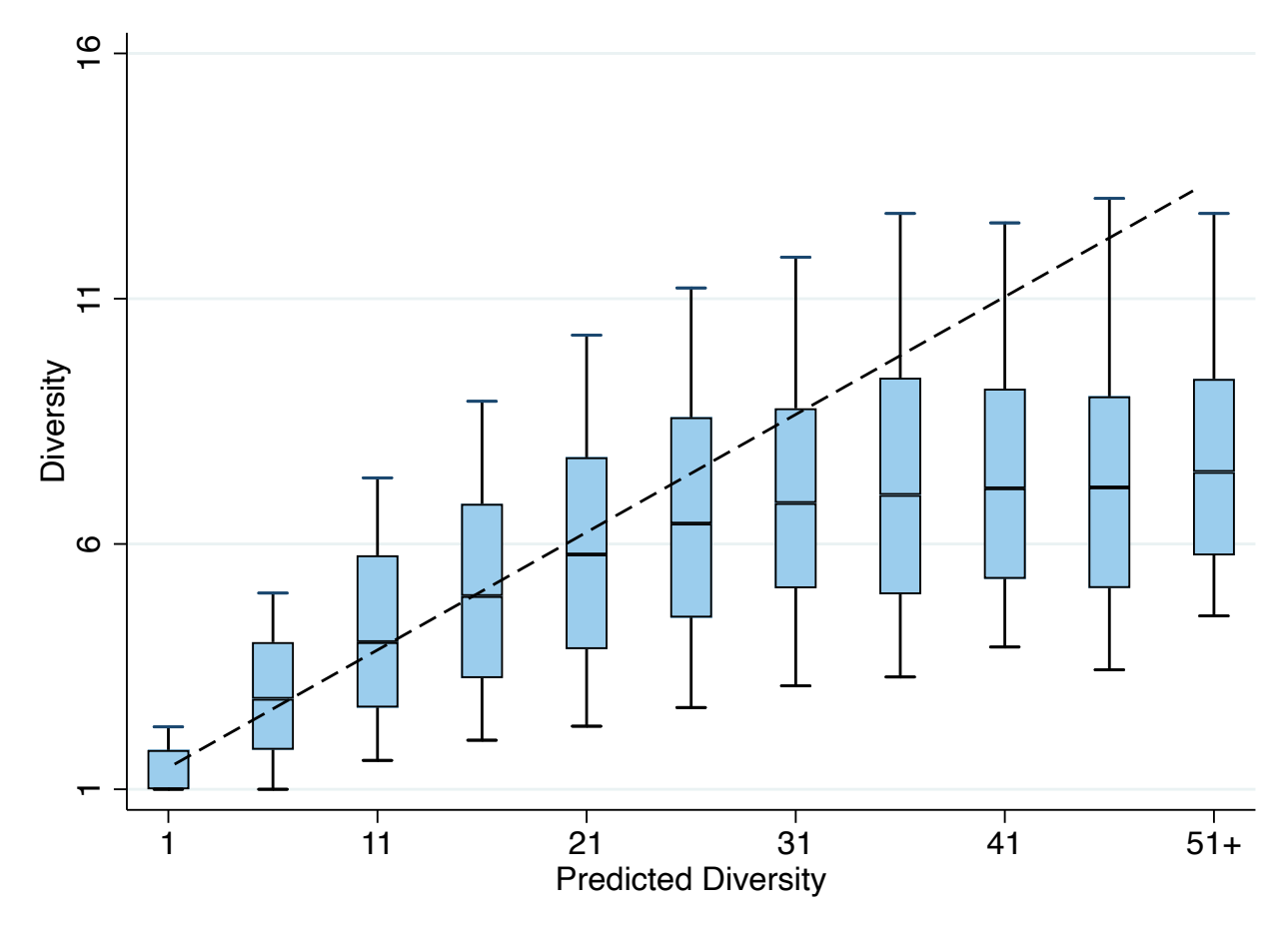
- Gap widens over time
- + Diversity, survival, activity, popularity

Conclusion

Conclusion

- Homophily $\Rightarrow$ team diversity has public benefits, private costs
- Outsiders are *less* likely to join low-diversity teams
- In equilibrium, diversity is too low relative to the social or even project-level optimum
- Policies to expand the candidate pool can backfire
- Policies targeted at low-diversity teams can have large payoffs, raise both diversity and productivity

Monotonicity in 2nd Stage



Back

Instrument works in both directions

	(1) First Stage $Diversity_{it}$	(2) IV $ProjectSurvives_{i,t+1}$	(3) IV $log(Commits)_{it}$	(4) First Stage $Diversity_{it}$	(5) IV $ProjectSurvives_{i,t+1}$	(6) IV $log(Commits)_{it}$
$EffectiveCountries_{it}$	0.211***			1.006***		
	(0.004)			(0.001)		
$\widehat{EffectiveCountries}_{it}$		0.027***	0.593***		0.027**	0.283***
		(0.003)	(0.013)		(0.013)	(0.050)
Subsample	$\hat{EC} > EC$	$\hat{EC} > EC$	$\hat{EC} > EC$	$\hat{EC} <= EC$	$\hat{EC} <= EC$	$\hat{EC} <= EC$
Observations	96,705	74,038	96,705	11,261	8,174	11,261
Project FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes

Back

Popularity

	(1) OLS	(2) OLS	(3) IV	(4) IV
	Dep. Var. $= ln(UserStars)_{it}$			
$EffCountries_{it}$	0.026***	0.024***	0.157***	0.143***
	(0.002)	(0.002)	(0.009)	(0.009)
$Coders_{i,t-1}$		0.000		0.000
		(0.000)		(0.000)
$TotalCommits_{i,t-1}$		0.000***		0.000***
		(0.000)		(0.000)
$ProjectAge_{it}$		0.361***		0.369***
		(0.066)		(0.068)
Observations	60,952	60,952	60,768	60,768
Adjusted R-squared	0.979	0.980		
Project FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Back

Userbase Diversity

	(1) OLS	(2) OLS	(3) IV	(4) IV
	Dep. Var. = $DivUsers_{it}$			
$EffCountries_{it}$	0.013**	0.013**	0.122***	0.130***
	(0.006)	(0.006)	(0.021)	(0.022)
$Coders_{i,t-1}$		-0.000		-0.000*
		(0.000)		(0.000)
$TotalCommits_{i,t-1}$		0.000		-0.000
		(0.000)		(0.000)
$ProjectAge_{it}$		0.327**		0.336**
		(0.141)		(0.143)
Observations	60,952	60,952	60,768	60,768
Adjusted R-squared	0.927	0.927		
Project FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Back

Coder Retention

	(1) OLS	(2) OLS	(3) IV	(4) IV
	Dep. Var. = $FractionCodersStay_{i,t+1}$			
$EffCountries_{it}$	-0.028***	-0.027***	-0.055***	-0.052***
	(0.001)	(0.001)	(0.002)	(0.002)
$Coders_{i,t-1}$		-0.000		-0.000
		(0.000)		(0.000)
$TotalCommits_{i,t-1}$		-0.000		0.000
		(0.000)		(0.000)
$ProjectAge_{it}$		-0.184***		-0.182***
		(0.020)		(0.019)
Observations	63,141	63,141	62,978	62,978
Adjusted R-squared	0.402	0.405		
Project FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Back

Project Forking

	(1) OLS	(2) OLS	(3) IV	(4) IV
	Dep. Var. = $HardFork_{it}$			
$EffCountries_{it}$	0.003***	0.003***	0.014***	0.013***
	(0.001)	(0.001)	(0.002)	(0.002)
$Coders_{i,t-1}$		0.000		0.000
		(0.000)		(0.000)
$TotalCommits_{i,t-1}$		0.000***		0.000**
		(0.000)		(0.000)
$ProjectAge_{it}$		-0.013***		-0.014***
		(0.005)		(0.005)
Observations	99,768	99,768	99,514	99,514
Adjusted R-squared	0.227	0.227		
Project FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes

Back