The Global Software Production Network

Carlo Birkholz[1]
David Gomtsyan[2]

August 28, 2024

[1] ZEW, University of Mannheim
[2] CERDI, Universite Clermont Auvergne

Introduction

Motivation

- Advanced economies greatly increase share of high-skilled services sector over the course of development, e.g. $>$50% of VA in the US [Buera and Kaboski, 2012]
- Many industries within this sector produce tradable outputs, e.g. IT, accounting, management services
- Tradable services are central to the debate of premature de-industrialization versus service-led growth [Rodrik, 2016]

RQ: Can developing countries benefit from exporting opportunities in the growing sector of tradable services, given the near free information flow via the internet and wage differentials relative to developed countries?

This paper

- Data on 2.55 million IT projects and 2.64 million users in 5,400 locations from GitHub
- Economic geography model of trade in tasks [Eaton and Kortum, 2002]
- Estimate productivity levels (skills of developers) at the level of locations
- Estimate distance elasticity of trade in tasks
- Study migration patterns of software developers over time

Preview of the results

Three factors limiting trade:

1. Significant productivity differences within and between countries correlated with income per capita levels
2. A notable decline in trade volumes with distance
3. Sorting patterns among software developers that are suggestive of brain drain

Literature

1. Human capital and income differences

   Clemens [2013]; Hendricks and Schoellman [2018]; Martellini et al. [2024]

2. Trade costs in services

   Gervais and Jensen [2019]; Eckert et al. [2019]; Anderson et al. [2014]; Kleinman [2022]

3. High skilled migration

   Akcigit et al. [2016]; Dauth et al. [2022]

Data

GitHub

- GitHub is the largest service for software development and version control
- We use two snapshots of all public activities on the plattform, 2021 and 2019
- **Users** ▸▸ Map
  - 45.8 million, 2.6 million with (cleaned) location information
  - Users with location information account for 36.5% of the trade volume (67.4% when quality adjusted)
  - Individuals can follow each other to receive updates about each others' activities
  - Professional users collaborate in projects and have strong monetary and reputational incentives

Data

- **Projects**
  - Projects have an owner and team members who can make contributions (commits)
  - Non-team members can also make contributions via pull requests
  - 189 million public projects, 47.3 million where the owners' location is identified, 2.55 million where the owners' and at least one contributors' location is identified
  - Successful open source projects generate revenue
- **Commits**
  - Sample of 219 million commits, which are versions of gradual changes to a project
  - Each commit has an author and a committer, most commonly being the same user
  - We define software production flows to originate from the author
  - Exclude commits by bots

Data

- **Forks and pull requests**
  - Forks are copies of existing projects which serve two purposes:
    - Propose changes to a project as a non-member via pull requests
    - Use of the project as final software product or as input for new independent software
- **Auxiliary data**
  - Geographic areas: Functional Urban Areas from GHS and admin-2 regions from GADM
  - Population from GHS 1km grid
  - Nightlights from VIIRS V2.1 Map
  - Income: US Metro areas from ACS, country level from Stack Overflow Survey and WDI Metro areas

Defining links (flows)

$$X_{ij} = \sum_{k \in K} commits_{jk} \times 1[owner_{ik} = 1],$$

- X_{ij} - the volume of the code that flows from location j to location i
- K - the set of projects
- $commits_{jk}$ - the number of commits on project k by users from location j
- $1[.]$ - an indicator function if the owner of project k is in location i

Owner centrality

Methodology

Tasks trade model

- Eaton and Kortum (2002) framework
- An individual produces differentiated computer code q in location i with efficiency $z_i(q)$
- Constant marginal disutility of labor supply
- Individual productivities are drawn from $F_i(z) = e^{-T_i z^{-\theta}}$
  - Location specific T_i
- Iceberg trade costs d_{ij}

Estimating equation

- $$ln\left(\frac{X_{ij}}{X_{ii}}\right) = \underbrace{ln\left(T_j\right)}_{\text{Exporter FE}} \underbrace{-ln\left(T_i\right) - \theta im_i}_{\text{Importer FE}} \underbrace{-\theta d_k - \theta a_{ij} - \theta b_{ij} - \theta Lang_{ij}}_{\text{Bilateral observables}} -\theta \nu_{ij}$$

- $exp(EFE_j) = T_j,$

- Alternative: recover productivities from importer FE ▸▸ Importer FE

Reduced form approaches

Approach 1: Page rank algorithm on trade links

- Locations are nodes of a directed graph
- X_{ij} represents the strength of a link between a pair
- Determine the centrality of each node

Approach 2: Page rank algorithm on follower network

- Calculate the centrality of each individual based on the following network
- Measure of individual quality
- Aggregate individuals' quality scores at the location level

Results

Distance elasticity in trade of tasks

	(1) X_{ij}/X_{ii}	(2) X_{ij}/X_{ii}	(3) X_{ij}/X_{ii}	(4) X_{ij}/X_{ii}	(5) $\hat{X}_{ij}/\hat{X}_{ii}$
Log distance in miles	-0.8081***	-0.8093***	-0.9129***	-0.6833***	-0.7311***
	(0.0811)	(0.0688)	(0.0834)	(0.1053)	(0.0071)
Controls	Yes	Yes	Yes	Yes	Yes
Same location dummy	No	No	No	Yes	No
Sample	FUA + Admin	FUA only	US FUA only	FUA + Admin	FUA + Admin
Observations	16,678,894	5,266,000	60,945	16,678,894	13,190,040
Pseudo R-squared	0.7067	0.7053	0.8419	0.7087	0.4920

Gains from Trade

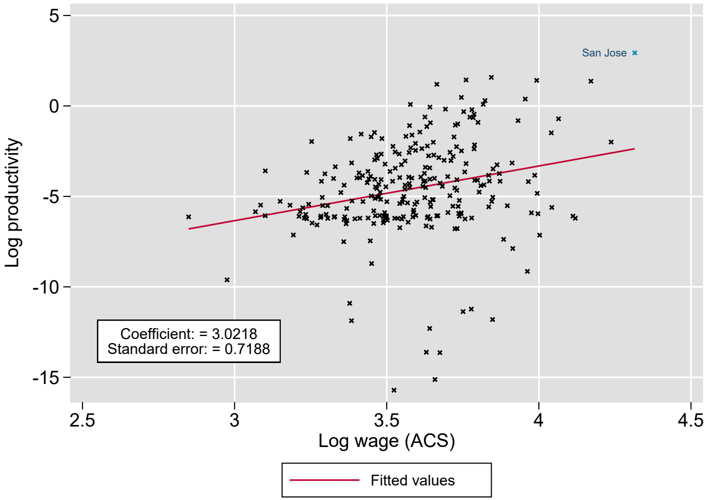
Ranking of the Top 35 Cities Across the World - Part I

Rank	Model	Approach 1	Approach 2
1	San Jose	San Jose	San Jose
2	Prague	New York	New York
3	Bengaluru	Seattle	London
4	Las Palmas de Gran Canaria	Boston	Beijing
5	Los Angeles	London	Seattle
6	Nuremberg	Washington D.C.	Shanghai
7	Portland (Oregon)	Los Angeles	Portland (Oregon)
8	Ottawa	Paris	Boston
9	New York	Beijing	Los Angeles
10	Seattle	Tokyo	Tokyo
11	Detroit	Atlanta	Berlin
12	Taichung	Chicago	Paris
13	Krasnoyarsk	Portland (Oregon)	Guangzhou
14	Toronto	Berlin	Toronto
15	Berlin	Denver	Austin
16	Ho Chi Minh City	Austin	Hangzhou
17	Sydney	Shanghai	Chicago

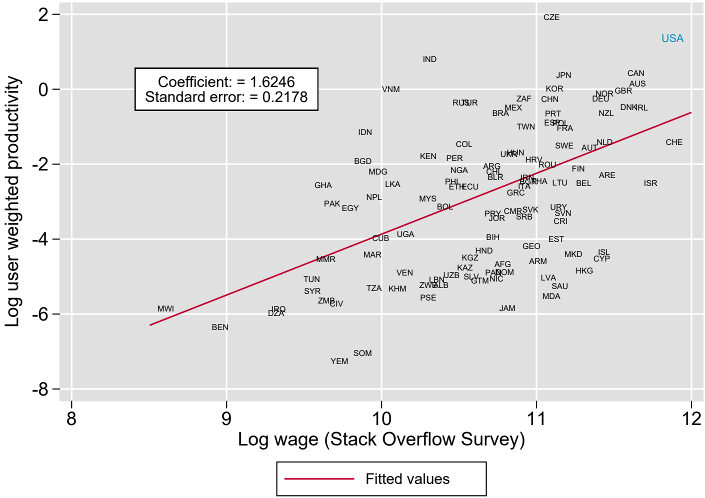
Ranking of the Top 35 Cities Across the World - Part II

Rank	Model	Approach 1	Approach 2
18	Tokyo	Toronto	Denver
19	Cape Town	Amsterdam	Washington D.C.
20	Cambridge	Bengaluru	Melbourne
21	Arrecife	Seoul	Pittsburgh
22	London	Philadelphia	Stockholm
23	Dallas	Tijuana	Moscow
24	São Paulo Nanjing	Guangzhou	Sydney
25	Krakow	Vancouver	Vancouver
26	Boston	Zurich	Bengaluru
27	Oslo	São Paulo	Montreal
28	Vancouver	Stockholm	Amsterdam
29	Moscow	Montreal	São Paulo
30	Beijing	Sydney	Atlanta
31	Dutchess County US (Poughkeepsie)	Cambridge	Philadelphia
32	Austin	Moscow	Madrid
33	Melbourne	Delhi [New Delhi]	Barcelona
34	Nanjing	Melbourne	Munich
35	Tijuana	Hangzhou	Seoul

Validation: US FUAs productivity and IT-related professions' wages



IT jobs


Coefficient: = 1.6246
Standard error: = 0.2178
Log user weighted productivity
Log wage (Stack Overflow Survey)
Fitted values
USA
CZE
IND
JPN
CAN
AUS
VNM
KOR
CHN
CHE
ISR
MWI
BEN
SOM
YEM
JAM

Validation: Top 35 Universities in the US, the UK and Germany

Rank	University	Rank	University
1	MIT	19	Northeastern University
2	University of California, Berkeley	20	University of Saarland
3	Carnegie Mellon University	21	Columbia University
4	University of California, Los Angeles	22	University of California, San Diego
5	Stanford University	23	University of Duesseldorf
6	University of Oxford	24	University of Applied Sciences Munich
7	Vanderbilt University	25	Arizona State University
8	Technical University Berlin	26	Harvard University
9	University of Wisconsin-Madison	27	Brown University
10	Johns Hopkins University	28	Purdue University
11	University of Edinburgh	29	California Institute of Technology (Caltech)
12	University of Washington	30	University of California, Davis
13	Cornell University	31	Technical University Munich
14	Brigham Young University	32	University of Cambridge
15	University of Colorado Boulder	33	University of Hawaii
16	University of Arizona	34	University of Essen
17	New York University	35	University of Michigan
18	Washington University in St. Louis		

Correlations of IT productivity and income per capita

	(1) Log productivity	(2) Log productivity	(3) Log productivity	(4) Log productivity
Log nightlights per capita	0.5248*** (0.0634)			
Log GDP per capita		0.8448*** (0.1162)	0.8367*** (0.1228)	0.9028*** (0.1259)
Sample	FUA	Country level	Country level	Country level
Aggregation method		Average of top 5%	Population weighted	User weighted
Observations	2,639	121	121	121
R-squared	0.0239	0.3252	0.3145	0.3251
F	68.45	52.86	46.45	51.40

Productivity gaps between rich and poor countries

Variables	Productivity gap
GDP per capita	4.61
Industry VA per worker	3.71
Services VA per worker	3.73
IT productivity, top 5%	4.15
IT productivity, population weighted	4.27
IT productivity, user weighted	4.64

Defining flows of trade in ideas

$$\tilde{X}_{ij} = \sum_{k \in K} fork_{ik} \times 1[owner_{jk} = 1],$$

- $\tilde{X}_{ij}$ - the flow of final software from location j to location i
- K - the set of projects
- $fork_{ji}$ - the number of forks on project k by users from location i
- $1[.]$ - an indicator function if the owner of project k is in location j

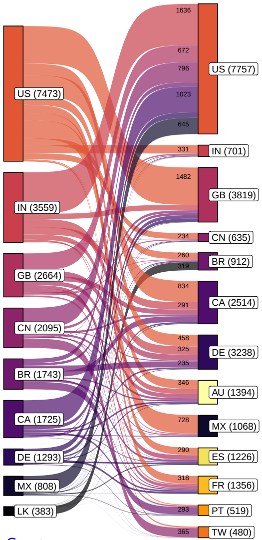
Trade in ideas

	(1)	(2)	(3)	(4)
	$\tilde{X}_{ij}/\tilde{X}_{ii}$	Comparative advantage in ideas over services		
Log distance in miles	-0.4376***			
	(0.0072)			
Log GDP per capita		0.8082***	0.3396***	0.1280
		(0.1751)	(0.1241)	(0.1048)
Controls	Yes	No	No	No
Sample	FUA + Admin	Country level	Country level	Country level
Aggregation method		Average of top 5%	Population weighted	User weighted
Observations	11,922,149	119	119	119
R-squared	0.6629	0.1363	0.0611	0.0139
F		21.30	7.492	1.493

Migration & Sorting

Migration descriptives

- 1.56 million users with cleaned location in 2021 and 2019
- 98,000 migrants
  - 38,000 cross-country
  - 60,000 within country



Individual quality and likelihood to migrate

	Migrated	Migrated	Migrated	Migrated within country	Migrated across country
Panel A:					
Log individual score	0.1902***	0.1639***	0.1898***	0.1902***	0.1838***
	(0.0091)	(0.0081)	(0.0052)	(0.0052)	(0.0123)
Observations	939,034	938,552	933,943	921,550	909,621
Pseudo R2	0.0175	0.0630	0.108	0.106	0.222
Panel B:					
2nd quartile	0.6303***	0.5971***	0.6201***	0.6804***	0.5001***
	(0.0224)	(0.0252)	(0.0188)	(0.0160)	(0.0404)
3rd quartile	0.9101***	0.8504***	0.8814***	0.9439***	0.7497***
	(0.0160)	(0.0215)	(0.0218)	(0.0184)	(0.0446)
4th quartile	1.2919***	1.1739***	1.1991***	1.2919***	1.0106***
	(0.0166)	(0.0278)	(0.0279)	(0.0219)	(0.0635)
Observations	1,566,353	1,565,559	1,558,279	1,539,900	1,519,561
Pseudo R2	0.0439	0.0902	0.133	0.123	0.244
Origin country FE	X	X			
Destination country FE		X	X		X
Origin city FE			X	X	X
Number migrants	97,438	97,438	97,438	60,122	37,316

Directional migration of individuals based on individual quality

	Up migration	Down migration	Up migration	Down migration
Panel A:				
Log individual score	0.2124***	0.1515***	0.0307***	-0.0343***
	(0.0064)	(0.0081)	(0.0034)	(0.0070)
Observations	872,287	878,591	69,184	66,393
Pseudo R2	0.186	0.128	0.0907	0.127
Panel B:				
2nd quartile	0.6368***	0.5832***	0.0104	-0.0276**
	(0.0214)	(0.0284)	(0.0104)	(0.0119)
3rd quartile	0.9155***	0.8246***	0.0558***	-0.0787***
	(0.0217)	(0.0356)	(0.0091)	(0.0107)
4th quartile	1.2668***	1.0687***	0.0954***	-0.1364***
	(0.0288)	(0.0452)	(0.0101)	(0.0148)
Observations	1,465,610	1,467,499	85,657	82,480
Pseudo R2	0.202	0.147	0.0927	0.131
Destination country FE	X	X	X	X
Origin city FE	X	X	X	X
Sample	All	All	Migrants	Migrants
Number migrants	52,256	37,763	52,256	37,763

Migration to higher and lower income countries based on individual quality

	Migration to > GDP per capita	Migration to < GDP per capita	Migration to > GDP per capita	Migration to < GDP per capita
Panel A:				
Individual quality	0.3021***	0.1936***	0.0196***	-0.0248***
	(0.0111)	(0.0116)	(0.0040)	(0.0070)
Observations	839,292	807,682	27,416	25,410
Pseudo R2	0.125	0.125	0.141	0.226
Panel B:				
2nd quartile	0.5330***	0.6941***	-0.0086	0.0049
	(0.0306)	(0.0379)	(0.0108)	(0.0153)
3nd quartile	0.8936***	0.9535***	0.0078	-0.0150
	(0.0272)	(0.0368)	(0.0090)	(0.0139)
4nd quartile	1.3681***	1.2778***	0.0344***	-0.0584***
	(0.0268)	(0.0490)	(0.0089)	(0.0150)
Observations	1,393,561	1,345,274	33,800	31,156
Pseudo R2	0.140	0.138	0.142	0.230
Origin city FE	X	X	X	X
Sample	All	All	Cross-country migrants	Cross-country migrants
Number migrants	22,913	14,403	22,913	14,403

Migrants comparative quality in the destinations

	Above median score in destination	Above median score in destination	Above median score in destination	Above median score in destination	Above median score in destination
Panel A:					
Migrated	0.3937***				
	(0.0091)				
Up migration		0.3469***			
(productivity)		(0.0079)			
Down migration			0.4332***		
(productivity)			(0.0134)		
Up migration				0.3284***	
(GDP per capita)				(0.0151)	
Down migration					0.3851***
(GDP per capita)					(0.0155)
Observations	1,560,104	1,553,869	1,553,869	1,560,104	1,560,104
Pseudo R2	0.0050	0.0033	0.0034	0.0025	0.0025
	Δ quartile individual score	Δ quartile individual score	Δ quartile individual score	Δ quartile individual score	Δ quartile individual score
Panel B:					
Migrated	-0.0496***				
	(0.0125)				
Up migration		-0.1224***			
(productivity)		(0.0121)			
Down migration			0.0561***		
(productivity)			(0.0192)		
Up migration				-0.1449***	
(GDP per capita)				(0.0201)	
Down migration					0.0039
(GDP per capita)					(0.0168)
Observations	1,566,039	1,553,926	1,553,926	1,566,039	1,566,039
R-squared	0.4388	0.1012	0.0714	0.4438	0.4346
Destination city FE	X	X	X	X	X
Number migrants	97,438	52,256	37,763	22,913	14,403

Migration flows at the country level

	Net migration	Out-migration	In-migration
Panel A:			
Log GDP per capita	0.0213*	0.0128**	0.0323**
	(0.0115)	(0.0055)	(0.0129)
Observations	146	146	146
R-squared	0.0177	0.0269	0.0442
Panel B:			
Log GDP per capita	0.0327***	-0.0042	0.0250***
	(0.0075)	(0.0060)	(0.0082)
Observations	108	108	108
R-squared	0.1053	0.0037	0.1028

Conclusions

- There are substantial gaps in skill levels between rich and poor cities
- Surprisingly large trade costs which suggest offline meetings and in-person networks matter quite a lot
- Evidence of brain drain
- Policy implications
  - Invest in human capital
  - Retain talent
- Outlook:
  - Agglomeration effects
  - Positive spill-overs from brain drain to origin location

Thanks for your attention!

carlo.birkholz@zew.de *https://carlo-birkholz.github.io/*

References

Ufuk Akcigit, Salomé Baslandze, and Stefanie Stantcheva. Taxation and the international mobility of inventors. *American Economic Review*, 106(10):2930–2981, 2016.

James E Anderson, Catherine A Milot, and Yoto V Yotov. How much does geography deflect services trade? canadian answers. *International Economic Review*, 55(3): 791–818, 2014.

Costas Arkolakis, Arnaud Costinot, and Andrés Rodríguez-Clare. New trade models, same old gains? *American Economic Review*, 102(1):94–130, 2012.

Francisco J. Buera and Joseph P. Kaboski. The rise of the service economy. *American Economic Review*, 102(6):2540–69, May 2012. doi: 10.1257/aer.102.6.2540. URL `https://www.aeaweb.org/articles?id=10.1257/aer.102.6.2540`.

Michael A Clemens. Why do programmers earn more in houston than hyderabad? evidence from randomized processing of us visas. *American Economic Review*, 103 (3):198–202, 2013.

Arnaud Costinot and Andrés Rodríguez-Clare. Trade theory with numbers: Quantifying the consequences of globalization. In *Handbook of international economics*, volume 4, pages 197–261. Elsevier, 2014.

Wolfgang Dauth, Sebastian Findeisen, Enrico Moretti, and Jens Suedekum. Matching in cities. *Journal of the European Economic Association*, 20(4):1478–1521, 2022.

Jonathan Eaton and Samuel Kortum. Technology, geography, and trade. *Econometrica*, 70(5):1741–1779, 2002. ISSN 00129682, 14680262. URL `http://www.jstor.org/stable/3082019`.

Fabian Eckert et al. Growing apart: Tradable services and the fragmentation of the us economy. *mimeograph, Yale University*, 2019.

Antoine Gervais and J Bradford Jensen. The tradability of services: Geographic concentration and trade costs. *Journal of International Economics*, 118:331–350, 2019.

Lutz Hendricks and Todd Schoellman. Human capital and development accounting: New evidence from wage gains at migration. *The Quarterly Journal of Economics*, 133(2):665–700, 2018.

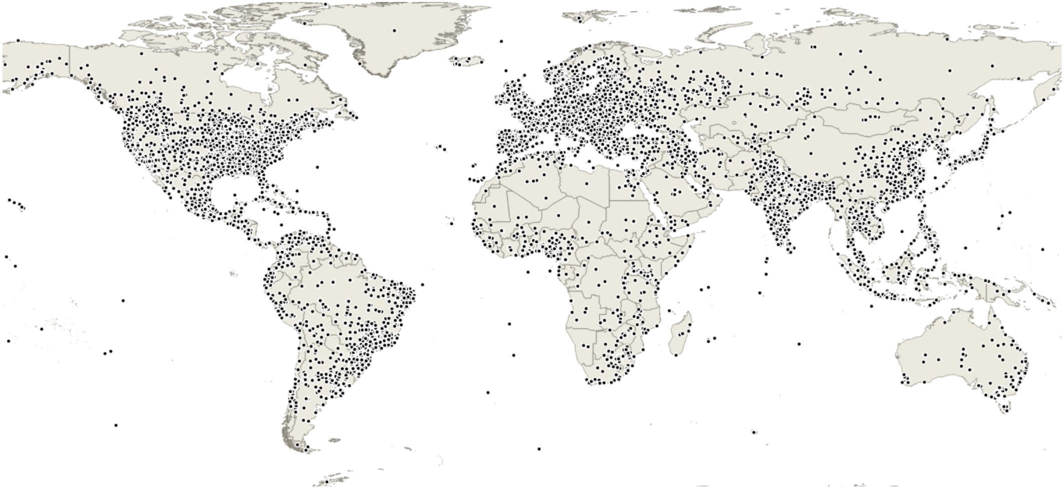
Benny Kleinman. Wage inequality and the spatial expansion of firms. Technical report, Princeton Working Paper, 2022.

Paolo Martellini, Todd Schoellman, and Jason Sockin. The global distribution of college graduate quality. *Journal of Political Economy*, 132(2):434–483, 2024.

Dani Rodrik. Premature deindustrialization. *Journal of economic growth*, 21:1–33, 2016.

Michael E Waugh. International trade and income differences. *American Economic Review*, 100(5):2093–2124, 2010.

Map of unique user locations across the world

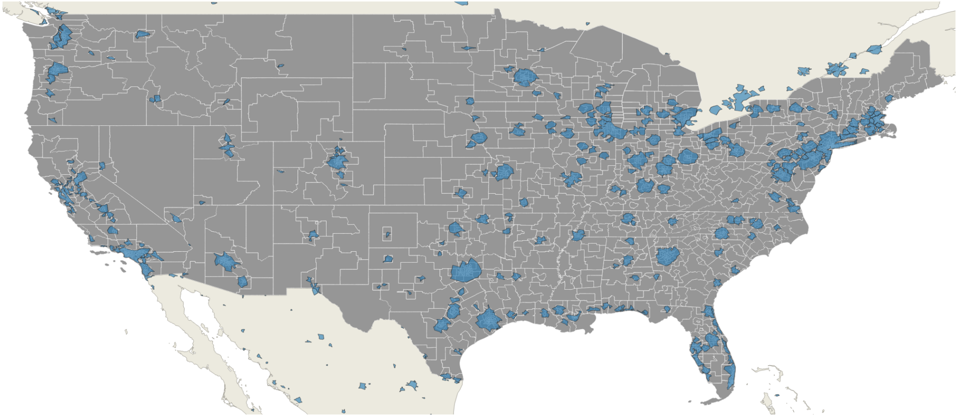


⏩ Data

Illustration data merge FUA, nightlights, users

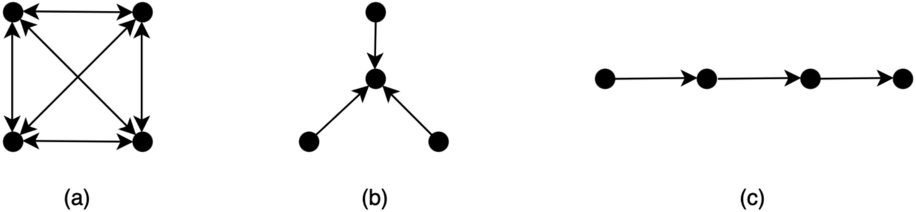
▸▸ Data

US Metro areas to FUAs



⏩ Data

The organization of teams

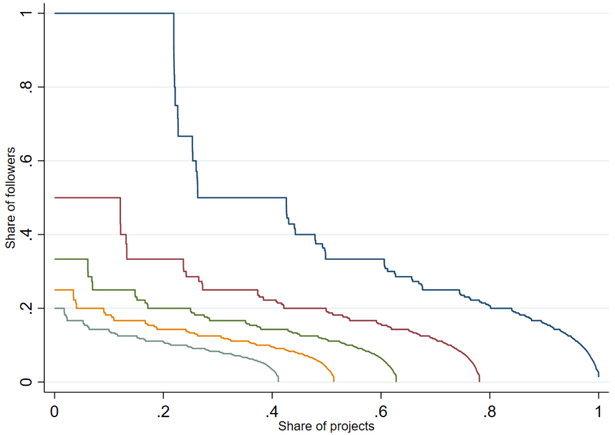


Notes: (a) fully connected network; (b) star network; (c) chain network.

The structure of collaboration

	(1)	(2)	(3)	(4)	(5)	(6)
	i follows j	i follows j	i follows j	i follows j	i follows j	Share of follows
$Owner_j$	2.0161***	2.1468***	1.4894***	1.3300***	1.2989***	0.9352***
	(0.0014)	(0.0015)	(0.0028)	(0.0036)	(0.0141)	(0.0018)
$Owner_i$		1.9697***	1.2169***	1.0627***	-7.2051***	
		(0.0016)	(0.0032)	(0.0041)	(0.9721)	
Same country			0.9506***	0.6787***	0.4621***	
			(0.0018)	(0.0027)	(0.0040)	
Same location				0.4514***	0.2389***	
				(0.0026)	(0.0047)	
Team size	> 2	> 2	> 2	> 2	> 100	> 2
Mean	0.015	0.015	0.030	0.031	0.015	0.161
Observations	244,177,260	244,177,260	47,869,198	30,712,310	24,947,588	3,419,080
Pseudo R^2	0.0303	0.0548	0.0517	0.0502	0.0106	0.0323

The hierarchy of following structures



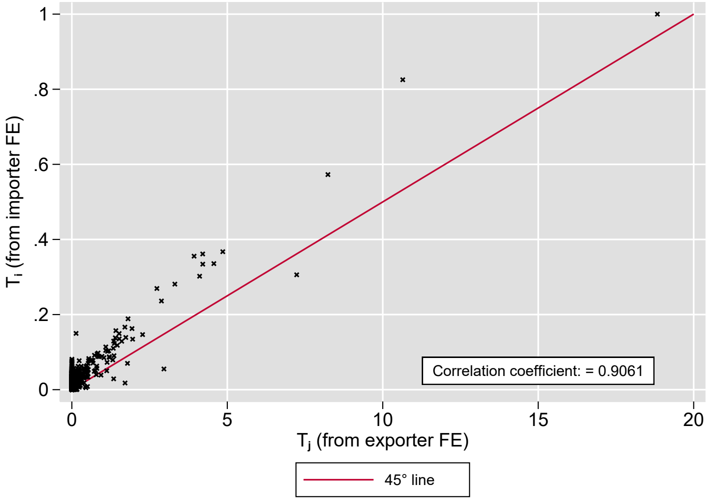
Notes: The cumulative distribution of the share of followers within projects.

Trade flows

Correlation between T_i and T_j

- $T_i = \left(\frac{FE_i}{FE_{SJ}}\right)\left(\frac{w_i}{w_{SJ}}\right)^{\theta}$
- $\theta = 0.18$ [Waugh, 2010]
- $w_i = \beta_{ACS} * pop_i + w_c$

Model



US Metro areas to FUAs

Code	Description
1005	Computer and information research scientists
1006	Computer systems analysts
1007	Information security analysts
1010	Computer programmers
1021	Software developers
1022	Software quality assurance analysts and testers
1031	Web developers
1032	Web and digital interface designers
1050	Computer support specialists
1065	Database administrators and architects
1105	Network and computer systems administrators
1106	Computer network architects
1108	Computer occupations, all other
1240	Other mathematical science occupations

Validation

Gains from trade

- Data from World Input-Output Database (WIOD), picking sector "Computer programming, consultancy and related activities; information service activities."
- 41 countries
- Perfect competition model [Arkolakis et al., 2012]
- Gains from trade compared with the hypothetical scenario in which the software development sector was autarkic: $G_i = 1 - \left(\frac{X_{ii}}{\sum_j X_{ij}} \right)^{1/\epsilon}$
- $\epsilon = 5$ [Costinot and Rodríguez-Clare, 2014]

Gains from trade

Country	WIOD consumption	WIOD investment	GitHub country level	GitHub US city level
USA	0.94	0.37	5.40	8.98
Mean	16.55	7.62	9.83	

- Correlation between WIOD and GitHub based calculations is 0.5 and 0.55 respectively
- US domestic gains are larger than US international gains

⏩ Distance elasticity