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

Gabor Bekes*, Julian Hinz**, Miklos Koren*, Aaron Lohmann**

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

Research questions

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

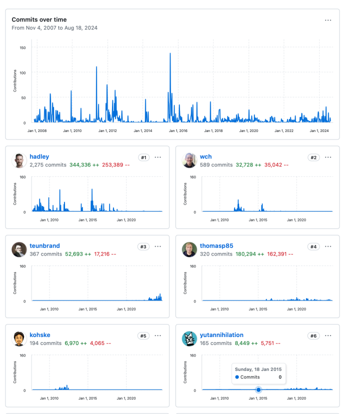


Figure 2: Commits in ggplot2

earning them fame.

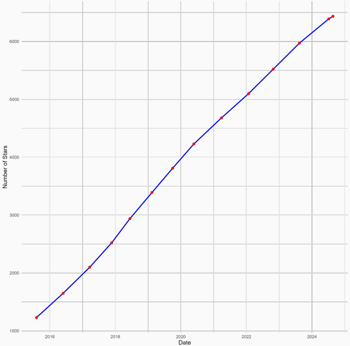


Figure 3: ggplot2 stars over time

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 novel data and combining several data sources.

Data

We use novel, large scale dataset provided by GitHub:

- $37,000,000$ software developers.
- $130,000,000$ projects (repositories).
- Contributions of developers to projects.
- Location of developers on a monthly basis geocoded based on IP addresses.
- Project outcomes:
    - Stars (A like)
    - Forks (Copying code from someone for personal reuse)

Map of developers

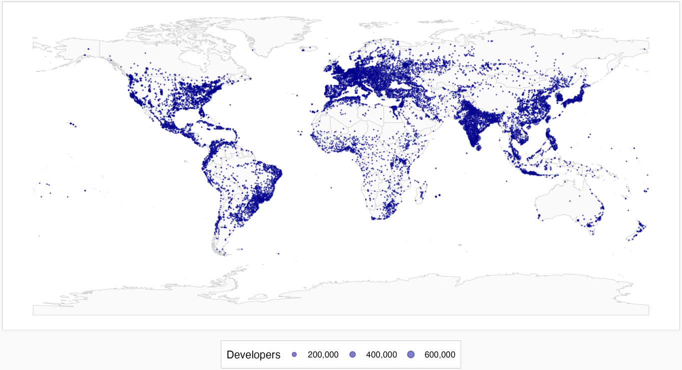


Figure 4: Map of developers around the world

Notes: Based on 30 million developers. Location for each developers based on main developer location.

Collaboration

Num. Developers	Num. Projects	Share in percent
1	115,813,905	90.8
2	7,465,995	5.85
3	2,389,951	1.87
4	1220,896	0.96
5	653,646	0.51

Notes: Only counts core team members. Core team members defined as those contributing in the first 6 months after start of project.

- Team size follows a power-law like relationship.
- The vast majority of projects is developed by one developer.
- Projects with some threshold amount of commits, much. higher percentage is developed by teams.

Pairwise city

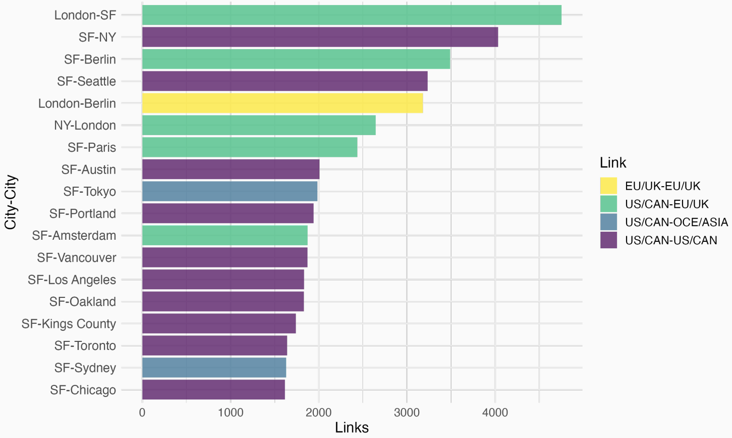


Figure 5: Pairwise collaboration between top cities in JavaScript language.

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: $V_p := e^{X_p}$

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

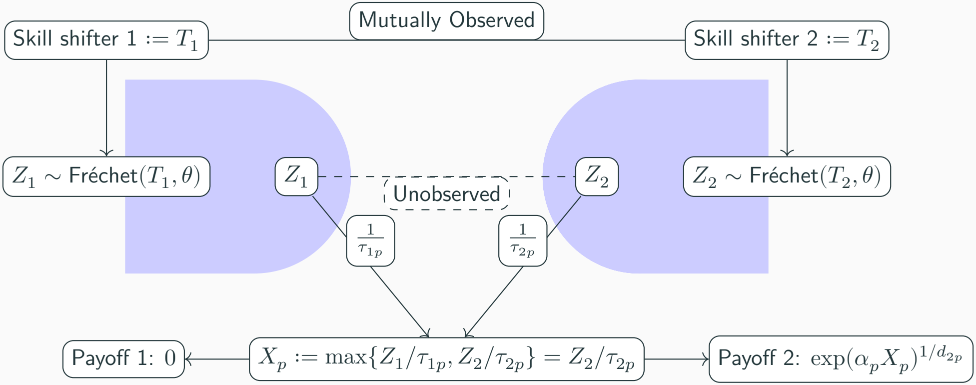
Join if I am likely to have the winning idea $\rightarrow$ positive selection.

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

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

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

Aggregate across city pairs d and o:

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

Estimate this with Poisson maximum likelihood.

Costs for collaboration - Gravity approach (Prediction 1)

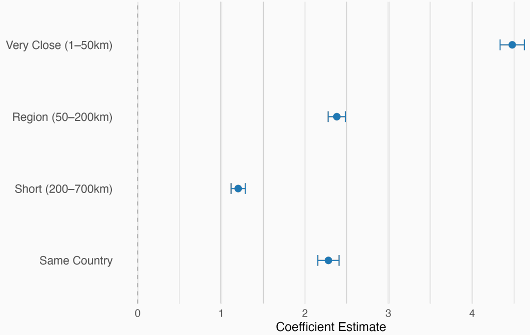


Figure 6: Estimates for different distance categories.

- Developers who are close are much more likely to collaborate.
- Reference category is $700+$

Participation in collaboration (Prediction 2)

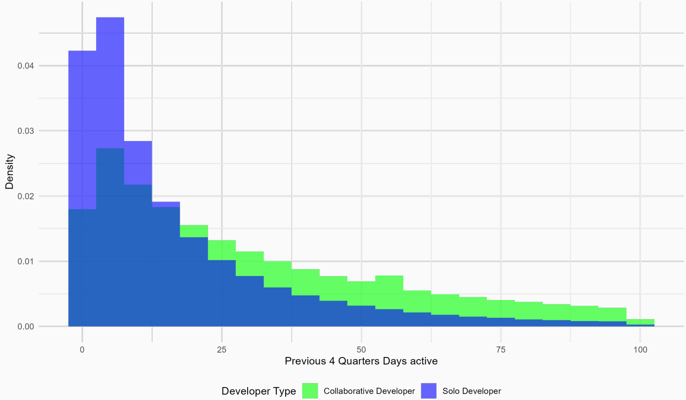


Figure 7: Work experience of developers who only work solo and those who work in collaboration.

- 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)

Dependent Variables:	log(Lag commits developer 1)	Commits/Dev 1 Age
Model:	(1)	(2)
Variables		
log(Lag commits developer 2)	0.2950***	
	(0.0014)	
Commits/Dev 2 Age		0.0849***
		(0.0119)
Fixed-effects		
Start Month ×Language	Yes	Yes
Fit statistics		
Observations	3,227,819	4,488,144
R^2	0.13888	0.00990
Within R^2	0.08834	0.00221

Clustered (Start Month ×Language) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

Team dispersion and quality

We run the following Poisson regression equation

$$Quality_{ljt} = \beta_1 \log \mathsf{dist}_j + \beta_2 \mathsf{coder\ experience}_{jt} + \lambda_t \times \delta_l + \varepsilon_{ljt}$$

where Quality can be:

1. Number of Stars
2. Number of public Forks

And the Fixed effects cover:

1. Language
2. Quarter

Higher success of dispersed teams (Prediction 4) – Teams of two

Dependent Variables:	Number Stars (after 12 months)		Number Forks (after 12 months)	
Model:	(1)	(2)	(3)	(4)
Variables				
log(distance)	0.3208***	0.2419***	0.2224***	0.1774***
	(0.0052)	(0.0048)	(0.0044)	(0.0040)
log(Age dev 1)		0.4667***		0.2484***
		(0.0149)		(0.0119)
log(Age dev 2)		0.4393***		0.2433***
		(0.0144)		(0.0105)
Fixed-effects				
Start Month $\times$Language	Yes	Yes	Yes	Yes
Fit statistics				
Observations	3,594,292	3,554,979	3,594,676	3,555,345
Squared Correlation	0.01035	0.01328	0.02615	0.02740
Pseudo R^2	0.20358	0.25902	0.11964	0.14604
BIC	72,598,211.8	66,970,411.6	21,982,816.1	21,024,483.2

Clustered (Start Month $\times$Language) standard-errors in parentheses
*Signif. Codes: ***: 0.01, **: 0.05, *: 0.1*

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 2018-2024.

Next steps

- 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?

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

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

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

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.