

## Success and geography: Evidence from open-source software

Gábor Békés (CEU, HUN-REN KRTK, and CEPR), Julian Hinz (Bielefeld University and Kiel Institute for the World Economy) Miklós Koren (CEU, HUN-REN KRTK, CEPR and CESifo), Aaron Lohmann (Bielefeld University and Kiel Institute for the World Economy)

April 20, 2024<sup>1</sup>

---

<sup>1</sup>This work was funded by the European Union under the Horizon Europe grant 101061123. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.

# Introduction

## Research question

- How and where is open source software developed?
- Can spatially dispersed developers produce quality software?

GitHub poll

```
if (poll == "no") {
```

# Why Open Source Software (OSS)?

## OSS is huge

- Software industry – 1% of global GDP
- 90+% of software has open source components

## OSS is everywhere

OSS plays an important roles in

- Websites (PHP, JavaScript)
- Operating systems (Linux, Android)
- Data (R Tidyverse, Python Pandas, Julia)
- Machine Learning and AI (PyTorch, LLaMA)

## OSS is observable

# Collaboration is done mostly online

 **git-extras** Public

 Watch **214**

 Fork **1.2k**

 Star **16.6k**

 main

 3 Branches

 53 Tags

Add file

<> Code

## About

GIT utilities -- repo summary, repl, changelog population, author commit percentages and more

git

 Readme

 MIT license

 Activity

 16.6k stars

 214 watching

 1.2k forks

Report repository

## Releases 22

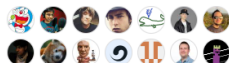
 **7.1.0 (Hauyne)** Latest  
on Oct 29, 2023


















+ 21 releases

## Packages

No packages published

## Contributors 224



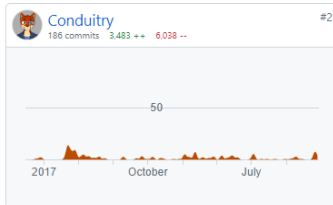
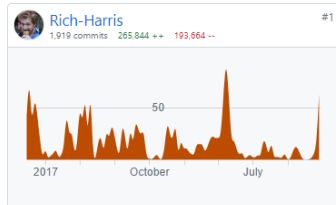
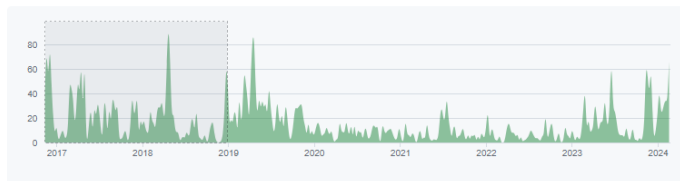
 <b>vanpipy</b>	test(browse-ci): add unit tests (#1130) ✓	5f19424 · 3 weeks ago	🕒 1,764 Commits
 .github	test(git-browse): add unit tests (#1127)	last month	
 bin	feat: add reverse option to git-brv (#1123)	2 months ago	
 etc	feat: add reverse option to git-brv (#1123)	2 months ago	
 helper	fix: No longer pollute env with GREP_OPTIONS	last year	
 man	feat: add reverse option to git-brv (#1123)	2 months ago	
 tests	test(browse-ci): add unit tests (#1130)	3 weeks ago	
 .editorconfig	Improve defaults for testing suite (#1104)	3 months ago	
 .gitignore	Improve defaults for testing suite (#1104)	3 months ago	
 .pytest.ini	test(git-authors): add unit test (#1098)	3 months ago	
 AUTHORS	maintenance: Add my name as maintainer in AUTHORS (#11...	3 months ago	
 CONTRIBUTING.md	chore: add poetry to handle the tests of the git extras (#1121)	3 months ago	
 Commands.md	feat: add reverse option to git-brv (#1123)	2 months ago	
 History.md	Version 7.1.0 (#1097)	4 months ago	
 Installation.md	Add more comprehensive dependencies (#1111)	3 months ago	
 LICENSE	Mention initial copyright year and add contributors to copyr...	9 years ago	
 Makefile	makefile: Allow bypassing conflict check (#1080)	5 months ago	

# Collaboration is done mostly online

Nov 13, 2016 – Dec 27, 2018

Contributions: Commits

Contributions to main, excluding merge commits





# Open Source vocabulary

**Package:** A unit of software, provision of a (bundle of) functionality

**Project:** A software project offering solution to a use case. Typically one package, but may be more.

**Repository:** A storage for one project (what we observe)

**Commit:** The smallest unit of contribution

**Git:** Distributed version control system for software projects

**GitHub:** A platform to collaboratively work on software projects

**Dependency:** An imported package that provides a functionality

}

## Related literature

- **Geographical Distance / Network formation / Agglomeration:**  
[@chaney2014network] [@bernard2019production] [@davis2019spatial]  
[@BaileyGuptaHillenbrandEtAl2021], [@Atkin\_2022\_F2F]
- **Gravity: Digital:** [@blum2006does] [@anderson2018dark]
- **Frictions in services:** [@stein2007longitude] [@bahar2020hardships]
- **Patents and science:** [@BircanJavorcikPauly2021],  
[@head\_li\_minondo\_math\_2019], [@jaffe1993geographic], Singh (2008)  
[@AlShebli\_nature\_2018], [@Li2014-patents-eer]
- **OSS:** [@lerner2002some], [@Laurentsyevea:2019] [@Wachs\_etal\_2022]  
[@fackler\_hofmann\_laurentsyevea\_2023]

# Data

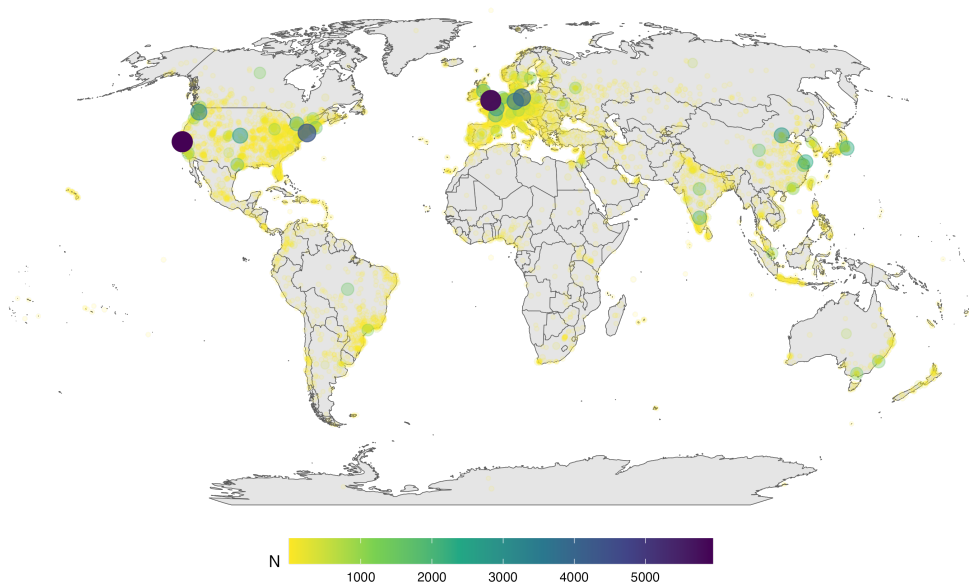
## GitHub

Snapshot of all public repositories on GitHub on 2019-06-01. Six largest languages: JavaScript, Python, Java, Ruby, PHP, and C++. Drop smallest and largest projects. 4.4m projects, 2.7m users. Self-reported location for about 1/3 of users.

## libraries.io

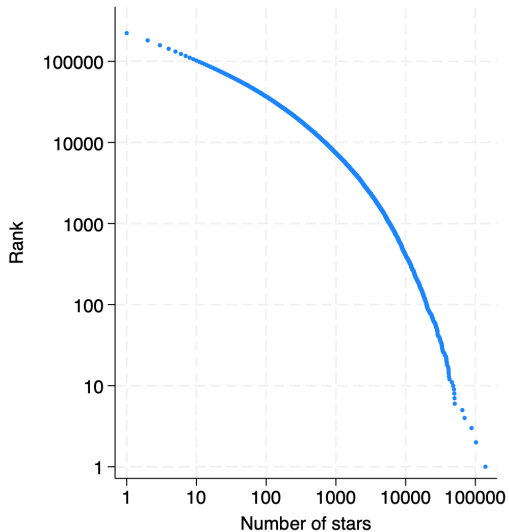
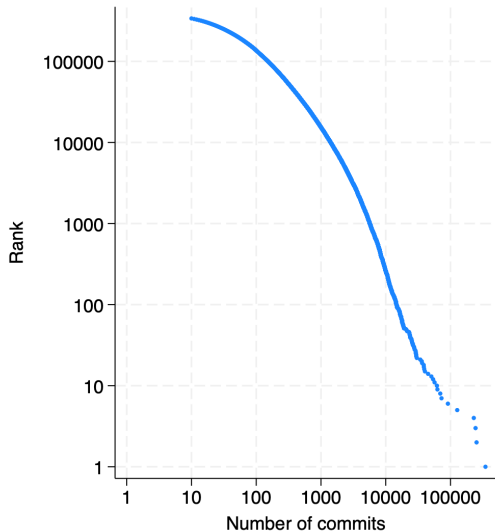
Dependency data for projects on major package managers (npm, PyPI, Maven, RubyGems, etc).

## Developer density around the globe

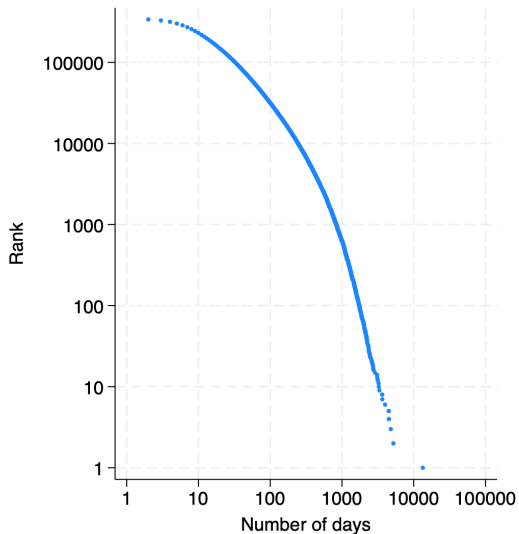
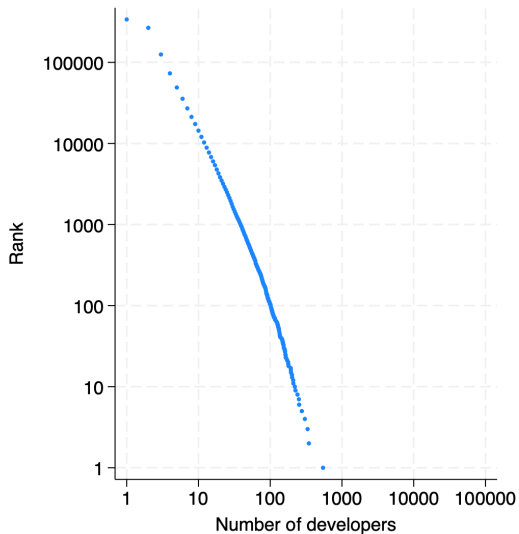


## Descriptives

## Project size and popularity

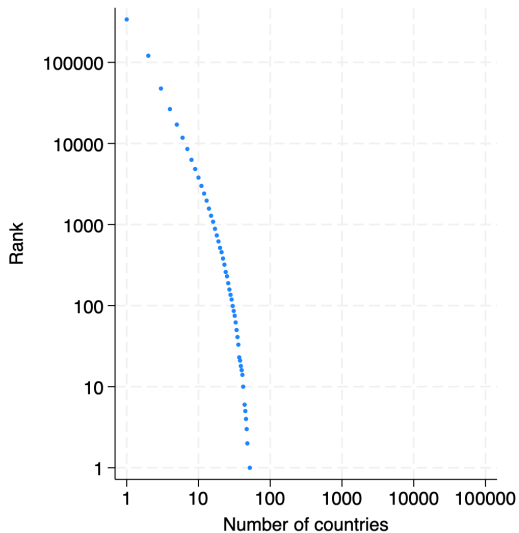
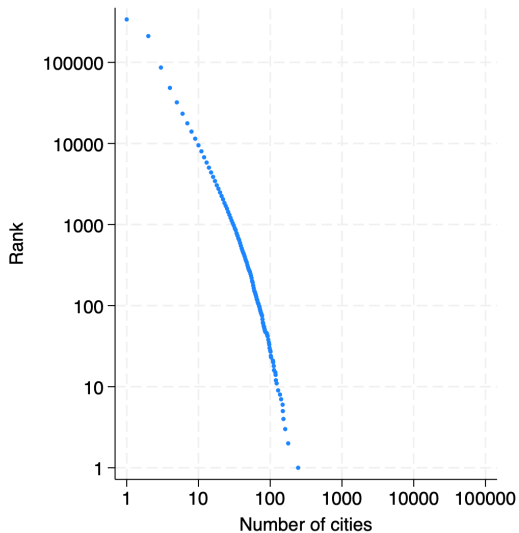


## Team size and total developer effort

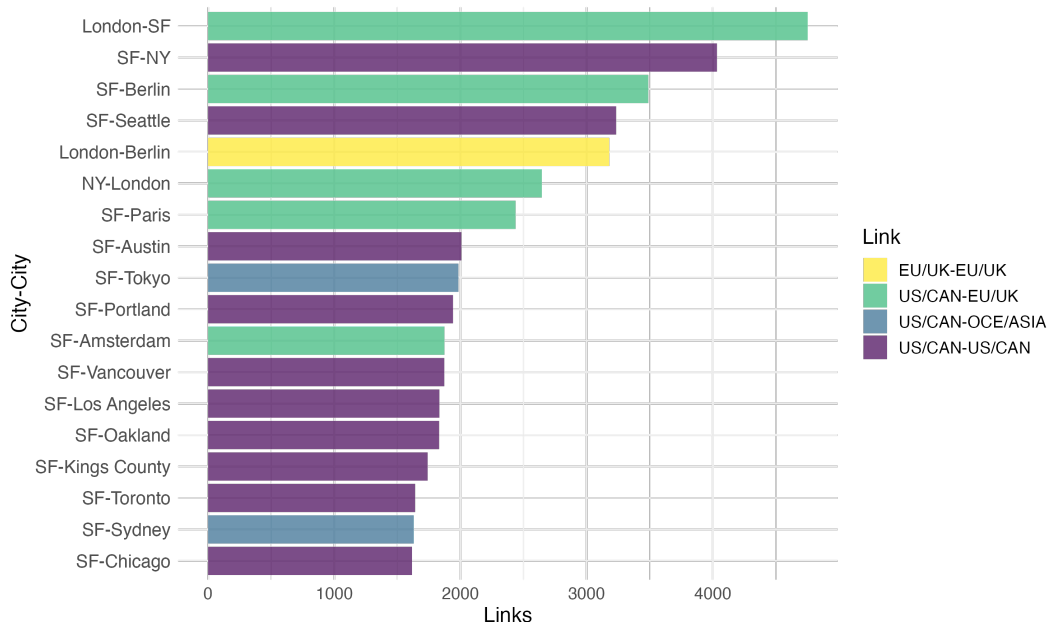




## Geographic diversity of teams



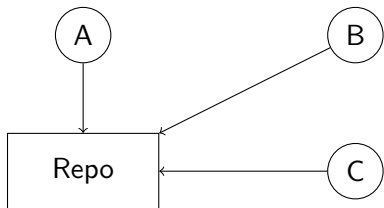
## Collaboration across cities is mostly North-North



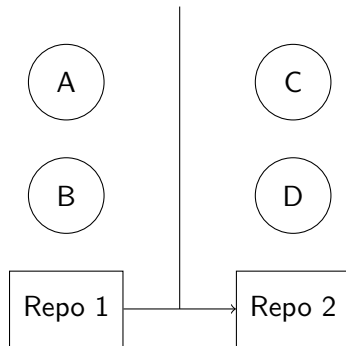
Most frequent city-pairs for repos developed from 2 cities

## Collaboration

## Measuring collaboration and dependencies

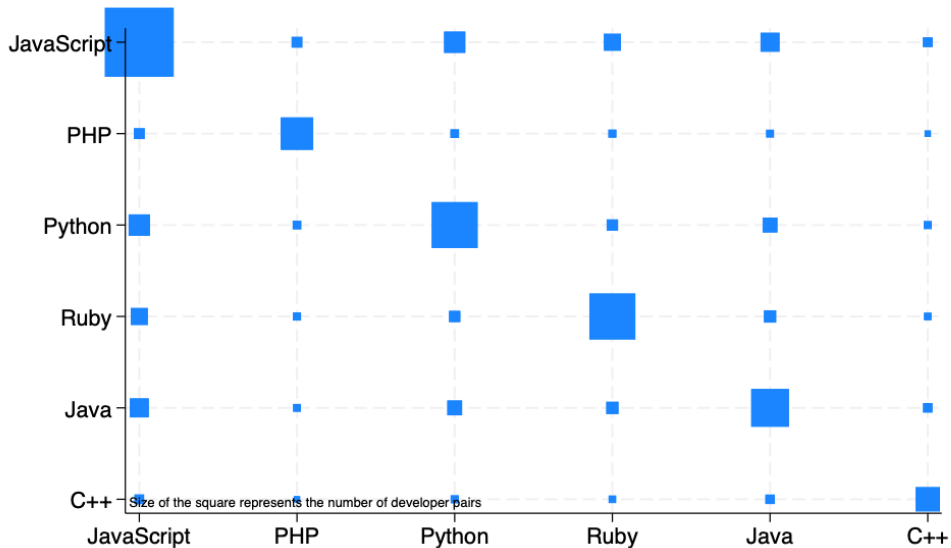


(a) Developers committing to a repository.

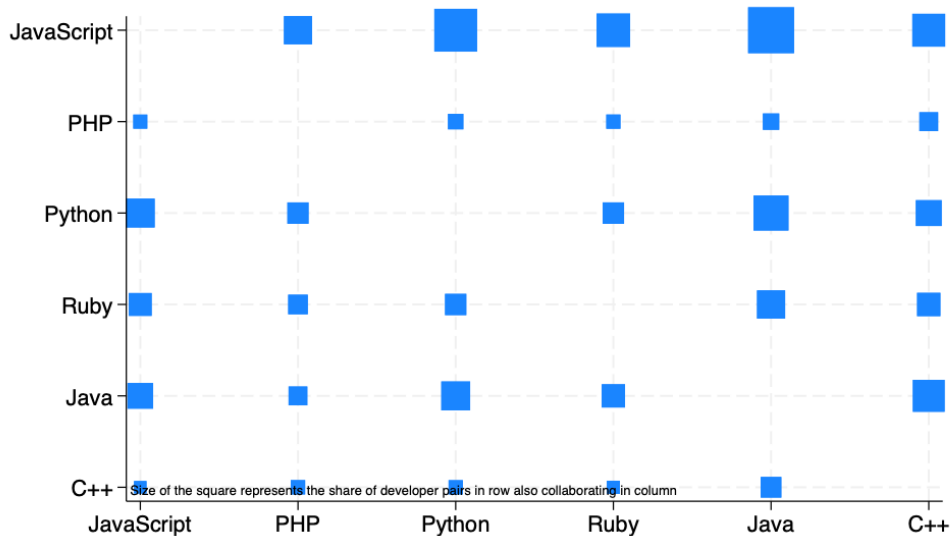


(b) Dependency of repository 1 on repository 2 with the respective developers.

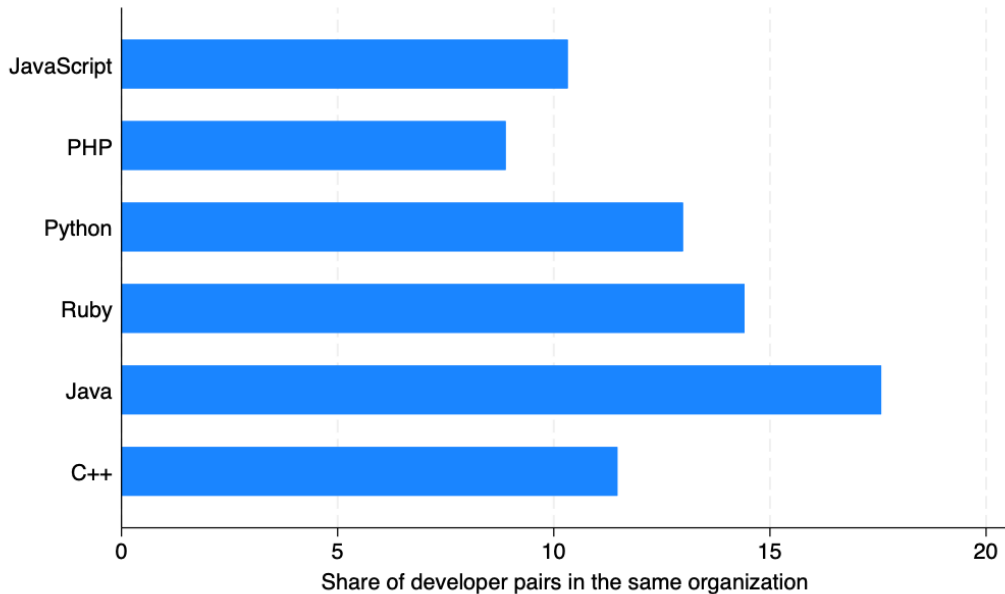
# Multiplexing



# Multiplexing



## Some work is done within organizations



## Collaboration in space



## Gravity model of collaboration

Developer  $i$  and  $j$  collaborate with probability

$$\Pr(\text{Collaboration}_{ij}) = \exp(\alpha_i + \beta_j - \gamma \times \text{distance}_{ij})$$

Aggregate across city pairs  $d$  and  $o$ :

$$E(N_{do}) = \exp(\tilde{\alpha}_d + \tilde{\beta}_o - \gamma \times \text{distance}_{do})$$

Estimate this with Poisson maximum likelihood.

# Four margins of collaboration

- 1 Committing to the same project
- 2 Commenting on the same issue
- 3 Editor rights on the same project
- 4 Members of the same organization

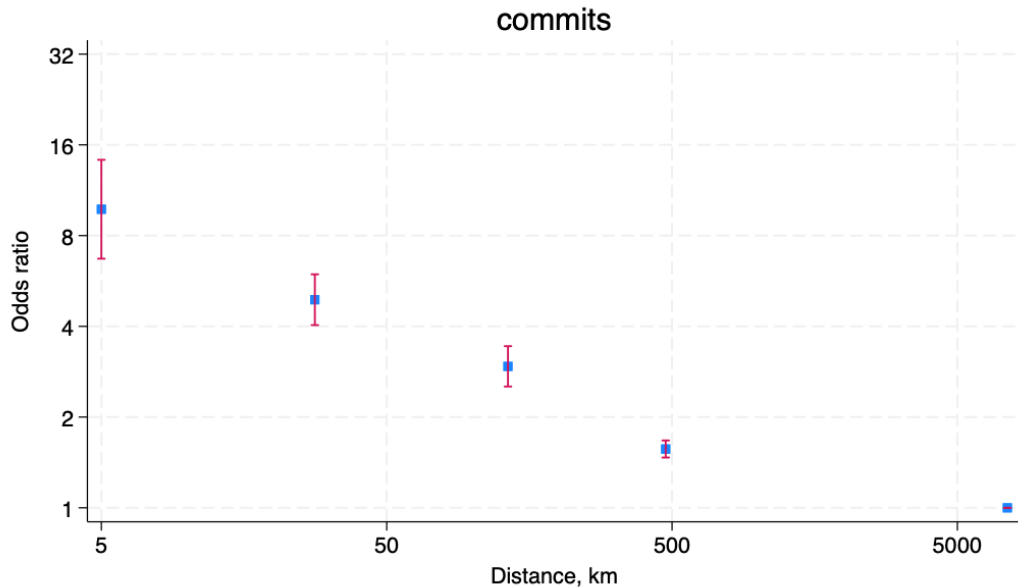
## Gravity model of the developer-to-developer network

VARIABLES	(1) commits	(2) comments	(3) editors	(4) organizations
Distance (log km)	-0.143*** (0.0129)	-0.107*** (0.0139)	-0.277*** (0.0178)	-0.183*** (0.0313)
Same country (dummy)	0.519*** (0.0528)	0.584*** (0.107)	1.058*** (0.0894)	0.706*** (0.135)
Same city (dummy)	0.846*** (0.0863)	0.433*** (0.0587)	0.856*** (0.0924)	0.350** (0.154)
Observations	498,472	781,201	139,464	408,094

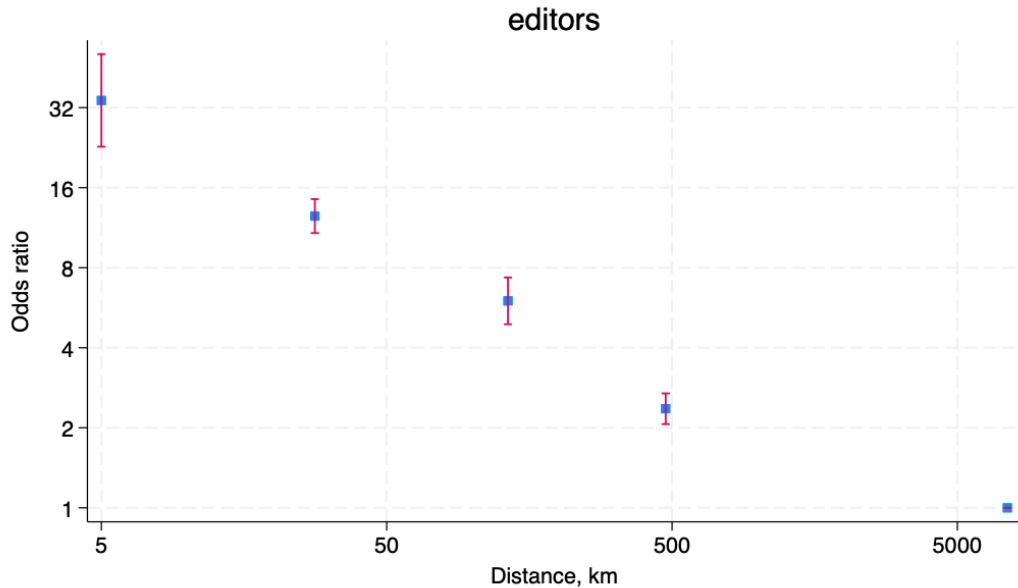
Robust standard errors in parentheses

\*\*\*  $p < 0.01$ , \*\*  $p < 0.05$ , \*  $p < 0.1$

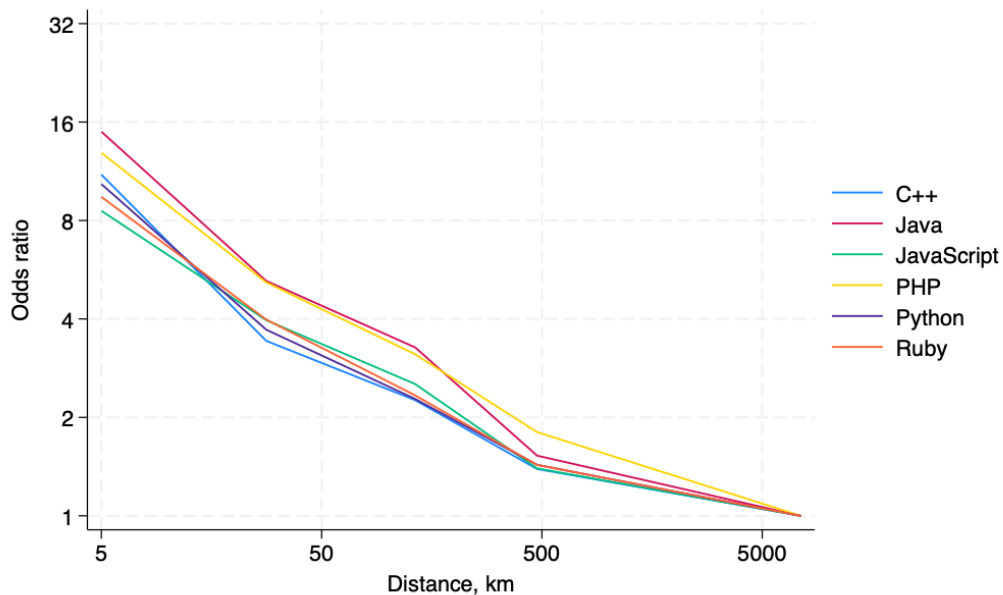
## Odds ratios of collaboration



## Editor rights are most localized



## Small differences across languages



Success

## Spatial diversity reduces amount of work

VARIABLES	(1) n_commits	(2) n_commits	(3) n_days	(4) n_days
avg_ln_distance	0.0409*** (0.00416)		0.0384*** (0.00271)	
ln_n_cities	-0.159*** (0.0302)		-0.218*** (0.0184)	
ln_n_countries	-0.244*** (0.0231)		-0.212*** (0.0111)	
ln_n_developers	1.241*** (0.0296)	0.935*** (0.0161)	1.257*** (0.0171)	0.929*** (0.00639)
share_organization		-0.236*** (0.0314)		-0.0602*** (0.00921)
share_different_cities		0.0534** (0.0255)		-0.0184* (0.0103)
share_different_countries		-0.124*** (0.0234)		-0.0955*** (0.0110)
Observations	267,086	267,086	267,086	267,086



## Spatial diversity associated with higher quality

VARIABLES	(1) n_stars	(2) n_stars	(3) n_downstream	(4) n_downstream
avg_ln_distance	0.198*** (0.0114)		0.299*** (0.0558)	
ln_n_cities	0.311*** (0.0780)		1.817*** (0.344)	
ln_n_countries	0.364*** (0.0447)		0.577*** (0.199)	
ln_n_developers	0.537*** (0.0797)	1.020*** (0.0134)	-1.428*** (0.320)	0.683*** (0.0409)
share_organization		-1.769*** (0.0446)		0.101 (0.147)
share_different_cities		0.944*** (0.0534)		4.144*** (0.768)
share_different_countries		0.731*** (0.0406)		1.217*** (0.184)
in_libraries = 0,			-	-