Goldbeck, Moritz

Working Paper

Bit by Bit: Colocation and the death of distance in software developer networks

ifo Working Paper, No. 386

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Goldbeck, Moritz (2022) : Bit by Bit: Colocation and the death of distance in software developer networks, ifo Working Paper, No. 386, ifo Institute - Leibniz Institute for Economic Research at the University of Munich, Munich

This Version is available at:
https://hdl.handle.net/10419/266603

ifo WORKING PAPERS

386
2022

November 2022

Bit by Bit
Colocation and the Death of Distance in Software Developer Networks

Moritz Goldbeck

ifo INSTITUTE
Leibniz Institute for Economic Research
at the University of Munich

Imprint:

ifo Working Papers
Publisher and distributor: ifo Institute – Leibniz Institute for Economic Research at the University of Munich
Poschingerstr. 5, 81679 Munich, Germany
Telephone +49(0)89 9224 0, Telefax +49(0)89 985369, email ifo@ifo.de
www.ifo.de

An electronic version of the paper may be downloaded from the ifo website:
www.ifo.de

ifo Working Paper No. 386

Bit by Bit
Colocation and the Death of Distance in Software Developer Networks*

Abstract

Digital tools potentially enable remote collaboration. Analyzing how some 191 thousand software developers in the United States collaborate on the largest online open-source code repository platform, I find 79.8% of users clustering in only ten economic areas. Conditional on economic-area characteristics, colocated users collaborate about nine times as much as non-colocated users. Apart from this colocation effect, distance is not significantly related to collaboration among software developers. Comparison to social networks shows the colocation effect is weaker for software developers and relative connectedness probability remains at a much higher (stable) level with increasing distance. Software developer and social networks show no significant regional overlap.

JEL Code: L84, O18, O30, R32
Keywords: Geography, digitization, online, open-source, high-skilled, collaboration

Moritz Goldbeck
ifo Institute – Leibniz Institute for
Economic Research
at the University of Munich,
University of Munich
Poschingerstr. 5
81679 Munich, Germany
goldbeck@ifo.de

* I thank Lena Abou El-Komboz, Oliver Falck, Edward Glaeser, Ricardo Hausmann, Johannes Stroebel, and Johannes Wachs for valuable comments and suggestions. I am grateful to Lena Abou El-Komboz and Thomas Fackler for sharing data from other projects. Further, I thank Svenja Schwarz and Gustav Pirich for excellent research assistance and gratefully acknowledge public funding through DFG CRC/TRR 190.

1 Introduction

Digital tools and the ICT revolution allow shifting collaboration entirely into the digital space leading to the "death of distance." This hypothesis has been prominently put forward by Cairncross (1997) at the heyday of the IT boom and has recently gained traction again through Baldwin (2019) while being further fueled by the rapid uptake of remote work during the pandemic. However, compelling empirical evidence supporting this view is scant, while there are numerous studies finding increased spatial concentration of knowledge-intensive economic activity in few large centers (see, e.g., Chattergoon and Kerr, 2022; Moretti, 2021; Catalini, 2018; Forman et al., 2016). Scholars proposed various explanations for this, including the importance of face-to-face interaction (Atkin et al., 2022; Battiston et al., 2021), positive industry-cluster spillovers (Greenstone et al., 2010), and size benefits of local labor markets (Dauth et al., 2022; Manning and Petrongolo, 2017).

This paper investigates the role of geographic proximity for collaboration of software developers in the United States, an already highly digitized occupation thus featuring high potential for remote collaboration. Drawing on detailed network data from the largest online code repository platform, I analyze regional concentration and collaboration patterns of about 191 thousand software developers in open-source projects between 2015 and 2021. In a first step, I provide descriptive evidence and fit gravity-type regression models to explain spatial concentration and to distinguish the colocation effect from general relevance of increased distance. In a second step, I ask if spatial clustering differs between software developer and other types of human networks. To this end, I compare software developer to social networks by benchmarking the results using regional social connectedness data from Bailey et al. (2018b). To analyze the relationship between distance and connectedness in the two networks, I apply fractional polynomial regression analyses to region-size-independent indices.

Results show high spatial concentration with 79.8% of users clustering in only 10 of 179 economic areas. This is a slightly stronger concentration than for computer science patent inventors (68.9%) and compares to 32.2% of the population concentrating in the same economic areas. Binned scatter plots show collaboration is strongly associated with economic-area characteristics, pointing to significant spillover effects from cluster size in line with recent findings by Abou El-Komboz and Fackler (2022). Conditional on economic-area characteristics, collaboration is essentially unrelated to distance apart from a strong benefit from colocation. Holding economic-area characteristics constant, gravity-type regression analysis suggests colocation is associated with nine times higher collaboration among software developers.

The comparison with regional connectedness in social networks shows no statistically significant regional overlap of software developer and social networks in the United States. Predicting the relative probability of connectedness between economic areas with geographic distance by use of fractional polynomial regression reveals a distinct difference between the two networks. While there is spatial decay in collaboration probability in both networks, the decrease in connectedness probability in distance is confined to very low

distances in the software developer network and remains at a much higher, stable level thereafter. This points to the colocation effect being weaker and suggests a higher level of remote connectedness independent of distance in software developer networks.

Data validity checks show the data sample features high regional fit with economic-area GDP in professional, technical, and scientific services as well as the share of inventors of computer science patents. This suggests no significant geographic bias is present in the analyzed sample. Robustness checks confirm the main results and show that they are robust to increased model flexibility and inclusion of geographically close economic areas in the definition of colocation.

Related literature. This work relates to the effects of distance on economic activity which originated form the trade literature (Tinbergen, 1962; Bergstrand, 1985). Inspired by the gravity model, other fields adopted similar research designs and find distance relevant, e.g., in scientific research (Catalini, 2018), patenting (Jaffe et al., 1993; Thompson and Fox-Kean, 2005), and business relations (Cristea, 2011). Studies on the impact of technology on trade show that improved ICT fosters trade by enhancing market integration (Steinwender, 2018; Jensen, 2007). Research on online software development, where new ICT and digital tools are used heavily, shows strong spatial clustering in Europe (Wachs et al., 2022) and suggests increased distance matters for global collaboration, but less than for trade flows (Fackler and Laurentsyeva, 2020).[1] Brucks and Levav (2022) demonstrate in the lab that virtual interaction comes with a cognitive cost for creative idea generation. This paper is, to the best of my knowledge, the first to show a setting in which, apart from a large colocation effect, increased distance does not matter for collaboration.

Another related literature discusses benefits from geographic proximity, building on the Marshallian notion of costs for moving goods, people, and ideas. Studies confirmed existence of local spillovers, e.g., for productivity (Greenstone et al., 2010; Baum-Snow et al., 2020), in customer-supplier relationships (Ellison et al., 2010), and for knowledge transmission (De La Roca and Puga, 2017). Recent evidence shows further positive spillovers from clustering in knowledge-intensive settings, e.g., for inventor (Moretti, 2021), firm (Nagle, 2019) and software developer productivity (Abou El-Komboz and Fackler, 2022), as well as for entrepreneurship (Wright et al., 2021). Yang et al. (2022) show remote collaboration of information workers made information sharing harder. This study confirms that local characteristics are a main driver of collaboration among software developers. Results suggest more opportunities for direct collaboration (as opposed to more indirect spillovers) in larger clusters contribute to agglomeration effects, in line with Azoulay et al. (2010).

Increased data availability allows researchers to measure inter-personal connectedness in great detail and comprehensively. Bailey et al. (2018a) construct regional connectedness from *Facebook* data. Analyses of this data reveal highly local clustering in social networks (Bailey et al., 2020b) and a strong association with

[1] In computer science, there is some anecdotal evidence of a colocation effect in software development driven by face-to-face interaction (Bird et al., 2009; Al-Ani and Edwards, 2008) and papers investigating the network structure of online coding platforms (Badashian et al., 2014; Thung et al., 2013) as well as specific features of particular platforms (Blincoe et al., 2016).

travel (Bailey et al., 2020a) and trade (Bailey et al., 2021). Also drawing on *Facebook* data, Chetty et al. (2022a,b) compute social capital measures showing substantial regional variation in social connectedness between people with high and low socio-economic status. I contribute to this literature by, first, showing that there is essentially no regional overlap between software developer and social networks. Second, I show that, compared to social networks, software developers' connectedness probability features a weaker colocation effect and remains at a higher (stable) level with increasing distance thereafter.

The remainder of this paper is organized as follows. In Section 2, I provide a brief background on online collaboration in software development and present the data. Section 3 explores the role of colocation and distance in online software development collaborations and in Section 4 the observed spatial collaboration pattern is compared to social networks. Section 5 concludes.

2 Background and data

Background. While software development can in principle be done alone, it is typically a collaborative effort of teams and research suggests this is increasingly the case in all high-skilled professions as projects become more complex (Jones, 2009; Wuchty et al., 2007). Hence, collaboration is an important driver of labor productivity (Hamilton et al., 2003). While software developers often interact face-to-face with their collaborators, they could interact completely remotely. Occupation-level estimates by Dingel and Neiman (2020) report 100% of jobs in related occupations can be done remotely.[2] High potential to work remotely has been confirmed during the COVID-19 pandemic, when the IT sector ranked among the industries with the highest work-from-home take-up in the United States (Dey et al., 2020). This makes software developers a particularly interesting group to study if digital tools help to overcome geographic distance.

Digital tools for collaborative software development drastically improve the workflow of developers to work together remotely in teams via cloud-based online code repositories. These repositories are maintained by using the integrated version control software *git*. Version control with *git* can be highly customized in combination with local code repository copies and can be controlled conveniently via the native or GUI-integrated command line. *GitHub* is by far the largest online code repository platform. It was founded in 2008, reached 10 million users by 2015, and in 2021 reported 73 million users worldwide (GitHub, 2021; Startlin, 2016). Since many developers routinely engage in open-source software development, a large number of repositories are public. Survey evidence generated by *GitHub* in 2021 suggests that approximately 19% of code contributions on the platform are to open-source projects (GitHub, 2021). Due to the nature of the version control system *git*, a detailed history of code changes and contributing users is available and openly visible online for public repositories. *GitHub* provides access to public user profiles and repositories via API.

[2]Related SOC occupations include, e.g., Computer and Information Research Scientists, Computer Systems Analysts, Computer Programmers, Software Developers (Applications), Software Developers (Systems Software), Web Developers, and Database Architects.

Data. Data analyzed in this paper originates from *GHTorrent*, a research project by Gousios (2013) that mirrors the data publicly available via the *GitHub* API and generates a queryable relational database in irregular time intervals.[3] The resulting snapshots contain data from public user profiles and repositories as well as the detailed activity stream capturing all contributions to and events in public repositories. This paper relies on ten *GHTorrent* snapshots dated between 09/2015 and 03/2021, i.e., roughly one snapshot every seven months.[4] Overall, the data contains 44.1 million users worldwide. For analysis of regional collaboration patterns of software developers in the United States, the sample of *GitHub* users is selected according to three criteria: (1) a user location is available and refers to a location within the United States; (2) the user is active in the observation period; and (3) the user contributes to at least one project with another in-sample user.

On their *GitHub* online profile, users can indicate their location. This self-reported indication is voluntary and is neither verified nor restricted to real-world places by *GitHub*. It is thus difficult to examine the accuracy comprehensively. However, researching profiles online that are connected to persons due to use of real name on the platform and with known location from other sources suggests that those who make a location available on *GitHub* to a large extent provide their correct location.[5] As *GitHub* also functions as a social network for software developers, users have an incentive to report their correct location for networking purposes since they are then more easily found by their local peers.

About 5.2% of users captured in the data (2.30 million) include a self-reported location in their public user profile. Thereof, 34% (778 thousand) can be georeferenced to a location within the United States.[6] This roughly corresponds with a survey conducted by *GitHub* in 2021, reporting a share of 31.5% of users being located in North America (GitHub, 2021). Of these users located in the United States, a portion of 46% (354 thousand) is active in public repositories, which I define as contributing at least once in two time intervals between data snapshots.[7] Finally, 54% of active U.S. users contribute in at least one project to which multiple users contribute in the observation period. This leaves a sample of 190,637 active, collaborating users geolocated in the United States during the observation period from 2015 to 2021. For the remainder of this paper, I refer to users and their activity in this sample.

For the purpose of regional analysis, each user is assigned to one of 179 economic areas in the United States as defined by the *Bureau of Economic Analysis* based on the self-reported geolocation on her user

[3] *GHTorrent* data contains potentially sensitive personal information. Information considered sensitive (e.g., e-mail address or user name) has been de-identified (i.e., recoded as numeric identifiers) by data center staff prior to data analysis by the author. Data from the *GHTorrent* project is publicly available at `ghtorrent.org`.

[4] Snapshots are dated 2015/09/25, 2016/01/08, 2016/06/01, 2017/01/19, 2017/06/01, 2018/01/01, 2018/11/01, 2019/06/01, 2020/07/17, and 2021/03/06.

[5] Due to de-identification of user names, the user profiles cannot be linked to other data to a larger extent in order to verify this anecdotal impression. I perform further aggregate plausibility checks below.

[6] This processing step also confirms above impression that most users provide correct location, as non-sense locations like, e.g., "the moon," together with other locations for which georeferencing to a country was unsuccessful, only make up 1.4% of users with non-empty location.

[7] New users in the last time interval are regarded as active if they contribute in this time interval.

profile. This regional level is chosen such that it is both sufficiently detailed to study colocation and distance effects and provides an adequate level of aggregation given the number of users in each economic area. The *Bureau of Economic Analysis* economic areas define the relevant regional markets surrounding metropolitan or micropolitan statistical areas (Johnson and Kort, 2004). Economic areas are similar to metropolitan statistical areas (MSA) in most cases. To capture entire economic regions, economic areas tend to be larger than corresponding MSAs for big cities.

Summary statistics. In-sample users contribute to about 4.29 million *repositories*, i.e., open-source code projects on the platform. In total, they make roughly 97.3 million single code contributions to these projects, so-called *commits*. The most popular programming languages used on the platform are JavaScript, Python, as well as C and related languages (see Figure A.1). As typical for digital platforms, activity in *GitHub*'s open-source projects is highly skewed, meaning that only a fraction of users contributes the majority of content.[8] See Figure A.2 for a visual impression.

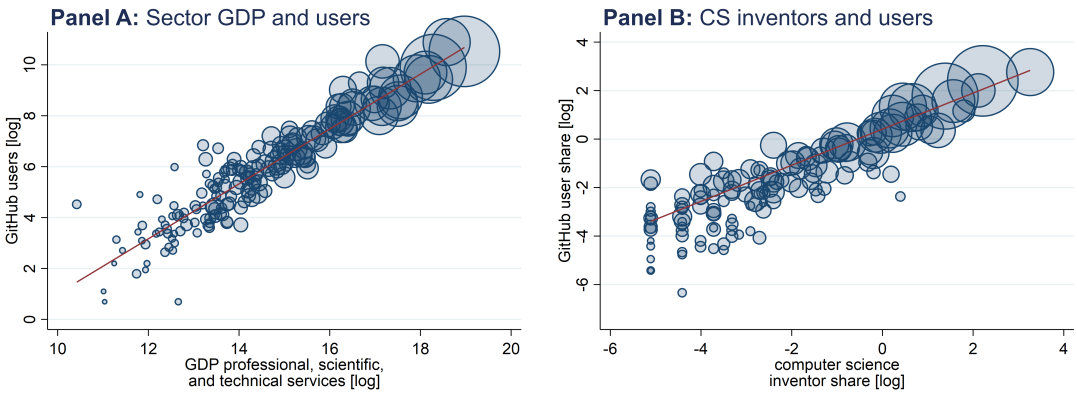
Each user on average contributes to 28.5 projects (median: 14) in the observation period. 28% of projects are one-time uploads with one (inital) *commit*. To projects that are not one-time uploads, users make on average 37.2 code contributions (median: 7). About 90% of observed projects are personal, i.e., only one user contributes to them. This leaves around 430 thousand projects run by teams. Although team projects account for only one tenth of all observed projects, they make up 45% of *commits* (≈43.3 million). Team projects have on average 3.6 (contributing) members (median: 2). In the observation period, a user on average makes 510 code contributions (median: 156), with an average of 18.4 *commits* in each of her projects (median: 3). 31% of *commits* are one-time contributions to a project.

I define users as being linked or collaborating with each other if they contribute to at least one joint project in the observation period. There are 2.94 million links between users in the sample. Each user on average is linked to 45.2 other in-sample users (median: 4). Overall, 12.4% of links are between users in the same economic area. For the average user, 34.7% of collaborations are with other local users (median: 14.3%) and two thirds of team projects are fully colocated, meaning that all contributing in-sample users are located in the same economic area. I define links between users that have more than one joint project strong ties. 19% of links between users are strong ties. More detailed summary statistics can be found in Table A.1.

Representativeness. Given the low share of users reporting a geolocation and the availability of only public user profiles and projects, I address potential concerns regarding the plausibility and representativeness of the sample by comparing the observed regional concentration patterns with other data. For this, I rely on two types of data associated with the regional concentration of knowledge workers and their activity footprint across U.S. economic areas: locations of inventors of computer science patents and regional GDP for professional, scientific, and technical services.

[8] See, e.g., Luca (2015) for a review of user content generation on social media platforms.

Figure 1: GitHub users, inventors, and GDP



Note: Plots show the relationship between (the share of) users per economic area and economic-area GDP in professional, scientific, and technical services (Panel A) and economic-area (share of) computer science inventors (Panel B). Bubble size represents economic-area population size. Red lines are best linear fits from user-weighted log-log regressions. The share of inventors in computer science by economic area is sourced from Moretti (2021). *Sources:* GHTorrent, Moretti (2021), Bureau of Economic Analysis, own calculations.

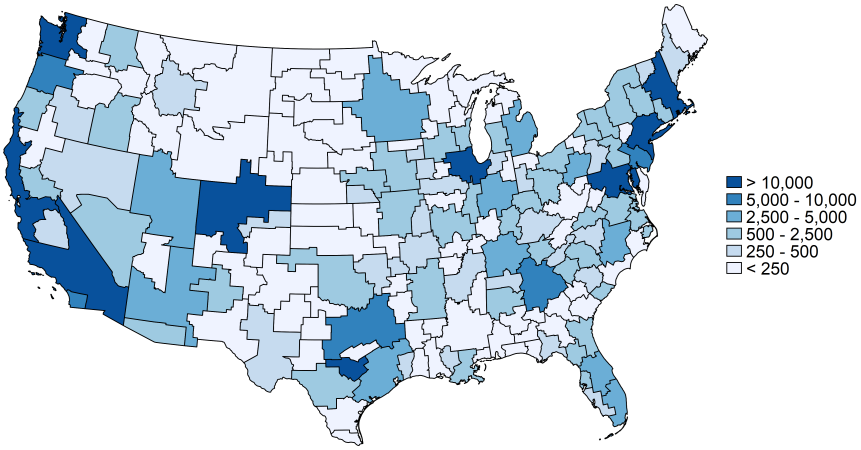
Figure 1 shows the relationship between the sampled *GitHub* users and both economic area GDP for professional, scientific, and technical services ("tech sector GDP") and inventor share of computer science patents. I find a strong positive association for both benchmarks on this regional level. Relating users to tech sector GDP in a simple user-weighted log-log regression explains 85.6% of regional variation. Similarly, the (logarithmic) share of computer science inventors explains 90% of regional variation in the (logarithmic) user share. Slope coefficients from user-weighted log-log regressions are .993 for tech sector GDP and .744 for computer science inventor share, both highly significant. This mitigates concerns regarding potential regional bias in the sample.

3 The role of colocation and distance for collaboration

Users are extremely concentrated in space. Figure 2 maps the number of active, collaborating users with a geolocation in the United States for each economic area. 79.8% of users concentrate in ten economic areas, all of which contain (at least) one major city: San Francisco, New York, Seattle, Los Angeles, Boston, Chicago, Washington D.C., Denver, Austin, and Atlanta. This is an even higher concentration in the largest hubs relative to inventors of computer science patents, where 68.9% cluster in the respective ten largest economic areas (Moretti, 2021). For comparison, the largest ten economic areas in terms of users account for only 32.2% of U.S. inhabitants.

Concentration is high even among the largest economic areas. While the largest economic area San Jose-San Francisco-Oakland, CA, hosts over 53 thousand users, only 16.3 thousand users are located in the

Figure 2: Geographic distribution of users



Notes: Map shows the number of (in-sample) users per economic area. The remote economic areas Anchorage, AK, and Honolulu, HI, are not shown. Classification method used is quantiles with six classes. *Sources:* GHTorrent, own calculations.

fifth-largest economic area Boston-Worcester-Manchester, MA-NH, and less than nine thousand users in the tenth-largest economic area Atlanta-Sandy Springs-Gainesville, GA-AL. On average an economic area contains 1,895 users with the median economic area hosting 302 users. Normalizing these numbers by economic area population size reveals user density in the general population. Three places stand out here: San Francisco, Austin, and Seattle; all with around 0.5% (in-sample) users in terms of population. Density is less than 0.25% for all other economic areas, for most of them much lower. Collaboration, measured in terms of the number of links users in an economic area are part of relative to the total number of links, is even more concentrated at the top than users. See Figure A.3 for more complete information on the largest twenty economic areas according to these metrics.

A notable property of collaborations is the extent to which they are local. 12.4% of links are between users in the same economic area. Hubs, the ten largest economic areas in terms of users, are involved in 67.9% of cross-economic area collaborations, a number with relatively little variation across economic areas.[9] Note that this is less than their combined user share of around 80%. Panel A of Figure 3 shows that the larger an economic area, measured by total collaboration share, the more of its users' collaborations are local. This strong relationship can be intuitively explained by increased opportunity for collaboration in a larger pool of users.

In Panel B of Figure 3 I benchmark this against a naïve hypothetical I call "flat world," a situation where links occur with equal probability irrespective of geography. "Local collaboration bias" divides the observed share of local collaborations of an economic area by its total share of collaborations. This means that an economic area part of 20% of total collaborations for which 20% of collaborations are local features a local

[9] See Figure A.4 for a distribution plot.

collaboration bias of one (with logarithm zero). Naturally, this measure is rarely smaller than one.[10] Local collaboration bias is larger the greater the share of local collaborations of an economic area relative to its size. Interestingly, Panel B of Figure 3 shows local bias is strongly negatively related to size. This implies while larger regions feature a greater their share of local collaborations, this effect is not proportional. Smaller economic areas, with respect to their size, disproportionately collaborate more with other local users. Overall, these descriptive findings suggests a high importance of being colocated for collaboration.

To assess the role of colocation and distance in collaborations, I calculate the geographic distance between economic areas and the number of links between each economic area pair. Distance between economic areas is computed as the (geodesic) distance in kilometers between the centroids of each economic area pair. Figure A.5 plots the histogram of distances between economic areas. The number of links (collaborations) between each economic area pair represents the count of connections between users in both economic areas as defined by active contribution to at least one project in the observation period.

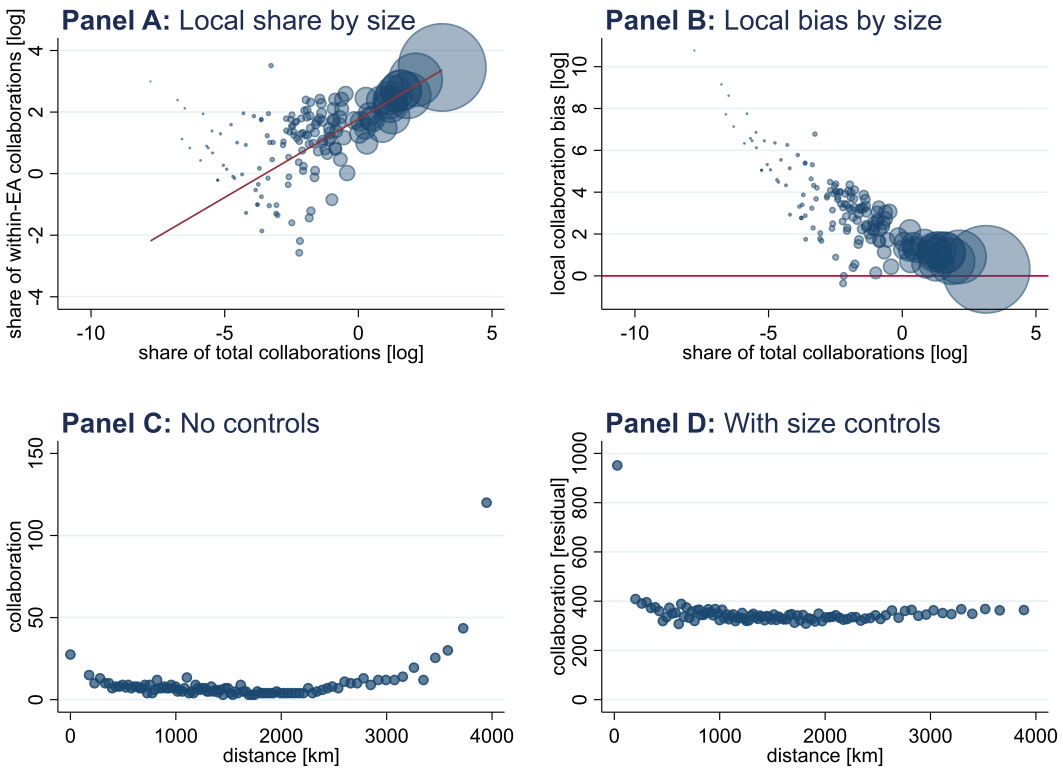
I study the relationship between distance and collaboration by constructing binned scatter plots. Panel C of Figure 3 shows a binned scatter plot for the median number of links between economic areas depending on geographic distance, with one point for each percentile. The graph shows a U-shaped relationship with a stronger increase in collaborations on the right side. I hypothesise this is likely driven by the collaborations between the large hubs on opposite coasts. Therefore, I construct another binned scatter plot (Panel D) after controlling for a set of variables measuring user size of each economic area pair: the number of users and users squared (to allow for nonlinear effects) for the two economic areas, respectively, and the number of users multiplied for each economic area pair as a representation of bilateral collaboration potential.

The results confirm the conjecture that the U-shape in Panel C is driven by the hubs' locations on opposite coasts. In Panel D, the pattern is essentially flat over the whole distance range, with the notable exception being the first distance percentile, for which (residual) collaborations are much higher. The mean distance between economic-area centroids in the first distance percentile is 28.6km.[11] Excluding the first percentile, residual medians range between 308 and 409 with a mean of 343. Being colocated (i.e., in the first distance percentile) increases median collaboration by a factor of 2.8 relative to the mean of other percentiles to a (residual) collaboration median of 951, conditional on user size controls. This suggests that, for region pairs with similar cluster size, being colocated is associated with almost three times as many collaborations for the median user.

[10] There are only two relatively small economic areas out of 179 with values smaller than one: Fayetteville-Springdale-Rogers, AR-MO-OK (.699), and Rapid City, SD-MT-ND-NT (.999).

[11] Note that this refers to centroid-based distance. Depending on the geographic expansion of economic areas actual distance between relevant users might differ.

Figure 3: (Local) collaboration and distance



Notes: Panel A shows the relationship between the share of collaborations of an economic areas users in all collaborations. The red line represents the best linear fit weighted by total collaboration share as economic area size measure. Panel B shows the deviation of the economic area user collaboration share from the benchmark of being equal to the percentage share in all collaborations. The horizontal red line (=0) represents this "flat-world" benchmark. Economic areas above the benchmark line feature a higher local collaboration share than their share in total collaborations, economic areas below the benchmark line have a lower share of local collaborations than their share in total collaborations. Bubble size indicates the collaborations of economic area users. Panels C and D show binned scatter plots of the median number of collaborations and the geographic distance between economic area pairs. The number of bins is 100, i.e., each point represents one percentile. Within economic area collaborations as well as Honolulu, HI, and Anchorage, AK, economic areas are excluded. Panel C plots the binned scatter without controls. Panel D plots the binned scatter after controlling for the following variables: users and users squared for both economic areas, respectively, and the multiplication of users of each economic area pair. Means are added back to residuals before plotting.
Sources: GHTorrent, own calculations.

To complement above analysis of the relationship between colocation, distance, and collaboration, I run simple gravity-type regression analyses of the form

$$\text{links}_{i,j} = \beta_0 + \beta_1 \mathbb{1}\{\text{coloc}_{i,j}\} + \beta_2 \text{dist}_{i,j} + \mathbf{X_i}\beta_3 + \mathbf{X_j}\beta_4 + \mathbf{X_{i,j}}\beta_5 + \varepsilon_{i,j} \tag{1}$$

where collaborations are explained by a colocation indicator marking collaboration between users in the same economic area, $\mathbb{1}\{\text{coloc}_{i,j}\}$, a distance term, and origin and destination economic-area characteris-

tics.[12] In all specifications I include the continuous centroid-based distance, $dist_{i,j}$. As control variables, I either include origin and destination economic-area characteristics, $\mathbf{X_i}$ and $\mathbf{X_j}$, or origin and destination economic-area fixed effects. Explicit controls include the number of users, GDP, and population. To control for collaboration potential between two economic areas, I further add the multiplication of origin and destination users, $\mathbf{X_{i,j}}$.

Table 1: Collaboration, colocation, and distance

Collaboration [log]	(1)	(2)	(3)	(4)	(5)	(6)
Colocation	2.921***	2.435***	2.387***	2.462***	2.373***	2.414***
	(0.227)	(0.181)	(0.180)	(0.175)	(0.157)	(0.074)
Distance	0.025***	-0.006***	-0.006***	-0.001	-0.007***	-0.004***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Users		×	×	×	×	
Users, multiplied			×	×	×	×
GDPs				×	×	
Populations					×	
Origin FE						×
Destination FE						×
Observations	31,329	31,329	31,329	31,329	31,329	31,329
Adj. R^2	0.016	0.404	0.405	0.465	0.591	0.913
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	17.56	10,41	9.87	10.73	9.73	8.92

Notes: The outcome variable is the natural logarithm of collaborations between two economic areas plus one. Colocation indicates collaboration between users in the same economic area. Distance is scaled in 100km. Users, GDPs, and Populations refers to the respective variables for both origin and destination. Users, multiplied, is the multiplication of the number of users in origin and destination. Collaboration with Anchorage, AK, and Honolulu, HI, are excluded. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, Bureau of Economic Analysis, own calculations.

The main results reported in Table 1 are stable over all specifications and confirm the findings from above analyses. Collaboration is strongly positively associated with being colocated. Effect size for colocation is large and statistically highly significant, suggesting colocated users collaborate about 8.9 to 10.7 as much as users that are not colocated, holding economic-area characteristics constant. Further, there is only a very weak, but statistically significant negative relation with distance. Depending on the specification and given equal economic-area characteristics, results suggest -0.1% to -0.7% fewer collaborations when distance increases by 100km.

Economic-area characteristics play an important role for collaboration, as shown by the naïve model in column (1) without controls. In line with the descriptive finding that a large part of collaborations happens within and between large hubs in terms of users, this specification overestimates both the role of colocation

[12] To deal with unconnected economic areas, I follow a common solution from the trade literature and avoid omission by adding one after logarithmic transformation of the number of links between each economic area pair.

and distance and is not able to explain variation in the data well. Once control variables for economic-area characteristics are subsequently added, the results remain robust and stable, while explained variation increases to around 40% with user controls and 59% with GDP and population controls. Adding origin and destination fixed effects that capture also unobserved economic-area characteristics further improves model fit to 91%. The fixed-effects model controlling for the multiplication of origin and destination users shown in column (6) is my preferred specification.

I conduct further robustness checks. For example, I vary the definition of colocation regarding to the distance cutoff, allow for more flexibility by adding various squared terms of the variables, and by including more economic area pair controls. Varying the definition of colocation by including not only users in the same economic area but also allowing for small centroid-based distances between nearby economic areas shows a comparable effect when including distances smaller than 100km and a much smaller effect when including distances smaller than 200km (see Table A.2). This points to the colocation effect being confined to small distances only and essentially vanishing thereafter, confirming findings from Panel D in Figure 3. Increasing model flexibility yields similar results (see Table A.3).

4 Spatial decay of collaboration in developer versus social networks

Human networks are generally known to have a strong spatial dimension. Due to the nature of software development as highly digitized activity with a large potential for remote collaboration, regional connectedness patterns are likely different to other types of human networks. To evaluate how the pattern between geographic distance and collaboration in online (open-source) networks of software developers relate to other human networks, I compare the results with patterns in social networks using data on regional connectedness from *Facebook*.

Connections on *Facebook* map to a large extent to real-world friendship, family and acquaintanceship ties. As such, observed regional network data constructed form active users on *Facebook* is an adequate representation of real-world social networks.[13] Social networks on *Facebook* feature a high degree of spatial clustering of connections (Bailey et al., 2018a). Bailey et al. (2018b) construct an openly available index of social connectedness on the United States' county-county level.[14] The so-called *Social Connectedness Index* (SCI) measures the relative probability of connections between users in two counties by

$$\text{index}_{i,j} = \frac{\text{links}_{i,j}}{\text{users}_i * \text{users}_j}. \tag{2}$$

[13] See Bailey et al. (2018a) for a detailed discussion.

[14] Data is retrieved online via data.humdata.org/dataset/social-connectedness-index.

Importantly, the index is independent of region size and scaled to numbers between 1 and 1,000,000,000.[15] I aggregate the county-county SCI to the economic-area pair level by using multiplied county population size as weights, since user counts are not available in the public data. After aggregation I rescale the index again. Further, I similarly compute a scaled index using the *GHTorrent* data sample, which I call *GH Connectedness Index* (GHCI). Figure A.6 provides a comparison of the two indices' key properties by depicting their histograms. Notable is the right-skewed distribution of both indices, which is stronger in the SCI. This is also reflected in the medians, which sits at maximum frequency for the GHCI and is shifted towards the right for the SCI. The fat right tails suggest many economic-area pairs feature a relatively low connectedness while few pairs are highly interconnected. Since the SCI has a fatter right tail, more economic-area pairs are highly connected as compared to the GHCI.

The two regional connectedness indices are essentially orthogonal to each other, with a low Pearson's correlation of 0.0248 that is not statistically significantly distinguishable from zero. This is also shown by Panel D of Figure 4 and a data example for the Los Angeles-Long Beach-Riverside, CA, economic area in Figure A.7 provides an illustration. While the (weighted) number of collaborations on *GitHub* is strongly associated with large clusters, this relationship vanishes for the GHCI since it is constructed analogous to the SCI and therefore is independent of economic-area size. This shows that software developer and general friendship networks measured through size-independent indices such as GHCI and SCI feature no significant regional overlap.

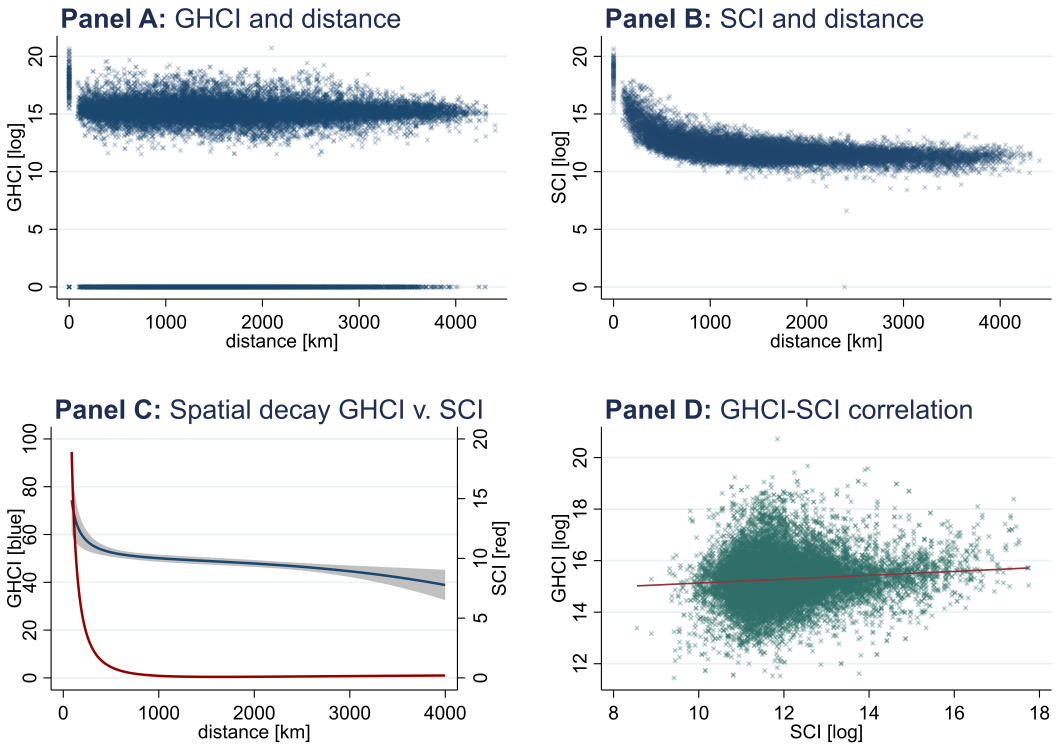
Data confirms the presence of a strong colocation effect in both networks. Panels A and B of Figure 4 plot raw data from scaled GHCI (Panel A) and SCI (Panel B) after logarithmic transformation. A large colocation effect is already visible in the raw data by the sharp upward shift of the (logarithmic) distribution at a distance of zero for both indices. Apart from the colocation effect, GHCI is essentially independent of distance, in line with the findings in Section 3. In contrast, the SCI features strong spatial clustering as depicted by the continued decrease over the whole distance range. The decrease in social connectedness with increasing distance is particularly strong for distances smaller than 500km.

For a model-based comparison of the relationship of the indices to geographic distance, I fit fractional polynomial regressions to flexibly model the relationship in the data. Panel C of Figure 4 graphs the predicted relationships. The fitted curve in blue, with a larger confidence interval, represents the relationship between the scaled GHCI and geographic distance while the fitted curve in red shows the same relationship for the scaled SCI. Spatial decay of the relative probability of a connection is present in both indices. It is, however, much more pronounced for predicted SCI, as shown by the steep and immediate decline of the index level, converging to very low index levels at a distance of 1,000km.[16] The predicted GHCI is also characterized by a steep decline in index level in the beginning, but the decline is less pronounced and values level off

[15] To (re)scale the indeces, following formula is applied: $\text{index} \rightarrow \frac{\text{index}-\min(\text{index})}{\max(\text{index})-\min(\text{index})} * [\max(\text{scale}) - \min(\text{scale})] + \min(\text{scale})$.

[16] The slight increase of the predicted SCI towards the right end reflects the connections between opposite coasts.

Figure 4: Relative collaboration probability and distance



Note: Panels A and B show scattered values of scaled GHCI and SCI after logarithmic transformation. Both indices are scaled between 1 and 1,000,000,000. Scaled SCI from Bailey et al. (2018b) is mean-aggregated from county-county level weighted by multiplied populations of each county-pair and rescaled between 1 and 1,000,000,000. Panel C shows the predicted relationship between scaled GHCI/SCI indices and distance as estimated by a fractional polynomial regression. Gray areas represent the 95% confidence interval. For purpose of readability I divide scaled GHCI by 100,000 and scaled SCI by 1,000,000. Panel D shows the correlation between scaled GHCI and SCI after logarithmic transformation with within-economic-area collaborations excluded. *Sources:* GHTorrent, Bailey et al. (2018b), Bureau of Economic Analysis, own calculations.

earlier (at around 300km). Afterwards, the predicted GHCI remains stable over a large part of the distance range and declines slowly and gradually starting at a distance of approximately 3,000km.

Underlying populations differ in both indices. GHCI refers to open-source software developers on *GitHub* and SCI more to the general population (for a discussion, see Bailey et al., 2018a). Recognizing this, a logical question is how much of the difference in the observed distance pattern is driven by composition differences in underlying populations. Software developers tend to be working-age, urban and more educated than the general population, but a separate SCI for a population with these characteristics is not publicly available. However, Bailey et al. (2018b) provide *Facebook* users' share of friends within 50, 200, and 500 miles distance for users with "some college" education, urban users, and users aged 35-55, as well as the full sample. These statistics show only small differences for these subgroups compared to the full sample,

with differences averaging 1.1%. This mitigates concerns that the observed difference in distance pattern is driven to a larger extent by composition differences in index populations.

Overall, although there is spatial decay in connectedness present in both indices, this analysis suggests that distance is much more relevant for social networks in general as compared to for software developer networks in the online open-source community. Further, decline in connectedness is fastest in the beginning and then levels off in both networks, but remains at a much higher level in software developer networks. In addition, the drop is much smaller, more sudden, and leveling off happens earlier in software developer networks suggesting fewer benefits from colocation to collaboration that are present only for small distances.

5 Conclusion

The results of investigating open-source software developer networks in the United States emphasize the overpowering importance of colocation for collaboration – even in the digital space and for a highly digitized occupation. However, apart from this colocation effect and in line with the long standing prediction that digital tools help to overcome geographic frictions in knowledge-driven sectors, I find strong evidence of further increased distance being only weakly associated with collaboration. These findings show that there is a setting where digital tools do indeed contribute to the "death of distance," but to date cannot replace the huge benefits for collaboration of being colocated.

Even in online collaboration networks of software developers data shows strong spatial concentration in few large clusters, driven to a large extent by local characteristics such as cluster size in terms of users, GDP, and population, as well as collaboration potential between economic areas. While this research cannot pinpoint what exactly causally explains the apparently present agglomeration effects, I find suggestive evidence that direct collaboration with other local users plays an important role. This does, of course, not rule out benefits from other (more indirect) local spillovers from interactions not observed here such as, e.g., informal exchange and networks or chance encounters.

This work implies that geographic clusters maintain enormous advantage compared to smaller regions due to benefits from colocation and cluster size, likely in part driven by improved direct local collaboration opportunities. However, there is no significant strong barrier from distance for collaboration among software developers in online open-source development. This allows software developers to engage with non-local peers even from remote regions. The big open question for non-cluster regional development is to identify drivers of the colocation effect and try to recreate similar experiences remotely.

References

Abou El-Komboz, Lena and Thomas Fackler, "Productivity Spillovers among Knowledge Workers in Agglomerations: Evidence from GitHub," *Working Paper*, 2022.

Al-Ani, Ban and H. Keith Edwards, "A Comparative Empirical Study of Communication in Distributed and Collocated Development Teams," in "IEEE International Conference on Global Software Engineering" 2008, pp. 35–44.

Atkin, David, M. Keith Chen, and Anton Popov, "The Returns to Face-to-Face Interactions: Knowledge Spillovers in Silicon Valley," *NBER Working Paper 30147*, 2022.

Azoulay, Pierre, Joshua S. Graff Zivin, and Jialan Wang, "Superstar Extinction," *The Quarterly Journal of Economics*, 2010, *125* (2), 549–589.

Badashian, Ali Sajedi, Afsaneh Esteki, Ameneh Gholipour, Abram Hindle, and Eleni Stroulia, "Involvement, Contribution and Influence in GitHub and Stack Overflow.," in "CASCON," Vol. 14 2014, pp. 19–33.

Bailey, Michael, Abhinav Gupta, Sebastian Hillenbrand, Theresa Kuchler, Robert Richmond, and Johannes Stroebel, "International Trade and Social Connectedness," *Journal of International Economics*, 2021, *129*, 103418.

_ , Drew Johnston, Theresa Kuchler, Dominic Russel, Johannes Stroebel et al., "The Determinants of Social Connectedness in Europe," in "International Conference on Social Informatics" 2020, pp. 1–14.

_ , Patrick Farrell, Theresa Kuchler, and Johannes Stroebel, "Social Connectedness in Urban Areas," *Journal of Urban Economics*, 2020, *118*, 103264.

_ , Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong, "Social Connectedness: Measurement, Determinants, and Effects," *Journal of Economic Perspectives*, 2018, *32* (3), 259–80.

_ , Ruiqing Cao, Theresa Kuchler, and Johannes Stroebel, "The Economic Effects of Social Networks: Evidence from the Housing Market," *Journal of Political Economy*, 2018, *126* (6), 2224–2276.

Baldwin, Richard, *The Globotics Upheaval: Globalization, Robotics, and the Future of Work*, Oxford University Press, 2019.

Battiston, Diego, Jordi Blanes i Vidal, and Tom Kirchmaier, "Face-to-Face Communication in Organizations," *The Review of Economic Studies*, 2021, *88* (2), 574–609.

Baum-Snow, Nathaniel, Nicolas Gendron-Carrier, and Ronni Pavan, "Local Productivity Spillovers," *Working Paper*, 2020.

Bergstrand, Jeffrey H., "The Gravity Equation in International Trade: Some Microeconomic Foundations and Empirical Evidence," *The Review of Economics and Statistics*, 1985, pp. 474–481.

Bird, Christian, Nachiappan Nagappan, Premkumar Devanbu, Harald Gall, and Brendan Murphy, "Does Distributed Development Affect Software Quality? An Empirical Case Study of Windows Vista," in "IEEE 31st International Conference on Software Engineering" 2009, pp. 518–528.

Blincoe, Kelly, Jyoti Sheoran, Sean Goggins, Eva Petakovic, and Daniela Damian, "Understanding the Popular Users: Following, Affiliation Influence and Leadership on GitHub," *Information and Software Technology*, 2016, *70*, 30–39.

Brucks, Melanie S and Jonathan Levav, "Virtual communication curbs creative idea generation," *Nature*, 2022, *605* (7908), 108–112.

Cairncross, Frances, *The Death of Distance: How the Communications Revolution Will Change Our Lives*, Harvard Business School Press, 1997.

Catalini, Christian, "Microgeography and the Direction of Inventive Activity," *Management Science*, 2018, *64* (9), 4348–4364.

Chattergoon, Brad and William R. Kerr, "Winner Takes All? Tech Clusters, Population Centers, and the Spatial Transformation of US Invention," *Research Policy*, 2022, *51* (2), 104418.

Chetty, Raj, Matthew O. Jackson, Theresa Kuchler, Johannes Stroebel, Nathaniel Hendren, Robert B. Fluegge, Sara Gong, Federico Gonzalez, Armelle Grondin, Matthew Jacob et al., "Social Capital I: Measurement and Associations with Economic Mobility," *Nature*, 2022, *608* (7921), 108–121.

_ , _ , _ , _ , _ , _ , _ , _ , _ , _ , and _ , "Social Capital II: Determinants of Economic Connectedness," *Nature*, 2022, *608* (7921), 122–134.

Cristea, Anca D., "Buyer-Seller Relationships in International Trade: Evidence from US States' Exports and Business-Class Travel," *Journal of International Economics*, 2011, *84* (2), 207–220.

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

De La Roca, Jorge and Diego Puga, "Learning by Working in Big Cities," *The Review of Economic Studies*, 2017, *84* (1), 106–142.

Dey, Matthew, Harley Frazis, Mark A. Loewenstein, and Hugette Sun, "Ability to Work from Home: Evidence from Two Surveys and Implications for the Labor Market in the COVID-19 Pandemic.," *Bureau of Labor Statistics Monthly Labor Review*, 2020.

Dingel, Jonathan I and Brent Neiman, "How many jobs can be done at home?," *Journal of Public Economics*, 2020, *189*, 104235.

Ellison, Glenn, Edward L. Glaeser, and William R. Kerr, "What Causes Industry Agglomeration? Evidence from Coagglomeration Patterns," *American Economic Review*, 2010, *100* (3), 1195–1213.

Fackler, Thomas and Nadzeya Laurentsyeva, "Gravity in Online Collaborations: Evidence from GitHub," *CESifo Forum*, 2020, *21* (03), 15–20.

Forman, Chris, Avi Goldfarb, and Shane M. Greenstein, "Agglomeration of Invention in the Bay Area: Not Just ICT," *American Economic Review*, 2016, *106* (5), 146–51.

GitHub, "The 2021 State of the Octoverse," 2021.

Gousios, Georgios, "The GHTorent Dataset and Tool Suite," in "IEEE 10th Working Conference on Mining Software Repositories (MSR)" 2013, pp. 233–236.

Greenstone, Michael, Richard Hornbeck, and Enrico Moretti, "Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings," *Journal of Political Economy*, 2010, *118* (3), 536–598.

Hamilton, Barton H., Jack A. Nickerson, and Hideo Owan, "Team Incentives and Worker Heterogeneity: An Empirical Analysis of the Impact of Teams on Productivity and Participation," *Journal of Political Economy*, 2003, *111* (3), 465–497.

Jaffe, Adam B., Manuel Trajtenberg, and Rebecca Henderson, "Geographic Localization of Knowledge Spillovers as Evidenced by Patent Citations," *The Quarterly Journal of Economics*, 1993, *108* (3), 577–598.

Jensen, Robert, "The Digital Provide: Information (Technology), Market Performance, and Welfare in the South Indian Fisheries Sector," *The Quarterly Journal of Economics*, 2007, *122* (3), 879–924.

Johnson, Kenneth P. and John R. Kort, "2004 Redefinition of the BEA Economic Areas," *Survey of Current Business*, 2004, *75* (2), 75–81.

Jones, Benjamin F., "The Burden of Knowledge and the "Death of the Renaissance Man": Is Innovation Getting Harder?," *The Review of Economic Studies*, 2009, *76* (1), 283–317.

Luca, Michael, "User-Generated Content and Social Media," in "Handbook of Media Economics," Vol. 1, Elsevier, 2015, pp. 563–592.

Manning, Alan and Barbara Petrongolo, "How Local are Labor Markets? Evidence from a Spatial Job Search Model," *American Economic Review*, 2017, *107* (10), 2877–2907.

Moretti, Enrico, "The Effect of High-Tech Clusters on the Productivity of Top Inventors," *American Economic Review*, 2021, *111* (10), 3328–75.

Nagle, Frank, "Open-Source Software and Firm Productivity," *Management Science*, 2019, *65* (3), 1191–1215.

Startlin, "History of GitHub," 2016.

Steinwender, Claudia, "Real Effects of Information Frictions: When the States and the Kingdom Became United," *American Economic Review*, 2018, *108* (3), 657–96.

Thompson, Peter and Melanie Fox-Kean, "Patent Citations and the Geography of Knowledge Spillovers: A Reassessment," *American Economic Review*, 2005, *95* (1), 450–460.

Thung, Ferdian, Tegawende F. Bissyande, David Lo, and Lingxiao Jiang, "Network Structure of Social Coding in GitHub," in "IEEE 17th European Conference on Software Maintenance and Reengineering" 2013, pp. 323–326.

Tinbergen, Jan, "An Analysis of World Trade Flows," *Shaping the World Economy*, 1962, *3*, 1–117.

Wachs, Johannes, Mariusz Nitecki, William Schueller, and Axel Polleres, "The Geography of Open-Source Software: Evidence from GitHub," *Technological Forecasting and Social Change*, 2022, *176*, 121478.

Wright, Nataliya, Frank Nagle, and Shane M. Greenstein, "Open-Source Software and Global Entrepreneurship," *Harvard Business School Technology & Operations Management Unit Working Paper 20-139*, 2021.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi, "The Increasing Dominance of Teams in Production of Knowledge," *Science*, 2007, *316* (5827), 1036–1039.

Yang, Longqi, David Holtz, Sonia Jaffe, Siddharth Suri, Shilpi Sinha, Jeffrey Weston, Connor Joyce, Neha Shah, Kevin Sherman, Brent Hecht et al., "The Effects of Remote Work on Collaboration among Information Workers," *Nature Human Behaviour*, 2022, *6* (1), 43–54.

A Appendix

Table A.1: Summary statistics

Statistic	Mean	Median	Min	Max	N
Users					
Projects per user	28.51	14	1	46,508	190,637
Links per user	123.65	7	1	14,739	190,637
Commits per user	510.42	156	1	388,287	190,637
Commits per user-project	18.40	3	1	364,397	5,286,886
Projects					
Commits per project	22.64	3	1	364,397	4,298,045
per personal project	13.97	3	1	364,397	3,867,611
per team project	100.52	18	2	209,214	430,435
Users per team project	3.64	2	2	147,236	430,435
Economic areas					
Users per economic area	1,895	302	2	53,818	179
Projects per economic area	26,924	3,328	4	831,728	179
Links per economic area	130,562	15,329	1	5,175,727	179
Links per economic-area pair	930	23	1	1,550,463	25,135
Commits per economic area	543,600	69,185	19	19,165,952	179

Notes: All statistics refer to the final sample of 190,637 active, collaborating users geolocated in the United States and retrieved from ten data snapshots dated between 09/2015 and 03/2021. Means are rounded to two decimal places for user and project statistics and to integers for economic-area statistics. Team projects are projects with more than one contributing user in the observation period and personal projects are projects with only one contributing user in the observation period. *Commits* per user-project is the number of *commits* to each project by each contributing user. Links refers to connections between users as defined by contributing to at least one joint project in the observation period. Links per economic-area pair excludes 6,906 ($= 2^{179} - 25,135$) unconnected economic-area pairs. *Sources:* GHTorrent, own calculations.

Table A.2: Robustness checks: colocation

Collaboration [log]	distance cutoff		
	(1)	(2)	(3)
	$= 0$ km	< 100 km	< 200 km
Colocation	2.414***	2.238***	0.908***
	(0.074)	(0.083)	(0.052)
Distance	-0.004***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)
Users, multiplied	×	×	×
Origin FE	×	×	×
Destination FE	×	×	×
Observations	31,329	31,329	31,329
Adj. R^2	0.913	0.912	0.909
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	10.18	8.38	1.48

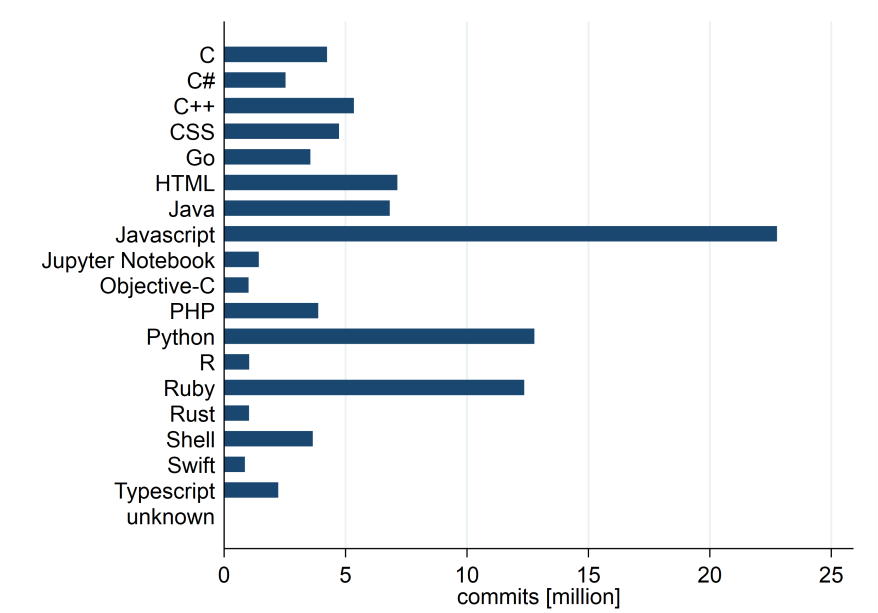
Notes: Model (1) is the preferred (fixed-effects) specification from Table 1, defining colocation as indicator of being in the same economic area. Models (2) and (3) extend this definition of colocation to include centroid-based distances of 100km and 200km, respectively. The outcome variable is the natural logarithm of collaborations between two economic areas plus one. Colocation indicates collaboration between users in the same economic area. Distance is scaled in 100km. Users, multiplied, is the multiplication of the number of users in origin and destination. Collaboration with Anchorage, AK, and Honolulu, HI, are excluded. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, Bureau of Economic Analysis, own calculations.

Table A.3: Robustness checks: model flexibility

Collaboration [log]	(1)	(2)	(3)	(4)
Colocation	2.295***	2.353***	2.433***	2.277***
	(0.075)	(0.082)	(0.074)	(0.079)
Distance	-0.022***	-0.004***	-0.004***	-0.020***
	(0.002)	(0.001)	(0.001)	(0.002)
Distance, squared	0.001***			0.000***
	(0.000)			(0.000)
Users, multiplied	×	×	×	×
Users, multiplied (squared)			×	×
GDPs, multiplied		×		×
GDPs, multiplied (squared)				×
Populations, multiplied		×		×
Populations, multiplied (squared)				×
Origin FE	×	×	×	×
Destination FE	×	×	×	×
Observations	31,329	31,329	31,329	31,329
Adj. R^2	0.913	0.915	0.913	0.917
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	8.92	9.52	10.39	8.74

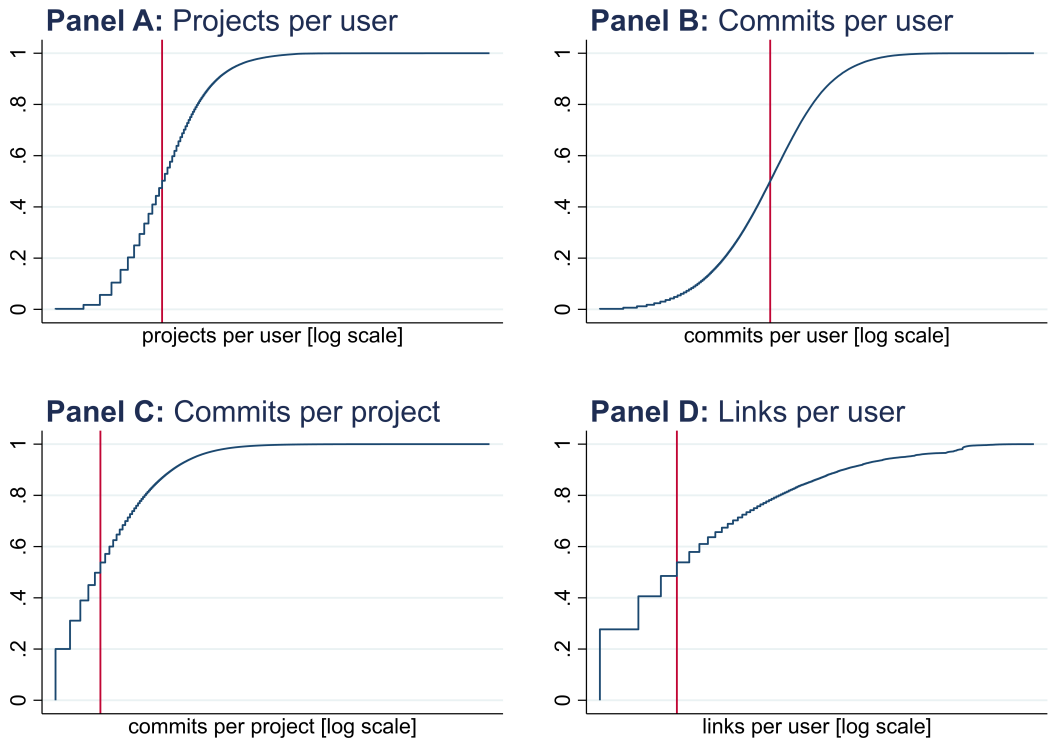
Notes: Table shows model variations allowing for increased model flexibility relative to the preferred specification in Table 1 by including: more economic-area pair characteristics and squared terms thereof as well as squared distance. The outcome variable is the natural logarithm of collaborations between two economic areas plus one. Colocation indicates collaboration between users in the same economic area. Distance is scaled in 100km. Multiplied refers to the multiplication of the respective metric in origin and destination. Multiplied (squared) refers to the squared multiplication of the respective metric in origin and destination. Collaboration with Anchorage, AK, and Honolulu, HI, are excluded. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, Bureau of Economic Analysis, own calculations.

Figure A.1: Programming languages



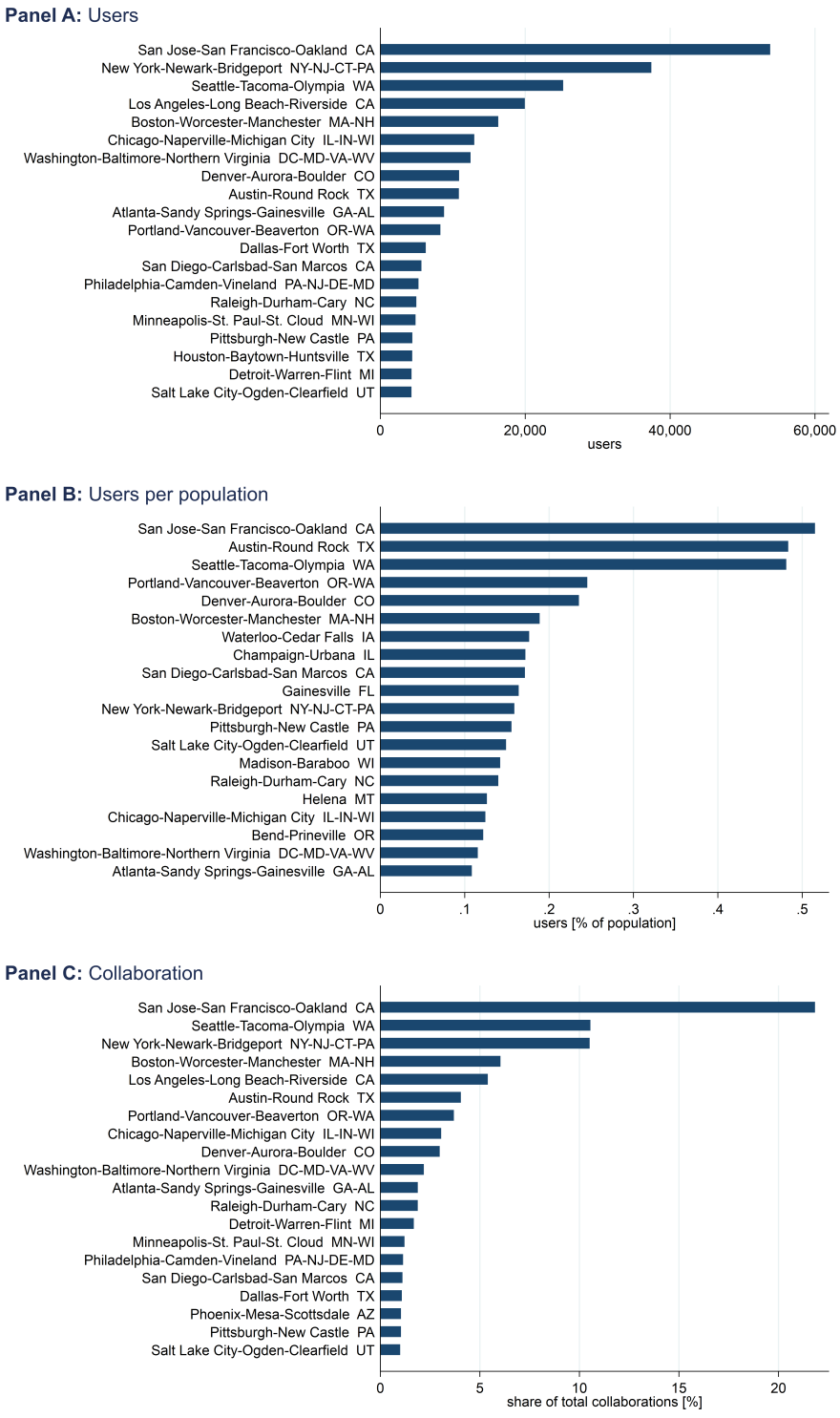
Note: Bars show the number of *commits* contributed to open-source projects by active, collaborating users in the United States in the observation period for each programming language. Unknown refers to *commits* that are not assigned to a programming language in the data. *Sources:* GHTorrent, own calculations.

Figure A.2: CDFs of user activity



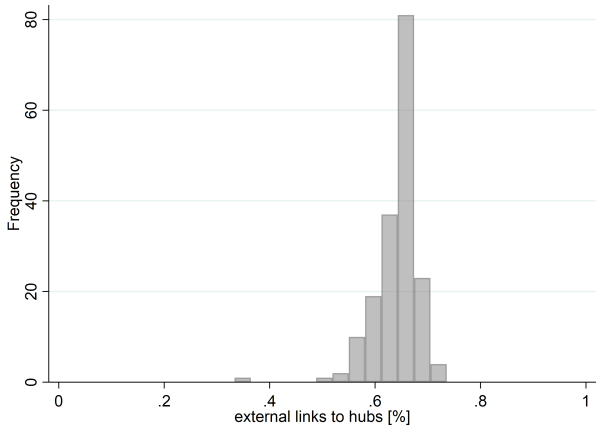
Note: Plots show cumulative density functions for different user activity metrics. Vertical red lines represent median values of each metric (i.e., projects per user: 14; *commits* per user: 156; *commits* per project: 7; links per user: 4). All x-axes are scaled logarithmically. The graph for *commits* per project excludes projects representing one-time uploads, i.e. projects with only one (initial) *commit*. *Sources:* GHTorrent, own calculations.

Figure A.3: Concentration at the top



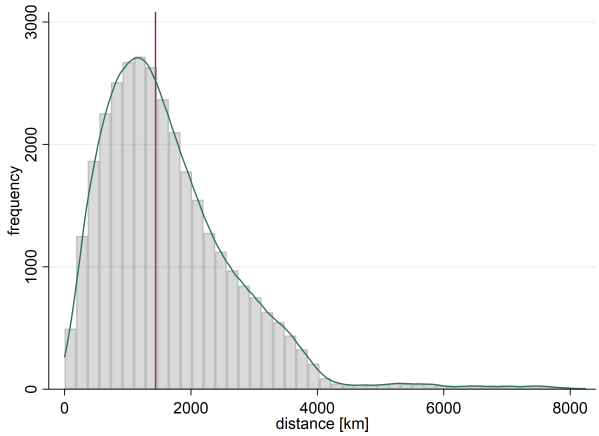
Notes: Plots show the values of different user and activity concentration metrics for the twenty largest economic areas in terms of respective metrics. *Sources:* GHTorrent, own calculations.

Figure A.4: Collaboration with hubs



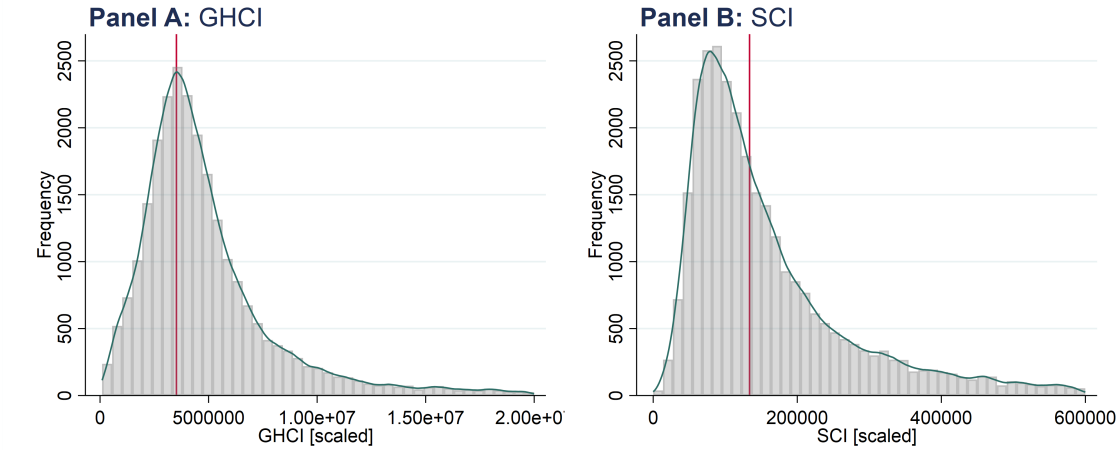
Notes: Plot shows the distribution of collaboration shares of each economic area with hubs, defined as the ten largest economic areas in terms of users. *Sources:* GHTorrent, Bureau of Economic Analysis, own calculations.

Figure A.5: Distance



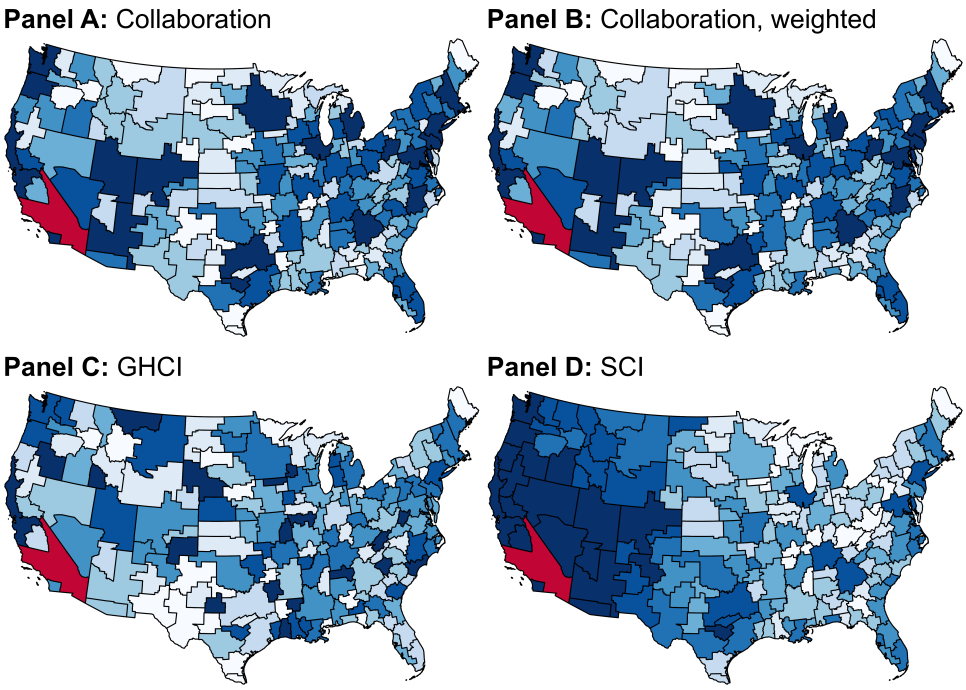
Notes: Plot shows the distribution of centroid-based geodesic distance between economic areas. The horizontal red line indicates the median distance of 1,439. The blue curve represents the Epanechnikov kernel density estimate. The right tail of the distribution starting approximately at distances greater than 4,000km is essentially driven entirely by the remote economic areas Anchorage, AK, and Honolulu, HI. *Sources:* GHTorrent, Bureau of Economic Analysis, own calculations.

Figure A.6: Histograms of scaled GHCI and SCI



Note: Plots show the distribution of scaled GHCI and SCI regional connectedness indices. The horizontal red lines indicate medians of 133,753 for the GHCI and 3,518,538 for the SCI. The blue curves represent the Epanechnikov kernel density estimates. Both indices are scaled between 1 and 1,000,000,000. Scaled SCI from Bailey et al. (2018b) is mean-aggregated from county-county level weighted by multiplied populations of each county-pair and rescaled between 1 and 1,000,000,000. As indices are highly skewed, I restrict the y-axes to maximum values of 20,000,000 for GHCI and 600,000 for SCI to achieve meaningful visualization. Scaled GHCI values of one, representing no links, are excluded from the histogram but not from the median. *Sources:* GHTorrent, Bailey et al. (2018b), Bureau of Economic Analysis, own calculations.

Figure A.7: Data example for Los Angeles-Long Beach-Riverside, CA



Notes: Maps show the connectedness of the Los Angeles-Long Beach-Riverside, CA, economic area with other U.S. economic areas according to different indicators. Anchorage, AK, and Honolulu, HI, are not shown. The classification method used for scaling is quantile with nine classes. Link weights used in the Panel B are the number of joint projects. *Sources:* GHTorrent, Bailey et al. (2018b), own calculations.