When dispersed teams are more successful: Theory and evidence from software

Gábor Békés*, Julian Hinz**, Miklós Koren*, Aaron Lohmann**

*CEU, KRTK and CEPR **University Bielefeld and IfW Kiel

Economics of open source

1. Why do people work for free? (literature in the early 2000s, not our main concern)
2. How do software teams form and collaborate in space? (This paper)

Why Open Source Software (OSS)?

- Software is everywhere and more specifically OSS is everywhere
  - 98% of commercial software uses OSS according to a report by Synopsis in 2023.
  - OSS is powering Machine Learning, AI development and embedded systems.
- OSS is huge
  - Hoffmann, Nagle, and Zhou (2024) estimate demand side as 8.8 triilion USD; GitHub nowadays has over 100 million developers
- OSS is observable
  - Due to the `git` paradigm almost everything is recorded!

What we see in the data: ggplot2-project as an example

Users living in cities

Figure 1: Hadley Wickham

are collaborating

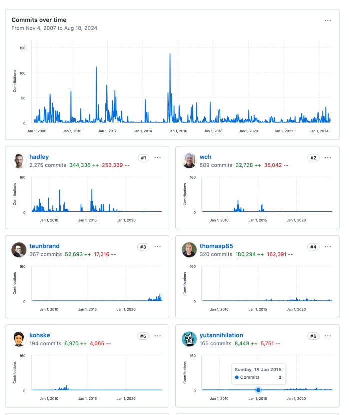


Figure 2: Commits in ggplot2

earning them fame.

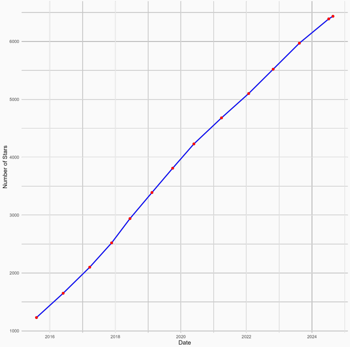


Figure 3: ggplot2 stars over time

Literature

- **Production in teams:** Jarosch, Oberfield, and Rossi-Hansberg (2021) ; Herkenhoff et al. (2024) ; Freund (2022) ; Kerr and Kerr (2018)
  Our contribution: A model for global team formation which has selection as a main mechanism.
- **Gravity/International Trade:** Eaton and Kortum (2002) ; Atkin, Chen, and Popov (2022) ; Head, Li, and Minondo (2019)
  Our contribution: Gravity estimates for team formation in OSS.
- **OSS**: Lerner and Tirole (2002) ; Fackler and Laurentsyeva (2020) ; Wachs et al. (2022)
  Our contribution: Providing more descriptive statistics, making use of data and combining several data sources.

Data

GHtorrent

Metadata from the GitHub (over 95 percent of OSS projects)

- $835,283$ projects.
- $347,767$ developers.
- over years from 2012 to 2019

Libraries.io

A data effort to collect upstream and downstream dependencies of OSS projects.

Analysis sample

- First quarter of each project.
- Developers who report their location.

Developers are globally dispersed

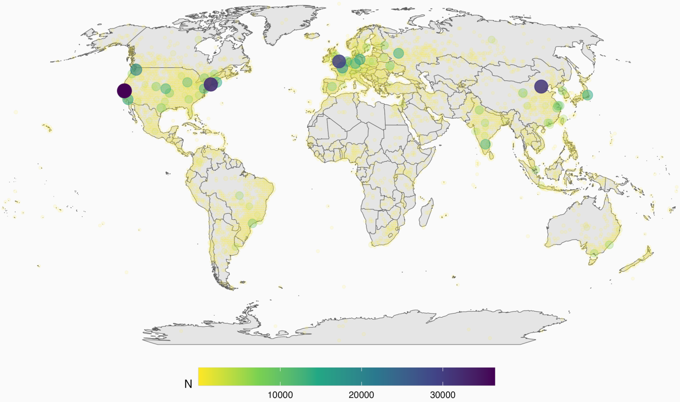


Figure 4: OSS developers around the world

Most developer teams are small

Number of Developers	Share
1	0.72
2	0.17
3	0.06
4	0.03
5	0.01

Table 1: Share of projects by number of developers.

- About 27% of projects are developed in collaborative teams.
- Team size follows a power-law like relationship.

Lots of “North-North” collaboration

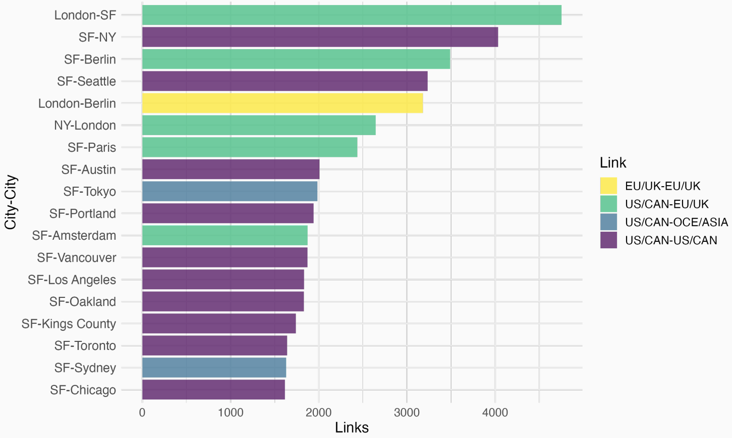


Figure 5: Pairwise collaboration between top cities in JavaScript language.

A model of global team formation

Features of OSS

- Developer differ in skills (partially observable).
- Team output is uncertain.
- Developers compete for “kudos.”

Endowments, technologies, and tastes

Developers have heterogenous skills Z_i which is drawn from a Fréchet distribution according to $\Pr(Z_i \leq x) = e^{-T_i x^{-\theta}}$

- observable skill T_i
- dispersion of unobserved skill $1/\theta$

Quality production function

The best idea determines software quality.

$$X_p = \max_{j \in p} \{Z_j / \tau_{jp}\}$$

Customer happiness

Overall customer happiness convex in software quality:

Frictions

Communication

Not all good ideas are heard (language, time zone, culture, clarity). $\tau_{ip} \geq 1$ iceberg cost of turning skills into ideas.

Participation

Not all benefits of distant projects can be captured (private cost of participation, time zones, misappropriate of credit). $d_{ip} \geq 1$ iceberg cost of turning kudos into utils.

Team formation

Attribution of kudos

Developer with the "winning idea" gets all the kudos for V_p.

Selection

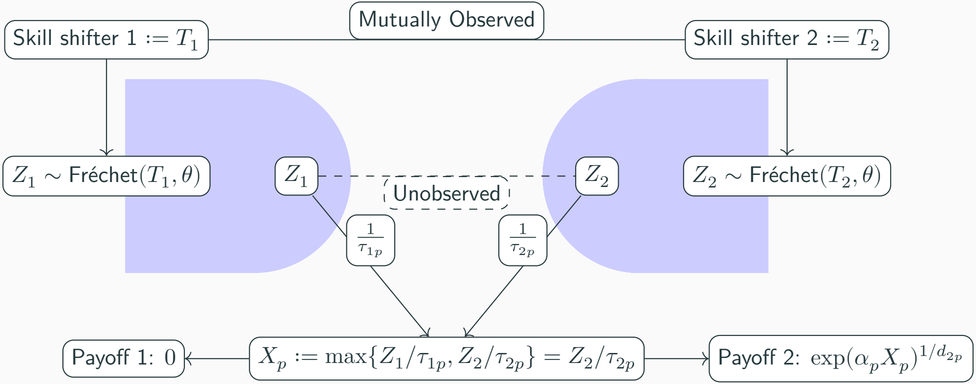
Join if I am likely to have the winning idea $\rightarrow$ positive selection.

$$Z_i > \frac{\tau_{ip} T_{jp}^{1/(\theta+1)}}{(\tau_{ip} d_{ip} - 1)^{1/(\theta+1)}} \xi_i$$

Team formation

Every project member has to say yes $\rightarrow$ assortative matching.

Visual representation



From theory to data

We derive the following empirical predictions from our model:

Prediction 1: Developers are **less likely** to collaborate across greater distances due to higher τ_{ip} and d_{ip}.

Prediction 2: Collaborating developers on average have higher skill.

Prediction 3: Skilled developers worked with skilled developers (PAM).

Prediction 4: Projects with **geographically diverse** teams tend to produce **higher quality** software, as measured by adoption or recognition.

Gravity approach for prediction 1

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:

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

Estimate this with Poisson maximum likelihood.

Collaboration decays with distance - Gravity approach (Prediction 1)

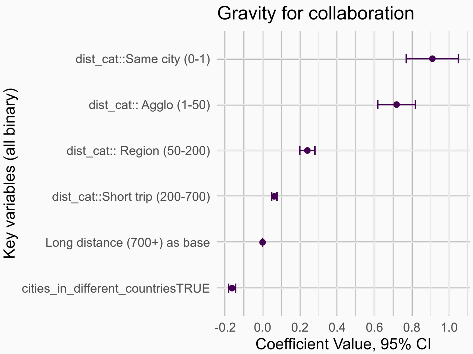


Figure 6: Estimates for different distance categories.

- Developers in the same city are much more likely to work on the same project.

Participation in collaboration (Prediction 2)

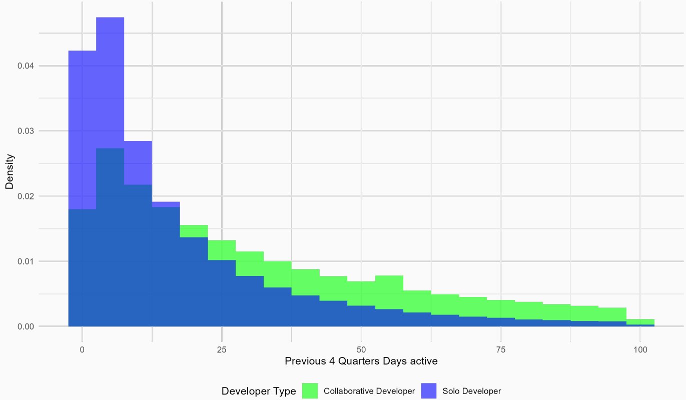


Figure 7: Work experience of developers who only work solo and those who work in collaboration.

- Developers who work in collaborative teams are on average more experienced.
- Experience works as a proxy here for skill.

Experienced developers work with experienced developers (Prediction 3)

Table 2: Assortative matching in developer experience

	Experience of Developer 2
	(1)
Log(Experience of Developer 1)	0.3190***
	(0.0547)
Observations	2,518,765
Squared Correlation	0.00018
Pseudo R^2	0.07283
BIC	102,122,219.0
Quarter x Language fixed effects	✓
Developer Count fixed effects	✓

Team dispersion and quality

Poisson regression

$$\text{Quality}_{pt} = \exp\left[\beta_1 \ln \text{distance}_{i,j\in p} + \beta_2 \text{experience}_{it} + \beta_3 \text{experience}_{jt} + f(n_{pt}) + \lambda_{lt}\right] + \varepsilon_{ljt}$$

where Quality can be:

1. Downstream Libraries
2. Stars on GitHub (3 Quarters Ahead)

And fixed effects cover:

1. Programming language $\times$ Quarter
2. Developer count n_{pt}

Higher success of dispersed teams (Prediction 4)

Table 3: Spatial dispersion and project success

	Shared as Library		Downstream Libraries		Stars on GitHub	
	(1)	(2)	(3)	(4)	(5)	(6)
Log(Distance Between Developers)	0.0321***	0.0185***	0.2638***	0.2528***	0.1792***	0.1649***
	(0.0047)	(0.0045)	(0.0386)	(0.0415)	(0.0110)	(0.0110)
Log(Max Developer Experience by Commits)		0.1326***		0.1833**		0.1494***
		(0.0104)		(0.0794)		(0.0113)
Log(Min Developer Experience by Commits + 1)		-0.0039		-0.0188		-0.0457***
		(0.0062)		(0.0299)		(0.0129)
Observations	513,197	489,211	45,045	44,030	603,918	576,324
Squared Correlation	0.07435	0.08139	0.06271	0.06792	0.01273	0.01348
Pseudo R^2	0.10285	0.10894	0.27119	0.27871	0.15470	0.16002
BIC	284,301.8	273,799.6	3,591,174.6	3,531,085.0	15,447,482.2	15,089,200.9
Quarter x Language fixed effects	✓	✓	✓	✓	✓	✓
Developer Count fixed effects	✓	✓	✓	✓	✓	✓

Standard errors in parentheses. * p<0.1, ** p<0.05, *** p<0.01

Conclusion

- We build a model of global team formation centering around selection on skill.
- This selection induces a positive correlation of distance and quality for software projects.
- Predictions are consistent with data from GitHub 2019.

Next steps

- Get more data.
- Estimate key parameters with "natural experiments" (policy changes on GitHub, war in Ukraine).
- Evaluate counterfactual policies.

Appendix

Expected developer payoff from project p

$$\mathcal{U}_{ip} = \begin{cases} e^{\xi_i Z_i/\tau_{ip}} & \text{if } Z_i/\tau_{ip} > Z_j/\tau_{jp} \\ 0 & \text{otherwise} \end{cases}$$

where ξ_i is a taste parameter for enjoying kudos. In expectation,

$$U_{ip} = \mathsf{E}\,\mathcal{U}_{ip} = e^{-T_{jp}\tau_{ip}^{\theta} Z_i^{-\theta}} e^{\xi Z_i/\tau_{ip}}$$

Increases in Z_i, decreases in T_{jp}, τ_{ip}.

Team formation

Does developer i join project p?

$$U_{ip}(Z_i, T_{jp}, \xi_i) > \mathsf{cost}_i(Z_i, d_{ip}) := e^{d_{ip}\xi_i Z_i}$$

Distribution cost

$d_{ip} \geq 1$. Not all benefits of distant projects can be captured (private cost of participation, time zones, misappropriate of credit).

Gravity

$$d_{ip} = \mathsf{distance}_{ip}^{\gamma_s}$$

where γ_s may be different from γ_k

Join team p if

$$Z_i > \frac{\tau_{ip} T_{jp}^{1/(\theta+1)}}{(\tau_{ip} d_{ip} - 1)^{1/(\theta+1)}} \xi_i^{-1/(\theta+1)}$$

Selection

1. Better skilled developers are more likely to join.
2. Spatial frictions reduce team formation.
3. Projects with high-skilled developers are more selective.

Fréchet magic

Assume Z_i is Fréchet with parameters T_i and θ,

ξ_i is Weibull with κ and $\theta/(\theta+1)$. Then

$$\Pr(Z_i \leq x | i \text{ joins project } p) = e^{-T_{ip}x^{-\theta}}$$

with

$$T_{ip} = T_i + \frac{1}{\kappa}\frac{\tau_{ip}^{\theta} T_{jp}^{\theta/(\theta+1)}}{(\tau_{ip}d_{ip} - 1)^{\theta/(\theta+1)}}$$

Closing the model

Both developers want to join, knowing what to expect from the other.

Mutual coincidence of wants

$$T_{1p} = T_1 + \frac{1}{\kappa} \frac{T_{2p}^{\theta/(\theta+1)}}{(d_{1p} - 1)^{\theta/(\theta+1)}}$$

$$T_{2p} = T_2 + \frac{1}{\kappa} \frac{\tau_{2p}^{\theta} T_{1p}^{\theta/(\theta+1)}}{(\tau_{2p} d_{2p} - 1)^{\theta/(\theta+1)}}$$

Team forms with probability

$$\frac{T_1}{T_{1p}} \frac{T_2}{T_{2p}}$$

References

Atkin, David, M Keith Chen, and Anton Popov. 2022. "The Returns to Face-to-Face Interactions: Knowledge Spillovers in Silicon Valley." National Bureau of Economic Research.

Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica* 70 (5): 1741–79.

Fackler, Thomas, and Nadzeya Laurentsyeva. 2020. "Gravity in Online Collaborations: Evidence from Github." In *CESifo Forum*, 21:15–20. 03. München: ifo Institut-Leibniz-Institut für Wirtschaftsforschung an der ....

Freund, Lukas. 2022. "Superstar Teams: The Micro Origins and Macro Implications of Coworker Complementarities." *Available at SSRN 4312245*.

Head, Keith, Yao Amber Li, and Asier Minondo. 2019. "Geography, Ties, and Knowledge Flows: Evidence from Citations in Mathematics." *Review of Economics and Statistics* 101 (4): 713–27.