Dispersed teams are more successful: Evidence from OSS

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

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

Research questions

1. How and where is **open source software** developed?
    - Descriptives on OSS developers on GitHub.
2. Do **spatially dispersed** developers produce **quality** software?
    - Model of global team formation with empirical predictions.

Economics of open source

- Incentives to produce and contribute to open source (literature in the early 2000s, not our main concern)
- Spatial dispersion and its interactions with quality (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

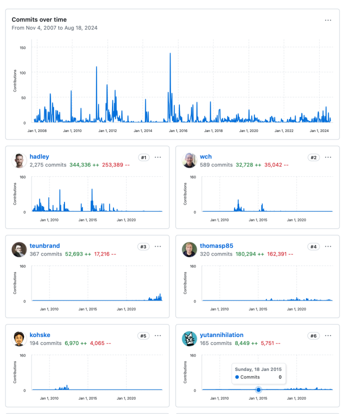

Figure 2: Commits in ggplot2

earning them fame.

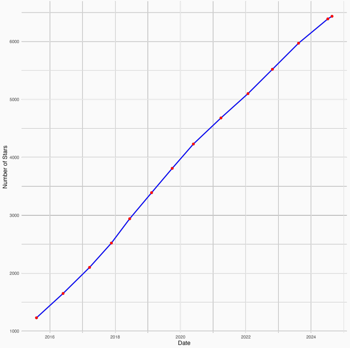

Figure 3: ggplot2 stars over time

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

We use data from two main data sources:

- GHtorrent: An effort to collect as much data as possible from GitHub. Our main sample will consist of
  - $835,283$ projects.
  - $347,767$ developers.
  - over years from 2012 to 2019
- Libraries.io: A data effort to collect downstream dependencies of OSS projects.

Map of developers

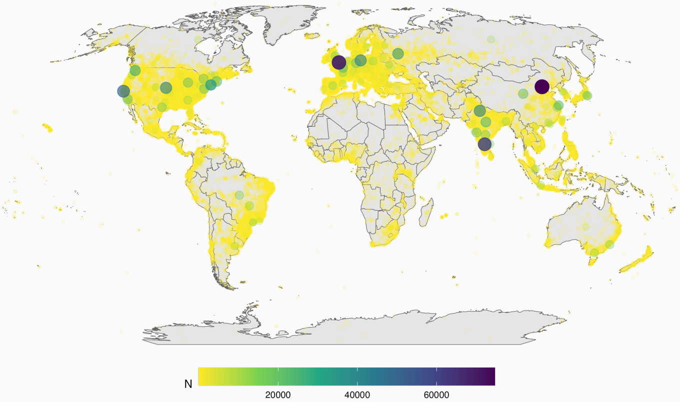

Figure 4: OSS developers around the world

Collaboration

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

Table 1: Share of projects by amount of developers.

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

Pairwise city

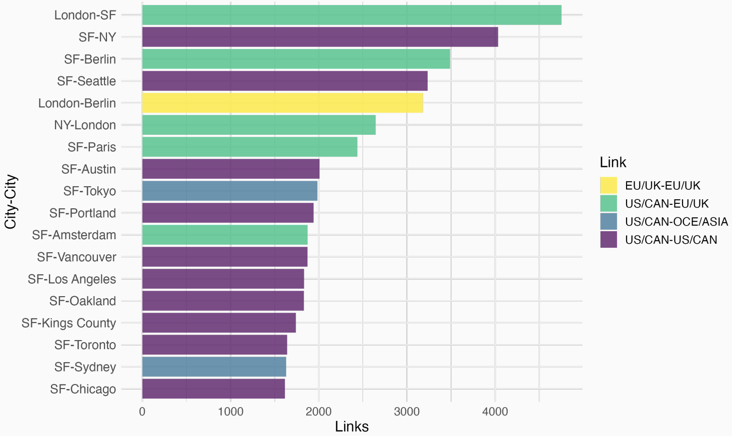

Figure 5: Pairwise collaboration between top cities.

Model

- From the modelling perspective OSS has two interesting features:
  - Potential spatial dispersion.
  - Self selection into collaboration by developers.

Motivates to build a model of global team formation.

Model - Economic environment

- Developers have heterogenous skills Z_i which is drawn from a Fretchet distribution according to $\Pr(Z_i \leq x) = e^{-T_i x^{-\theta}}$
- Developers can work in teams which is subject to costs:
    - Communication: $\tau_{ip} = \text{distance}_{ip}^{\gamma_k}$
    - Participation: $d_{ip} = \text{distance}_{ip}^{\gamma_s}$
- Production is based on the best idea of the developers: $X_p = \max_{j \in p}\{Z_j / \tau_{jp}\}$

Model - Visual representation

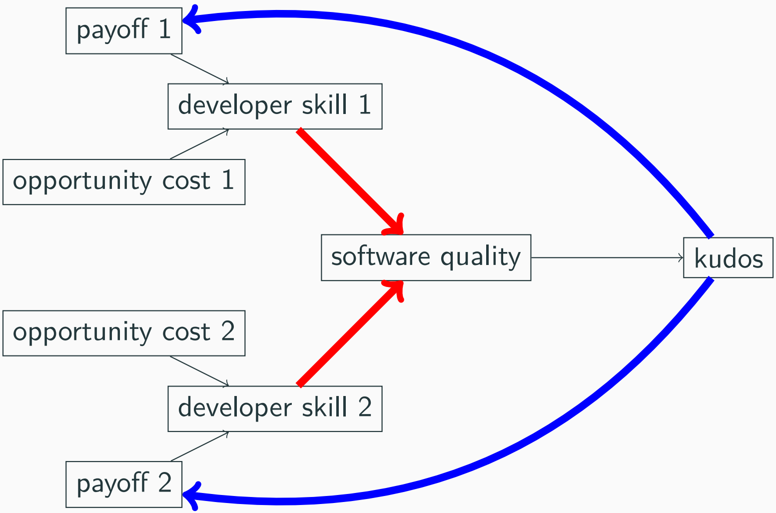

Figure 6: Visualisation of the model.

Sharing kudos

Overall customer happiness increases in software quality:

$$V_p := e^{X_p}$$

Attribution of kudos

The better-skilled developer gets all the kudos for V_p. ($\approx$ "First author bias")

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: 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.

Costs for collaboration - Gravity approach (Prediction 1)

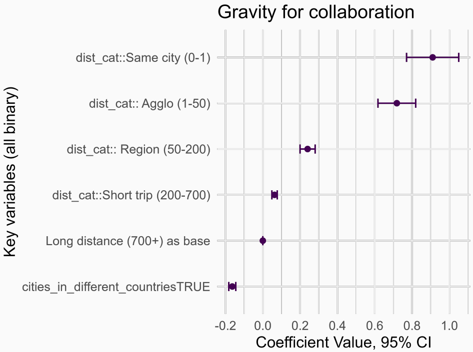

Figure 7: Estimates for different distance categories.

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

Participation in collaboration (Prediction 2)

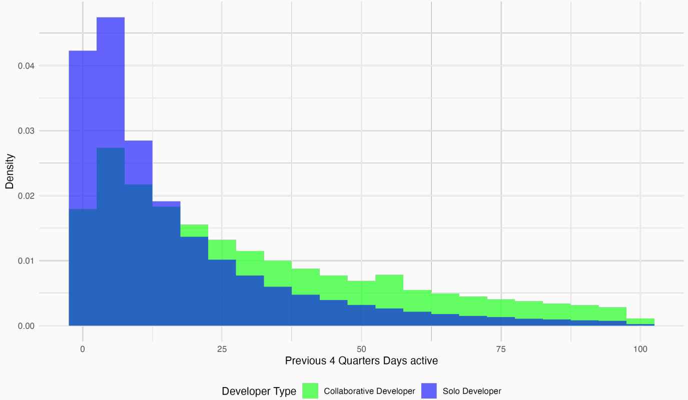

Figure 8: Work experience of developers who only work solo and those who work in collaboration.

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

Team dispersion and quality

We run the following Poisson regression equation

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

where Quality can be:

1. Downstream Libraries
2. Stars

And the Fixed effects cover:

1. Language
2. Quarter
3. Developer Count

Higher success of dispersed teams (Prediction 3)

Dependent Variables:	Downstream Libraries		Stars Count (3 Quarters Ahead)	
Model:	(1)	(2)	(3)	(4)
Variables				
Log(Distance Between Coders)	0.2638***	0.2528***	0.1792***	0.1649***
	(0.0386)	(0.0415)	(0.0110)	(0.0110)
Log(Maximum Coder Quality (Commits))		0.1833**		0.1494***
		(0.0794)		(0.0113)
Log(Minimum Coder Quality (Commits + 1))		-0.0188		-0.0457***
		(0.0299)		(0.0129)
Fixed-effects				
Quarter and Language Fixed Effects	Yes	Yes	Yes	Yes
Developer Count	Yes	Yes	Yes	Yes
Fit statistics				
Observations	45,045	44,030	603,918	576,324
Pseudo R^2	0.27119	0.27871	0.15470	0.16002

Clustered (Quarter and Language Fixed Effects) 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 of quality for software projects.
- We provide descriptive statistics on OSS development and showcase the informative insights of the data.

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.