Bit by Bit

Colocation and the Death of Distance in Software Developer Networks

Moritz Goldbeck*

October 4, 2024

Abstract

Digital work settings potentially facilitate remote collaboration and thereby decrease geographic friction in knowledge work. Here, I analyze spatial collaboration patterns of some 191 thousand software developers in the United States on the largest code repository platform *GitHub*. Using a gravity framework that accounts for cluster size, I show that colocated developers collaborate about nine times as much as non-colocated developers. This colocation effect is much smaller than in less digital social or inventor networks. Additionally, further increased geographic distance is of little relevance to collaboration. Heterogeneity analyses demonstrate the colocation effect is smaller within larger organizations, for high-quality projects, among experienced developers, and for sporadic interactions. Overall, this suggests geographic proximity is indeed less important for collaboration in a digital work setting.

Keywords: geography, digitalization, networks, knowledge economy, colocation

JEL-Codes: L84, O18, O30, R32

*ifo Institute & University of Munich; goldbeck@ifo.de.

I thank Lena Abou El-Komboz, Gabriel Ahlfeldt, Dany Bahar, Raj Chetty, Thomas Fackler, Oliver Falck, Lisandra Flach, Richard Freeman, Ed Glaeser, Shane Greenstein, Ricardo Hausmann, Anna Kerkhof, Bill Kerr, Frank Nagle, Giacomo De Nicola, Megan MacGarvie, Claudia Steinwender, Johannes Stroebel, Enrico Vanino, and Johannes Wachs as well as seminar participants at the 6th CRC Rationality and Competition Retreat, Harvard Growth Lab, ifo Institute, the 2nd CESifo Workshop on Big Data and the 12th European Meeting of the Urban Economics Association for valuable comments and suggestions. I am grateful to Lena Abou El-Komboz and Thomas Fackler for sharing data. Further, I thank Raunak Mehrotra, Svenja Schwarz and Gustav Pirich for excellent research assistance and gratefully acknowledge public funding through DFG grant number 280092119.

1 Introduction

Digitization 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. Unlike previous transformations in the labor market, online collaboration affects especially white-collar occupations in the knowledge economy that are driving innovation and thus long-run economic growth (Romer, 1986; Harrigan et al., 2021, 2023). However, compelling empirical evidence supporting the 'death of distance' hypothesis is scant, while there are numerous studies finding increased spatial concentration of knowledge-intensive economic activity in a few large centers (see, e.g., Chattergoon and Kerr, 2022; Moretti, 2021; 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 (Arkolakis et al., 2023; Greenstone et al., 2010), and benefits from local labor market size (Moretti and Yi, 2023; Dauth et al., 2022; Manning and Petrongolo, 2017). Still, with digital tools rapidly evolving and their growing adoption, it remains an open question to what extent 'distance is dying.'

Knowledge work is expected to be particularly susceptible to the 'death of distance' since many tasks are already digitized. Here, study software development as an integral and increasingly important part of the knowledge economy: software is not only a key sector on its own (Korkmaz et al., 2024) but also an omnipresent element to other products (Nagle, 2019; Andreessen, 2011). Yet, comprehensive empirical evidence on spatial collaboration of software developers is lacking.[1] Software development also is characteristic for knowledge work more generally since it is typically a collaborative effort, which research suggests is increasingly the case in all high-skilled professions as work becomes more specialized and complex (Jones, 2009; Wuchty et al., 2007). This makes collaboration an important driver of high-skilled labor productivity (Hamilton et al., 2003; Simon, 1979; Arrow, 1974). Additionally, even within the knowledge economy, the 'death of distance' hypothesis applies particularly strongly to software development for two reasons: First, software development is already routinely performed using an ecosystem of digital tools that facilitate cloud-based collaborative development in teams. Thus, it is a prototypical setting where collaboration theoretically can be shifted completely into the virtual space (Emanuel et al., 2023).[2] Second, software development is by nature codified to a higher degree than other knowledge work, which facilitates knowledge transmission over distance (Carlino and Kerr, 2015).

[1] The main reasons for this are that software is generally harder to patent and easy to keep as a trade secret, and therefore incompletely and selectively observed in widely-used patent data (Jedrusik and Wadsworth, 2017).

[2] Occupation-level estimates by Dingel and Neiman (2020) report 100% of jobs in related occupations can be done remotely. 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. 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).

In this article, I ask if there is empirical evidence of a subdued relevance of geographic distance for collaboration in software development. Drawing on detailed georeferenced network data from the largest code repository platform, *GitHub*, I analyze regional collaboration patterns of some 191 thousand U.S. software developers in public projects between 2015 and 2021. I focus on the U.S. as a large and integrated market with relatively few cultural and language barriers and thus lower barriers to collaboration across space. The data is representative of the overall activity of software developers and offers unique and comprehensive insights into the industries' production process. In a first step, I estimate non-parametric and gravity-type regression models to explain spatial collaboration patterns and distinguish the colocation effect from the general relevance of increased distance and cluster size. In a second step, I compare the observed patterns to two other networks that are arguably less digital, albeit to a different degree: the (computer science) inventor network and the social network. A third step aims to unravel potential drivers of the observed spatial collaboration pattern. To this end, I leverage detailed information on the type of collaboration and individuals' characteristics to estimate the group-specific colocation effect depending on organizational affiliation, user and project characteristics, as well as collaboration intensity and quality.

I find colocation is on average associated with about nine times higher collaboration among software developers, conditional on economic-area characteristics. Further increases in geographic distances are of little importance to collaboration. Although the colocation effect in digital knowledge work is sizable, compared to less digital networks it is relatively small. First, the colocation effect in the closely related collaboration network of computer science inventors is about three times larger while both networks feature a dichotomous geographic pattern with a large colocation effect but further increased geographic distance being of little relevance. As the general mode of working and underlying population overlap, these results are in line with higher face-to-face interaction requirements as computer science inventors work on more creative, novel, and innovative projects (see, e.g., Akcigit et al., 2018). Second, the colocation effect for software developers is about four times smaller than in social networks of the general working-age population, a benchmark where physical proximity is essential. While further increased geographic distance is of little relevance in the knowledge worker network, it remains a strong and defining force for regional connectedness probability in the social network. Granular data on the type of collaboration reveals that collaborating users colocate less if they belong to the same (large) organization. Moreover, sporadic collaboration is less colocated than intensive interactions, suggesting it is harder to establish and maintain in-depth work relationships remotely. Further, inexperienced users tend to collocate more than their experienced peers and users match with similarly experienced peers locally while they typically find more experienced developers remotely.

The contribution of this study is threefold. First, while existing works (e.g., Azoulay et al., 2010; Catalini, 2018; Head et al., 2019) provide consistent evidence that colocation increases collaboration, comprehensive insight into spatial collaboration patterns in a setting with the potential to be fully virtual is lacking. This article presents representative evidence for such a setting and reveals a dichotomy with respect to geography

in the sense that there is a large colocation effect, but apart from that geographic distance is not an important driver of collaboration. Second, I show the colocation effect in a prototypical setting of digital knowledge work is much smaller compared to arguably less digital environments. This provides empirical evidence in line with the 'death of distance' hypothesis that counters the otherwise strong agglomeration effects leading to geographic clustering (e.g., Jaffe et al., 1993; Keller and Yeaple, 2013; Moretti, 2021). Third, previous studies focus on challenges for organizations in managing remote teams (e.g., Gray et al., 2015; Bloom et al., 2022; Yang et al., 2022) while works that compare collaboration within organizations to collaboration between or outside firms is scarce (Duede et al., 2024; Giroud et al., 2022). My findings emphasize the role of large organizations in facilitating remote collaboration as opposed to collaboration outside or between organizations. Large organizations, especially big tech firms, are systematically associated with much smaller colocation effects. At the same time, data suggests that there is still substantial cost associated with remote collaboration as it tends to be less intense than colocated interaction. In line with Emanuel et al. (2023), results point to colocation being especially valuable for inexperienced workers.

The remainder of this paper is organized as follows. In Section 2, I present the data and Section 3 outlines the empirical approach. Section 4 reports the results and Section 5 concludes with a brief discussion.

2 Data

In the last two decades, the adoption of new digital tools for collaborative software development drastically improved workflow and organization of software development projects and enabled developers to work together both on-site and remotely in teams via cloud-based online code repositories. These repositories are maintained using the integrated version control software `git`. Version control with `git` can be highly customized in combination with local code repository copies and is 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 (GitHub, 2021). Due to the nature of the version control system `git`, a detailed history of code changes and contributing users is available online for public repositories. I tap this information as novel data source to measure spatial collaboration patterns of software developers.

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 user profiles and repositories as well as a detailed activity stream capturing all contributions to and events in public repositories. I rely on ten *GHTorrent* snapshots dated between 09/2015 and 03/2021, i.e., roughly one snapshot every seven months.[4] Overall, the data

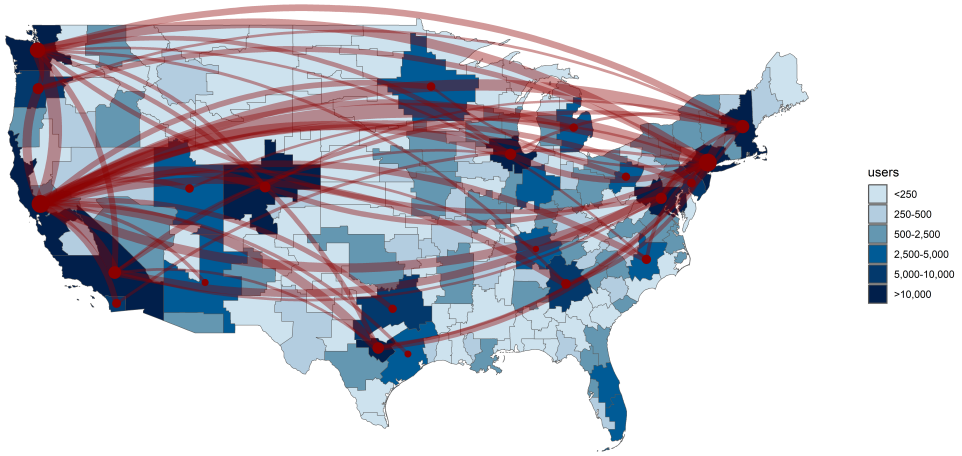
[3] Data from the *GHTorrent* project is publicly available at `ghtorrent.org`.

[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.

contains 44.1 million users worldwide. For my spatial analysis of software developer collaboration in the United States, I select the sample of *GitHub* users according to three criteria: (1) the user reports a location that refers to a city-level location within the United States; (2) the user is active in the observation period, i.e., contributes at least once in two time intervals between data snapshots[5]; and (3) the user collaborates, i.e., contributes to at least one project with another in-sample user. This yields a sample of 190,637 active, collaborating users geolocated in the United States during the observation period from 2015 to 2021, who contribute to about 4.3 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*, and form 10.1 million links among each other.

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 profile. Locations are georeferenced via exact string matching to U.S. cities in the *World Cities Database* and then assigned to respective economic areas via their latitude and longitude and *Bureau of Transportation Statistics*'s economic-area shapes. I choose this regional level such that it is both sufficiently detailed to study colocation and distance effects, provides an adequate level of aggregation given the number of users in each economic area, and respects the precision of users' location input. 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.

Figure 1: Geographic user distribution and collaboration network



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

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

Figure 1 maps the spatial distribution of users and their collaborations. Darker blues represent clustering of a high number of users, with the ten largest economic areas accounting for 79.8% of users. This compares to 68.9% for inventors of computer science patents (Moretti, 2021) and 32.2% for economic area population. Red edges represent inter-regional links with above 20,000 collaborations. The strongest inter-regional links are formed between the largest economic areas, with the Bay Area as the central hub. As a result of the location of the central nodes, many important inter-regional links span long distances between the opposite coasts. A notable property of collaborations is the extent to which they are local. Although the average economic area contains only 0.6% of users, 4.7% of all links of economic-area users are local, i.e., between users that are both located within the economic area. This implies collaborations are, compared to random link formation, on average disproportionally local by a factor of 7.8.

For comparison, I tap two additional data sources. First, I use patent filings from *Patstat* between 2015 and 2021 and source inventor locations from Seliger et al. (2019) and extract inventors of collaborative patents located in the U.S. With this information, I define inventor collaborations similar to the definition of software developer collaboration, i.e., as having filed at least one joint patent. To get a sample that is as similar as possible to software developers, I select inventors of computer science patents.[6] I arrive at a sample of around 17,000 U.S. inventors that filed a collaborative computer-science patent in the observation period.

As a second benchmark, I use regional connectedness in the social network 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* are an adequate representation of real-world social networks. Bailey et al. (2018) construct a regional index of social connectedness for the United States. The so-called *Social Connectedness Index* (SCI) measures the relative probability of connection between users in two regions i and j by

$$\text{index}_{i,j} = \frac{\text{links}_{i,j}}{\text{users}_i * \text{users}_j}, \tag{1}$$

scaled to numbers between 1 and 1,000,000,000. I similarly compute a scaled index using the *GHTorrent* data sample, which I call *GH Connectedness Index* (GHCI).[7] Importantly, the index is independent of region size by construction.

3 Empirical approach

To assess the relation between collaboration an geographic distance, differences in collaboration potential have to be accounted for. In particular, regional collaboration patterns are likely driven by collaboration

[6] More information on data preparation is provided in the Appendix.

[7] For details on index construction, and aggregation see the Appendix. Figure A.7 shows histograms of scaled GHCI and SCI.

potential, i.e., the number of users in the origin and destination region. Therefore, I apply residualized binscatter regression analysis as a non-parametric estimation procedure (Stepner, 2013) that partials out covariates using the Frisch-Waugh-Lovell theorem (Frisch and Waugh, 1933). The conditional expectation function (CEF) is

$$\mathbb{E}[\text{links}_{i,j} \,|\, \mathbf{X}_i, \mathbf{X}_j, \mathbf{X}_{i,j}], \tag{2}$$

where $\text{links}_{i,j}$ denotes the median number of collaborations between regions i and j including $i = j$ for colocated links. To account for collaboration potential, I condition on a vector of cluster size controls $\mathbf{X}_{i,j}$, specifically, the number of origin and destination users, their squared terms (to allow for nonlinear effects), and their logarithmic multiplication to capture bilateral collaboration potential. The binscatter representation of the CEF mapping residualized collaboration against the geodesic distance between origin and destination centroids displays a consistent non-parametric estimate of the relationship between collaboration and geographic distance. To capture local behavior adequately while retaining straightforward interpretation, I choose the number of bins $J = 100$, i.e., each bin representing one percentile of observations.

To quantify the relationship between colocation, distance, and collaboration in a more principled way, I follow the vast literature originating from Tinbergen (1962) and estimate a parsimonious gravity model of the form

$$\ln(\text{links}_{i,j}) = \beta_0 + \beta_1 \mathbb{1}\{\text{coloc}_{i,j}\} + \beta_2 \text{dist}_{i,j} + \mathbf{X}_i\boldsymbol{\beta_3} + \mathbf{X}_j\boldsymbol{\beta_4} + \mathbf{X}_{i,j}\boldsymbol{\beta_5} + \varepsilon_{i,j} \tag{3}$$

where logarithmic collaborations $\ln(\text{links}_{i,j})$ are explained by a colocation indicator marking collaboration between users in the same economic area, $\mathbb{1}\{\text{coloc}_{i,j}\}$, a distance term $\text{dist}_{i,j}$, and origin and destination economic-area characteristics.[8] 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. To control for collaboration potential, I add the multiplication of origin and destination users $\mathbf{X}_{i,j}$. The coefficient β_1 captures the colocation effect, i.e., how much higher local collaboration is relative to non-local collaboration, conditional on covariates. Likewise, the semi-elasticity with respect to distance, β_2, informs how collaboration relates to an increase in geographic distance, accounting for the colocation effect and covariates. The error term is denoted by $\varepsilon_{i,j}$ and I use heteroskedasticity-robust standard errors.

I am interested in differences in the spatial collaboration pattern between a digital work setting, i.e., software development, and arguably less digital settings. Therefore, I compare spatial collaboration patterns among software developers to the (computer science) inventor collaboration network and the social network. Both benchmark networks are less digital than software development because they are more intensive in face-to-face interaction, but arguably to very different degrees. Although there are other differences than their

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

degree of face-to-face intensity as well, these comparisons can offer suggestive evidence on the impact of digital work settings and provide more context to the observed colocation effect in the software developer network.

Computer science inventors are a natural comparison group for software developers for multiple reasons. First, both groups are comprised of high-skilled individuals. Second, both perform similar work in the same field that is mostly characterized by non-routine cognitive tasks. Third, both typically work in an office setting with high computer use intensity. Still, the work of inventors is more creative, innovative, and novel and therefore more face-to-face intensive, i.e., cannot be done virtually to the same extent (see, e.g., Atkin et al., 2022; Yang et al., 2022; Brucks and Levav, 2022; Gibbs et al., 2023). Furthermore, all developers on GitHub by definition use digital tools while this is unlikely true for inventor teams. Hence, I put the effect size observed for software developers in context by comparing the regional collaboration pattern in the software developer network to the pattern in the inventor network by using the same methods as for software developers.

Compared to both the inventor and the software developer network, social relationships are arguably even more demanding in terms of physical proximity even though digital tools such as online social networks greatly facilitate (remote) communication. In that sense, they are the least digital setting among the three networks studied. A comparison of spatial collaboration patterns in software developer and inventor networks to social networks informs on general differences between professional digital collaboration and face-to-face intensive social interaction. For comparing the developer to the social network, I employ a slightly different approach since social connectedness is only available as connectedness index. For the purpose of flexibly estimating the relationship between the indices and distance, I follow Royston and Altman (1994) and fit regressions with fractional polynomials x allowing for the standard set of (repeatable) powers p_i suggested in Royston and Sauerbrei (2008) by

$$x^{(p_1,p_2,\ldots,p_m)}\beta = \beta_0 + \beta_1 x^{(p_1)} + \beta_2 x^{(p_2)} + \ldots + \beta_m x^{(p_m)} \tag{4}$$

where $x^{(0)} = \ln x$ and each repeated power multiplies with another $\ln x$. I then estimate the colocation effect for both the GHCI and SCI as the relation of the predicted values at a distance of zero to the smallest non-zero distance of the respective connectedness index $\hat{\text{CI}}$, i.e.,

$$\frac{\hat{\text{CI}}(\text{dist} = 0)}{\hat{\text{CI}}(\min\{\text{dist} \,|\, \text{dist} \neq 0\})}. \tag{5}$$

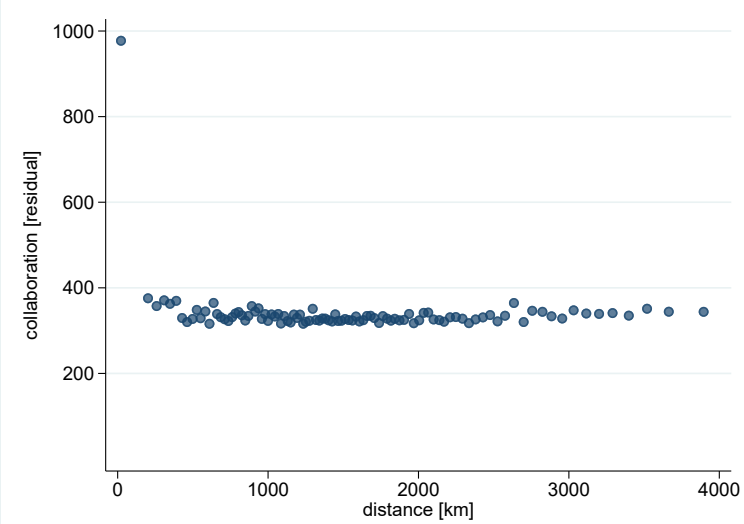
Note that this approximation is conservative in the presence of differences between GHCI and SCI in further spatial decay with geographic distance beyond $\min\{\text{dist} \,|\, \text{dist} \neq 0\}$ due to the smoothing in fractional polynomial estimation.

4 Results

4.1 Main results

Figure 2 plots the binscatter representation of the residualized relationship between collaboration and geographic distance. The first distance percentile, which essentially captures colocation, is clearly elevated.[9] Apart from this colocation effect, the conditional expectation function is flat over the whole distance range. 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 cluster size controls. This suggests that, for region pairs with similar cluster size, being colocated is associated with almost three times more collaborations at the median.

Figure 2: Collaboration and distance



Notes: Figure depicts a residualized binned scatter plot of the conditional expectation function in Equation 2. Means are added back to residuals before plotting. Within-economic area collaborations as well as Honolulu, HI, and Anchorage, AK, economic areas are excluded. *Sources:* GHTorrent, own calculations.

Gravity regression results in Table 1 based on Equation 3 confirm and quantify this pattern more formally. Estimates of the colocation effect are remarkably stable across all specifications. The effect size for colocation is large and statistically highly significant, suggesting colocated users collaborate on average about 8.8 to 9.7 times as much as users that are not colocated, holding economic-area characteristics constant. Further, there is only a very weak, statistically significant negative relation with distance. Depending on the specification and given equal economic-area characteristics, results suggest 0.1% to 0.6% fewer collaborations

[9] The mean centroid-based distance between economic-area centroids in the first distance percentile is 28.6km.

when distance increases by 100km. The fixed-effects model controlling for the multiplication of origin and destination users in column (6) is my preferred specification. The large colocation effect points to direct collaboration with other locals as an important driver of spillover effects among software developers.

Table 1: Collaboration, colocation, and distance

Collaboration [log]	(1)	(2)	(3)	(4)	(5)	(6)
Colocation	2.825***	2.354***	2.298***	2.371***	2.286***	2.329***
	(0.223)	(0.176)	(0.177)	(0.171)	(0.153)	(0.071)
Distance	0.024***	-0.006***	-0.006***	-0.001	-0.006***	-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	0.409	0.409	0.469	0.595	0.922
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	15.87	9.53	8.96	9.71	8.83	9.26

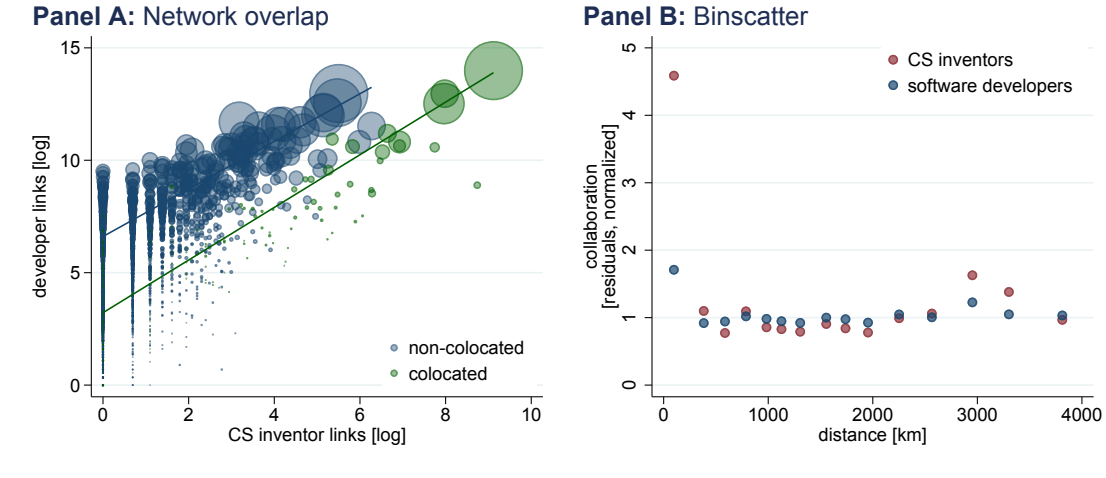
Notes: The outcome variable is the natural logarithm of collaborations between two economic areas plus one. Colocation indicates collaboration between users in the same economic area. Distance is scaled in 100km. Users, GDPs, and Populations refer to the respective variables for both origin and destination. 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.

Note that results show that agglomeration – represented by economic-area characteristics, most importantly cluster size – play a major role for collaboration. The naïve model in column (1) of Table 1 without controls illustrates this: In line with the descriptive finding that a large part of collaborations happens within and between large hubs, this specification overestimates both the role of colocation and distance, even suggests a positive relation between distance and collaboration, and generally is not able to explain variation in collaboration well. Once control variables for economic-area characteristics are subsequently added, the results are robust and stable, while model fit increases to an adjusted R^2 of around 40% with user controls and 60% with GDP and population controls. Adding origin and destination fixed effects that capture also unobserved economic-area characteristics and non-linearity further improves model fit to 92%.

Inventor networks I examine the size of the colocation effect in software developer collaboration via comparison to arguably less digital settings. Panel A of Figure 3 plots the relation between software developer and computer-science inventor networks and differentiates between (blue) and within (green) economic-area collaborations. Marker size represents a measure of economic-area size. There is a strong linear relationship between the two networks. This high inter-regional network overlap implies that software developers and

inventors exhibit a similar inter-regional collaboration pattern.[10] This indicates computer science inventors indeed are a viable comparison group for software developers.

Figure 3: Colocation effect relative to inventors



Note: Panel A shows the relationship between the number of collaborations between economic areas in the software developer and computer-science inventor network. Marker size represents the logarithm of the multiplication of cluster size. The blue and green line are best linear fits from weighted log-log regressions. Panel B shows residualized binned scatter plots of the median number of collaborations and geographic distance between economic-area pairs for both computer-science inventors (red) and software developers (blue), with the number of bins $J = 15$. Residuals are normalized to the mean of bin values, excluding the first distance bin. Means are added back to residuals before plotting. Unconnected economic areas as well as collaborations with Honolulu, HI, and Anchorage, AK, economic areas are excluded. *Sources:* GHTorrent, PatStat, own calculations.

Importantly, within-economic area (i.e., colocated) collaborations, marked in green, are systematically shifted to the right. Size-weighted linear regression lines for within (green) and between (blue) economic area observations formally confirm this. This parallel shift implies that, while exhibiting a comparable pattern otherwise, inventor collaborations are systematically more colocated than collaborations in the software developer network. To quantify the difference in colocation effect size between the two networks, Panel B of Figure 3 shows the relationship between collaboration and geographic distance in a binned scatter plot for both software developers (blue) and computer-science inventors (red) after controlling for economic-area characteristics. Residual values are normalized by the mean values of all distance bins but the first (which represents colocation). There is a clearly visible colocation effect in both networks while increased distance is essentially irrelevant thereafter. The colocation effect is much higher in the inventor network, shown by the larger elevation in median collaboration in the first distance bin for inventors compared software developers. This comparison suggests the colocation effect is about 2.7 times larger in the computer-science inventor network relative to the software developer network.

[10] Figure A.6 shows a similar plot for all inventors, a larger sample of around 76,000 individuals.

Table A.6 reports results of gravity regression analyses and compares variations of the baseline model for the software developer to the inventor network. Model (2) is the preferred (fixed-effects) specification from Table 1, defining colocation as indicator of being in the same economic area. I run specifications for inventors and software developers both on the full sample of observations and for connected economic-area pairs only. The relative effect size is the ratio between estimated colocation effects from the same specification for inventors relative to software developers. Results confirm the binscatter representation, also pointing to a two to three times larger colocation effect for inventors, who are about 26 to 28 times more likely to collaborate locally.

Table 2: Colocation effect for developers and inventors

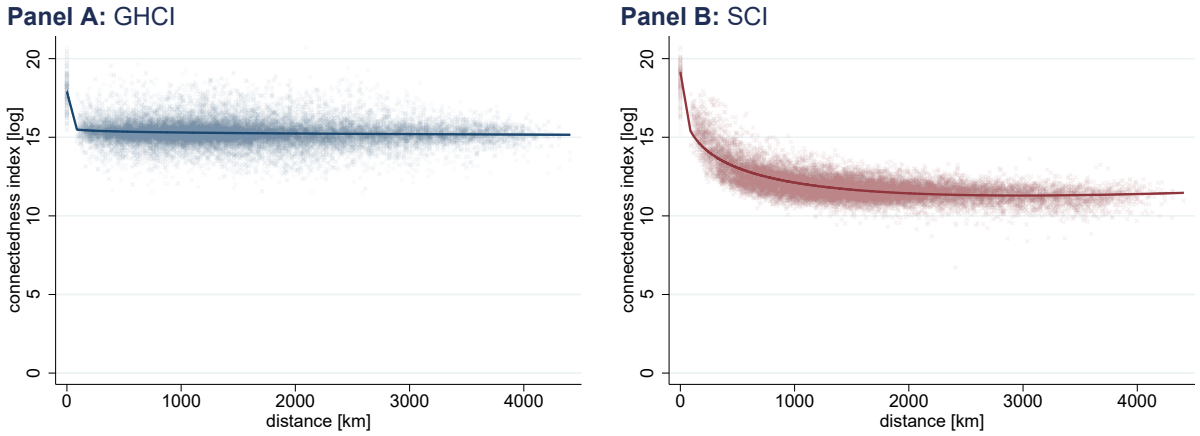
Collaboration	all		connected	
	(1) inventors	(2) developers	(3) inventors	(4) developers
Colocation	3.373***	2.329***	3.292***	2.478***
	(0.138)	(0.071)	(0.102)	(0.081)
Distance	-0.009***	-0.004***	-0.018***	-0.001***
	(0.001)	(0.001)	(0.001)	(0.001)
Users, multiplied	×	×	×	×
Origin FE	×	×	×	×
Destination FE	×	×	×	×
Observations	31,329	31,329	6,662	6,662
Adj. R^2	0.566	0.922	0.593	0.975
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	28.18	9.26	25.90	10.91
Relative effect size	3.04		2.37	

Notes: Distance is scaled in 100km. 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, PatStat, Bureau of Economic Analysis, own calculations.

Intuitively, a larger colocation effect for inventors of computer science patents compared to software developers is explained by three main differences between the two groups. First, inventors' work results in a patent (filing) and therefore always claims novelty and, as a result, requires more creativity and innovation in collaboration processes (Akcigit et al., 2018). And while software development is often a creative and innovative process, as well, this is not always necessary to the degree required for a patent grant. Second, software consists of program code and thus software development tends to be, by nature, more codified than inventing, which increases transferability. Third, while we know by definition developer teams on *GitHub* use digtal tools for collaboration, this is not neccessarily true for inventor teams. All these factors make inventing an activity that is more intensive in face-to-face interaction and thus less susceptible to remote collaboration in an entirely digital work setting.

Social networks As a second benchmark, I investigate the social network. Figure 4 plots predictions of the fractional polynomial regressions from Equation 4 and the underlying index values for the GHCI (left) and SCI (right panel). In both networks, a large colocation effect is clearly visible in the raw data, represented by the sharp upward shift of the (logarithmic) distribution at a distance of zero. Apart from the colocation effect, developer connectedness is essentially independent of distance, in line with the previous findings. In contrast, social connectedness features strong and decreasing spatial clustering as depicted by the continued decrease over the whole distance range. Fractional polynomial regression predictions show the colocation effect as discontinuity at a distance of zero. Comparing predicted index values at a distance of zero to the smallest non-zero distance as in Equation 5 yields a 11.2-fold increase in relative connectedness probability for developer connectedness. This is larger but comparable to the colocation effect estimated in the gravity model, which includes more controls. For the social connectedness, the colocation effect is 41.4 and thus 3.7 times larger than for developer connectedness. Given the continued strong spacial decay in social connectedness and not for developer connectedness, this represents a conservative estimate.

Figure 4: Relative collaboration probability and distance



Note: Panels show fractional polynomial predictions (lines) and values (markers) of scaled GHCI (blue) and SCI (red) between connected economic-area pairs. Scaled SCI from Bailey et al. (2018) is mean-aggregated from county-county level weighted by multiplied populations of each county-pair and rescaled between 1 and 1,000,000,000. *Sources:* GHTorrent, Bailey et al. (2018), U.S. Census Bureau, own calculations.

Hence, compared to the professional networks of (digital) knowledge work by developers or inventors, social connectedness is much more strongly related to geography. Appropriate digital tools are the precondition for remote collaboration and, as a result, enable the difference in observed spatial collaboration patterns between the social and professional networks. In particular, not only is the colocation effect in the social network larger, there is also a strong and continued spatial decay in connectedness for social networks that is not present in knowledge worker networks. Overall, the comparisons to the inventor and social network show that even though the colocation effect in knowledge work is large, it is significantly smaller than in less digital networks.

4.2 Heterogeneity

Collaboration is potentially colocated to a different extent depending on the type of user and/or project. I use the rich data on user activity and affiliation to separately estimate the colocation effect from Equation 3 by organizational affiliation, quality, user and project types, as well as collaboration intensity. Table 3 reports the estimated colocation effects along those dimensions, comparing networks for below- and above-threshold collaborations.

Organizations Large organizations might facilitate remote collaboration (Giroud et al., 2022). I draw on user-indicated affiliation in the data (Panel A).[11] The colocation effect for users with affiliation is 5.67, meaning that users with affiliation are 39% less colocated compared to the full sample. In a naive comparison of the colocation effect into intra- and inter-organizational collaboration, links within organizations are 41% more colocated. However, many firms are small and thus have little scope to facilitate remote collaboration.[12] Therefore, the appropriate comparison is inter- and intra-organizational links of users affiliated with large firms, defined as having more than 200 affiliated users. For large firms, the colocation effect is generally significant but smaller. Specifically, the colocation effect for within-large firms collaborations is 0.59 and 0.78 for between-firm collaborations where at least one user is affiliated with a large firm. This implies a 15% smaller colocation effect for intra-organizational collaboration in this group. Similarly, looking at only users affiliated with one of the big tech firms (Amazon, Google, Apple, Microsoft, or Facebook) yields within-firm collaborations 35% less colocated compared to between-firm links with involvement of a big tech firm user. Interestingly, not all multi-establishment firms seem to facilitate remote collaboration. Defining multi-establishment organizations as firms with users in more than five different economic areas yields no differences in the estimated colocation effect. Overall, these findings provide direct evidence that in particular the largest organizations facilitate remote collaboration.

Quality Colocated and non-colocated collaboration potentially systematically differs in quality. On *GitHub*, there are multiple quality indicators. First, users can be *followed* by other users so that they receive updates on their latest work on the platform. The results shown in Panel B suggest the colocation effect is 28% smaller for high-quality links with above-median followers. A second measure of quality on *GitHub* are *forks*. Users can fork projects on the platform, i.e., copy the current version into another repository. This is typically done when the original project is useful in other projects and, therefore, indicates user interest and usefulness. Alos with forks as quality measure, high-quality collaborations are less colocated, specifically by 19%. As a third quality measure on the platform, I use *stars*. Users can award stars to repositories on *GitHub* to bookmark them for future reference. Hence, stars on a project are an indication of interest in the project. Collaborations in starred projects feature a significantly smaller colocation effect and with a 59%

[11] Around 30% of users provide their affiliation to an organization.

[12] The organization size distribution is plotted in Figure **??** in the Appendix.

Table 3: Colocation effect heterogeneity

Dimension	colocation effect	relative effect	relative to baseline
Panel A: Organizations			
intra-organization	5.26	1.41	0.57
inter-organization	3.73		0.40
within big-tech firm	0.13	0.65	0.01
big-tech firm involved	0.20		0.02
within multi-establishment firm	3.48	0.99	0.38
multi-establishment firm involved	3.51		0.38
within large firm	0.59	0.76	0.06
large firm involved	0.78		0.08
Panel B: Quality			
above-median followers	6.64	0.72	0.72
below-median followers	9.16		0.99
above-median forks	8.97	0.81	0.97
below-median forks	11.07		1.20
with stars	6.49	0.41	0.70
no stars	15.80		1.71
Panel C: User type			
above-median user experience	6.00	0.62	0.65
below-median user experience	9.75		1.05
above-median experience differential	4.36	0.39	0.47
below-median experience differential	11.08		1.20
common programming language	8.02	0.99	0.87
no common programming language	8.13		0.88
Panel D: Collaboration intensity			
strong tie, via project	11.23	1.57	1.21
weak tie, via project	7.16		0.77
above-median project commits	13.00	4.36	1.40
below-median project commits	2.98		0.32
strong tie, via commits	13.05	2.54	1.41
weak tie, via commits	5.12		0.55
Panel E: Project type			
above-median users	6.13	0.33	0.66
below-median users	18.47		1.99
above-median commits	8.64	0.69	0.93
below-median commits	12.47		1.35
above-median project age	6.38	0.38	0.69
below-median project age	16.99		1.83

Notes: Table shows coefficient estimates of the colocation effect in Equation 3 for above- and below-threshold collaboration networks with respect to different characteristics. The relative effect indicates the ratio between the colocation effect in above- and below-threshold networks. The relative-to-baseline effect is the relation to the colocation effect from the preferred model of 9.26. More detailed information on each model is provided in separate tables in the Appendix. *Sources:* GHTorrent, Bureau of Economic Analysis, own calculations.

smaller colocation effect, this effect is even larger using this measure. Since most projects do not receive any stars this measure is a relatively strong sign of quality.

User type Another dimension along which the colocation effect might differ is user characteristics (Panel C). Results show that the colocation effect for experienced users, i.e., users with above-median tenure on the platform, is 38% smaller. This is in line with learning effects for remote collaboration or higher face-to-face requirements for inexperienced developers. Interestingly, collaboration between experienced and inexperienced users is 61% more distributed than collaboration between equally experienced users, maybe because inexperienced users are more willing to incur remote collaboration costs for learning opportunities (Akcigit et al., 2018). Lastly, there is no significant difference in the colocation effect among users with the same main programming language and users with different main programming language.

Collaboration intensity Panel D reports differences in the colocation effect for different measures of strong and weak ties. Strong ties feature a 57% larger colocation effect compared to weak ties, defined as links between users collaborating on only one joint project. Likewise, collaborations in projects with above-median number of commits compared to the average number of commits in joint projects colocate 4.4 times more than collaborations in projects with below-median commits. Defining a weak tie as user pairs where at least one user commits less than two times in all joint projects yields similar results, with 2.5 times higher colocation effect for strong ties. These results suggest that local collaborations typically are much more intense than non-colocated collaborations. Remote collaboration is more sporadic, pointing towards occasional contributions to other (open-source) projects than to core project team membership.

Project type I assess heterogeneity by project type by estimating the colocation effect in networks for large and small projects in terms of users, commits, and project duration. Results in Panel E show that the colocation effect in projects with below-median team size is 77% smaller. When measured through commits, the colocation effect for below-median size teams is 31% smaller. Similarly, longer-running projects exhibit a 72% smaller colocation effect compared to projects with above-median project age. These results suggest that large and long-running projects are more spatially distributed while small and shorter-running projects are more likely to be colocated.

5 Conclusion

I document spatial collaboration patterns of software developers in the United States to study the relevance of geographic distance in a digital work setting. Conditional on economic area characteristics, colocated users collaborate about nine times as much as non-colocated users. However, apart from the colocation effect I find strong evidence of further increased distance being only of limited relevance for software developer collaboration. Importantly, the size of the colocation effect is relatively small compared to less digital

networks; both social networks and computer science inventor networks exhibit colocation effects more than twice as large. The colocation effect is particularly small within large organizations, for high-quality projects, sporadic interactions and experienced users. These findings suggest the relevance of geographic distance for collaboration is indeed subdued in digital knowledge work, which counteracts otherwise strong agglomeration effects.

The broad scope and descriptive nature characterizing the contribution of this analysis have limitations. Although controlling for a multitude of observed and unobserved factors, it ultimately remains unclear to what extent the colocation effect is causally reduced by digitization. Further, the cross-sectional analysis implies a partial equilibrium framework as it takes the observed spatial distribution of developers as given. While unraveling ample suggestive evidence on the mechanism and drivers of the colocation effect, no causal claims can be made. Additionally, data limitations constrain this analysis. More granular definitions of colocation are infeasible, although heterogeneity analyses with respect to shared affiliation point to colocation effects operating at a finer scale and through face-to-face interaction. More direct measurement of face-to-face interaction and a higher spatial resolution would further enhance our understanding of the drivers behind the colocation effect. In addition, especially as organizations seem to be important, it would be desirable to study activity in private repositories, which are not available to date. Moreover, additional information on user characteristics could help to disentangle individual selection effects from aggregate heterogeneity.

These findings have important implications, notably for the governance and spatial organization of knowledge worker teams in the information technology sector. Importantly, findings suggest that colocation is important for knowledge worker teams, but to a lesser extent compared to less digital settings. However, heterogeneity in colocation prevalence indicates that remote collaboration is feasible to a different degree for certain types of collaboration and in different environments. Results point to a crucial role of large organizations in facilitating remote collaboration, and that high-quality projects are often associated with spatially distributed teams. Conversely, data suggests that colocation is more important for intensive collaboration while non-colocated collaboration is typically sporadic. For inexperienced workers colocation with their team seems to be essential. These findings have wider implications for policy making, in particular that ICT could play a significant role in attenuating the strong agglomeration forces in high-skilled labor markets. Not only management but also innovation policy makers should consider that different types of collaboration require different degrees of colocation.

References

Akcigit, Ufuk, Santiago Caicedo, Ernest Miguelez, Stefanie Stantcheva, and Valerio Sterzi, "Dancing with the Stars: Innovation through Interactions," *NBER Working Paper*, 2018.

Andreessen, Marc, "Why Software Is Eating the World," *Wall Street Journal*, 2011, *20* (2011), C2.

Arkolakis, Costas, Federico Huneeus, and Yuhei Miyauchi, "Spatial Production Networks," *NBER Working Paper*, 2023.

Arrow, Kenneth J, *The Limits of Organization*, WW Norton & Company, 1974.

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

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

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.

_ , 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.

Bloom, Nicholas, Ruobing Han, and James Liang, "How Hybrid Working from Home Works Out," *NBER Working Paper*, 2022.

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.

Carlino, Gerald and William R. Kerr, "Agglomeration and Innovation," *Handbook of Regional and Urban Economics*, 2015, *5*, 349–404.

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.

Chauvin, Jasmina, Prithwiraj Choudhury, and Tommy Pan Fang, "Working Around the Clock: Temporal Distance, Intrafirm Communication, and Time Shifting of the Employee Workday," *Organization Science*, 2024.

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

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.

Duede, Eamon, Misha Teplitskiy, Karim Lakhani, and James Evans, "Being Together in Place as a Catalyst for Scientific Advance," *Research Policy*, 2024, *53* (2), 104911.

Emanuel, Natalia, Emma Harrington, and Amanda Pallais, "The Power of Proximity: Training of Tomorrow or Productivity Today?," *Working Paper*, 2023.

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.

Frisch, Ragnar and Frederick V Waugh, "Partial time regressions as compared with individual trends," *Econometrica: Journal of the Econometric Society*, 1933, pp. 387–401.

Gibbs, Michael, Friederike Mengel, and Christoph Siemroth, "Work from Home and Productivity: Evidence from Personnel and Analytics Data on Information Technology Professionals," *Journal of Political Economy Microeconomics*, 2023, *1* (1), 7–41.

Giroud, Xavier, Simone Lenzu, Quinn Maingi, and Holger Mueller, "Propagation and Amplification of Local Productivity Spillovers," *NBER Working Paper*, 2022.

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.

Gray, John V., Enno Siemsen, and Gurneeta Vasudeva, "Colocation Still Matters: Conformance Quality and the Interdependence of R&D and Manufacturing in the Pharmaceutical Industry," *Management Science*, 2015, *61* (11), 2760–2781.

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.

Harrigan, James, Ariell Reshef, and Farid Toubal, "The March of the Techies: Job Polarization Within and Between Firms," *Research Policy*, 2021, *50* (7), 104008.

_ , _ , and _ , "Techies and Firm Level Productivity," *NBER Working Paper*, 2023.

Head, Keith, Yao Amber Li, and Asier Minondo, "Geography, Ties, and Knowledge Flows: Evidence from Citations in Mathematics," *Review of Economics and Statistics*, 2019, *101* (4), 713–727.

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.

Jedrusik, Anita and Phil Wadsworth, "Patent Protection for Software-implemented Inventions," *WIPO Magazine*, 2017, (1), 7–11.

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.

Keller, Wolfgang and Stephen Ross Yeaple, "The Gravity of Knowledge," *American Economic Review*, 2013, *103* (4), 1414–1444.

Korkmaz, Gizem, J Bayoán Santiago Calderón, Brandon L Kramer, Ledia Guci, and Carol A Robbins, "From GitHub to GDP: A framework for measuring open source software innovation," *Research Policy*, 2024, *53* (3), 104954.

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.

_ and Moises Yi, "Size Matters: The Benefits of Large Labor Markets for Job Seekers," *Working Paper*, 2023.

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

Romer, Paul M., "Increasing Returns and Long-run Growth," *Journal of Political Economy*, 1986, *94* (5), 1002–1037.

Royston, Patrick and Douglas G. Altman, "Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling," *Journal of the Royal Statistical Society Series C: Applied Statistics*, 1994, *43* (3), 429–453.

_ **and Willi Sauerbrei**, *Multivariable Model-Building: A Pragmatic Approach to Regression Anaylsis Based on Fractional Polynomials for Modelling Continuous Variables* 2008.

Seliger, Florian, Jan Kozak, and Gaétan de Rassenfosse, "Geocoding of Worldwide Patent Data," 2019.

Simon, Herbert A., "Rational Decision Making in Business Organizations," *The American Economic Review*, 1979, *69* (4), 493–513.

Startlin, "History of GitHub," 2016.

Stepner, Michael, "BINSCATTER: Stata module to generate binned scatterplots," 2013.

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

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, and Jamie Teevan, "The Effects of Remote Work on Collaboration Among Information Workers," *Nature Human Behaviour*, 2022, *6* (1), 43–54.

A Appendix

A.1 Supplementary information

Representativeness. I validate the plausibility and representativeness of the sample in two ways. First, I compare the observed regional concentration pattern with other regional data. For this, I rely on types of data associated with the regional concentration of knowledge workers and their activity footprint across U.S. economic areas: GDP, inventors, establishments, employees, and employee payroll. Where available, I use these metrics both for professional, scientific, and technical services and for computer science. I find a precise and strong positive association for all benchmarks.[13] Relating *GitHub* users to these measures in simple user-weighted log-log regressions explains 77.5 to 90.1% of regional variation and yields an average slope coefficient of 0.99 ranging from 0.74 to 1.20, all highly significant. Relationships are plotted in Figure A.1. These tight and linear relationships centering around one-to-one are reassuring and mitigate potential concerns regarding regional bias in the sample.

Second, I compare the number of connections between users in the software developer network to connections between inventors of collaborative patents in *PatStat*. Although inventors are presumably more focused on creative, novel, and innovative activities resulting in a patent and only represent a subset of the broader community of software developers active on *GitHub*, one would expect to see at least some overlap of the two networks; the fact that regional concentration of inventors and software developers is highly correlated supports this presumption (see Figure A.1). Figure A.2 shows the correlation between inter-regional collaborations of in-sample users and inventors, with all inventors in Panel A and inventors of computer science patents in Panel B. Similar to the definition of a link in the software developer network, I define inventors as linked if they patented jointly at least once.[14] Naturally, there are much less inventors than developers and thus many economic-area pairs feature zero or few inventor links. Despite the differences, there is a strong positive and statistically significant relationship between inter-regional collaboration in the networks which provides additional reassurance of the samples' representativeness also on the (regional) network level.

Connectedness indices. GHCI and SCI indices are calculated using Equation 1. SCI data on the county-county level is taken from Bailey et al. (2018)[15] and aggregated to economic-area level using the methodology suggested in Bailey et al. (2021):

$$\text{SCI}_{i,j} = \sum_{r_i \in R(i)} \sum_{r_j \in R(j)} \text{PopShare}_{r_i} * \text{PopShare}_{r_j} * \text{SCI}_{r_i,r_j} \tag{6}$$

where SCI_{r_i,r_j} is the SCI between sub-regions i and j, sub-regions within region i are indexed $r_i \in R(i)$, and sub-regional population share in region i is denoted by PopShare_{r_i}. For SCI, I aggregate the county-county

[13] For detailed information on supplementary data used here see the Appendix.
[14] For detailed information on supplementary data used here see the Appendix.
[15] Data is retrieved online via data.humdata.org/dataset/social-connectedness-index.

data to the economic-area pair level by using population shares derived from *U.S. Census Bureau* county-level population data as weights, since *Facebook* user counts are not available. After aggregation I rescale the index. To (re)scale GHCI and SCI indices I apply

$$\mathrm{I} \rightarrow \frac{\mathrm{I} - \min(\mathrm{I})}{\max(\mathrm{I}) - \min(\mathrm{I})} * [\mathrm{S}_{\max} - \mathrm{S}_{\min}] + \mathrm{S}_{\min} \tag{7}$$

where I is the index value and minimum (maximum) scale values are denoted by $\mathrm{S}_{\min}$ and $\mathrm{S}_{\max}$ set at 1 and 1,000,000,000, respectively.

Index aggregation. Here I reproduce the derivation of Equation 6 used to aggregate the index to economic-area level from Bailey et al. (2021):

$$\begin{aligned}
\mathrm{SCI}_{i,j} &= \frac{\mathrm{links}_{i,j}}{\mathrm{pop}_i * \mathrm{pop}_j} \\
&= \frac{\sum_{r_i \in R(i)} \sum_{r_j \in R(j)} \mathrm{links}_{r_i,r_j}}{\sum_{r_i \in R(i)} \mathrm{pop}_{r_i} * \sum_{r_j \in R(j)} \mathrm{pop}_{r_j}} \\
&= \sum_{r_i \in R(i)} \sum_{r_j \in R(j)} \frac{\mathrm{pop}_{r_i}}{\sum_{r_i \in R(i)} \mathrm{pop}_{r_i}} \frac{\mathrm{pop}_{r_j}}{\sum_{r_j \in R(i)} \mathrm{pop}_{r_j}} \frac{\mathrm{links}_{r_i,r_j}}{\mathrm{pop}_{r_i} * \mathrm{pop}_{r_j}} \\
&= \sum_{r_i \in R(i)} \sum_{r_j \in R(j)} \mathrm{PopShare}_{r_i} * \mathrm{PopShare}_{r_j} * \mathrm{SCI}_{r_i,r_j}
\end{aligned} \tag{8}$$

where SCI_{r_i,r_j} is the SCI between sub-regions i and j, links between two sub-regions are denoted by links_{r_i,r_j}, sub-regions within region i are indexed $r_i \in R(i)$, sub-regional population is denoted by pop_{r_i}, and sub-regional population share in region i is denoted by $\mathrm{PopShare}_{r_i}$.

Supplementary data. Analyses of *GHTorrent* data is enriched with supplementary data both on the economic area- (i.e., regional) and the economic area pair- (i.e., network) level. At the economic area-level, I use data from the *Bureau of Economic Analyses*, *U.S. Census Bureau*, Moretti (2021), and *County Business Patterns*. From the *Bureau of Economic Analyses* I aggregate yearly county-level data on GDP in "Professional, Scientific, and Technical Services" (NAICS Rev. 2 code 54, "tech GDP") to the economic-area level using the crosswalk between counties and economic areas from Moretti (2021)[16] and take averages for the years 2014 to 2020.[17] From the *U.S. Census Bureau* I use county-level population estimates and apply the same aggregation procedure.[18] From the online replication package of Moretti (2021), I use the number of computer science inventors in each economic area in 2007. From *County Business Patterns*, I use county-level data on the number of workers and establishments as well as payroll for both the "Professional, Scientific, and Technical Services" (NAICS Rev. 2 code 54, "tech") and the "Computer Systems Design and

[16] Retrieved at https://www.openicpsr.org/openicpsr/project/140581/version/V1/view;jsessionid=2BBE031DF440387A3F4EA8416E38D449.

[17] Retrieved at https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas.

[18] Retrieved at https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html.

Related Services" (NAICS Rev. 2 code 5415, "computer science") industry. Here, as well, I aggregate this data to the economic area-level using the procedure described above.

At the economic area pair-level, besides the *Facebook* SCI data discussed above, I merge data on inventors of patents with an application filed from 2015 until 2021 from *PatStat*. Here I first geolocate inventors using the fifth version of the inventor location file in the "Geocoding of Worldwide Patent Data" by Seliger et al. (2019).[19] Inventor latitude and longitude are assigned to economic areas using the economic area shape file by the *Bureau of Transportation Statistics*.[20] Using the location information, I select inventors of collaborative patents located in the U.S. (i.e., patents with at least two inventors). For analysis, I use data on both all inventors and inventors of computer science patents, defined as either having NACE Rev. 2 codes 62 ("Computer Programming, Consultancy and Related Activities") or 63 ("Information Service Activities"), or IPC code H04 ("Electric Communication Technique"). There are around 76,000 inventors with a location in the U.S. that filed a collaborative patent in this time period, of which about 17,000 filed a computer science patent.

A.2 Robustness

Colocation There is no universal method to conceptualize colocation, but literature suggests that commutable geographic distances are often economically meaningful for economic applications and colocation effects are even stronger at the microgeographic level. Here I opt for economic areas for two reasons. First, they represent commutable economic markets surrounding cities. Second, users often indicate their location as a city's "metropolitan area" or "area", so that there typically is not more precision in their exact location available. However, since economic areas are of different geographic size, a potential concern is that small neighboring economic areas might be commutable and therefore should be included in the definition of colocation. Therefore, I run Model (6) from Table 1 with alternative definitions of colocation. The results are shown in Table A.2. Including centroid-based distances of less than 100km captures only seven economic-area pairs but leads to a substantially smaller colocation effect of 7.73. Allowing distances up to 200km includes 207 economic-area pairs and causes a sharp drop in the estimated colocation effect to 1.38. This confirms that the colocation effect is indeed confined to small geographic distances and decays rapidly after 100km.

Functional form In the main specification, I impose a (linear) functional form assumption on the distance effect. A potential concern here is that the relationship between collaboration and distance exhibits a different, possibly non-linear, pattern. To check for this possibility I increase model flexibility by specifying distance in a non-parametric way, i.e., using indicator variables for different distance bins. Figure A.4 plots the resulting coefficient estimates of these distance bin indicators. The coefficient for distances greater than 3200km is omitted as reference. Also here, the colocation effect clearly stands out, measured by the coef-

[19] Retrieved at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OTTBDX.
[20] Retrieved at https://maps.princeton.edu/catalog/harvard-ntadbea.

ficient on the first indicator for distances equal to zero. The other distance bins are of little importance in comparison. The bin for distances between zero and 100km is estimated less precisely than others and is not significantly different from zero. Except for the last estimate, the coefficient estimates tend to gradually become smaller for higher distances. This shows that the colocation effect is confined to small distances only and essentially vanishes thereafter, confirming findings from Panel B in Figure 2. The results thus provide further support of the colocation definition and, given the generally monotonous behavior with increasing distance, justify a simple parametric distance specification. Other parametric models that allow for non-linear distance effects by adding a squared distance term do not improve model fit or impact the main effect significantly (Table A.3).

Individual-level models Alternative model specifications are individual-level probability models, which I avoid as main specification for two reasons. First, at the individual level, the largest part of a developers' network is unobserved in the data while at the economic-area pair level, the representativeness is given and validated. Second, data becomes extremely large and sparse as the adjacency matrix features less than 0.5% non-zero values, a known characteristic of social networks. Nevertheless, I run several probability models for a specification with non-parametric distance. To be computationally efficient I draw a random sample of about 20,000 users which yields a model with about 5.6% of collaborating users and 33 million observations. All three types of models (Linear Probability, Poisson Pseudo-maximum Likelihood, and Probit) presented in Table A.4 exhibit a similar pattern with respect to distance as the preferred specification (see Figure A.5).

Time zones Omitted variables related to distance and collaboration are potential concerns when assessing effects of geography. In particular, reductions of collaboration could be caused by differences in time zones, i.e., business hour overlap (Chauvin et al., 2024). Repository-based software development generally allows for a high degree of asynchronous collaboration. As a result, time zone differences might be less important. Nevertheless, Table A.5 reports regression results from specificaitons including time zone controls. Reassuringly, the effect size remains virtually unchanged across all specifications. Still, time zones significantly affect collaboration. Results from model one suggests about 8.2% higher collaboration within time zones. Using time zone differences, I estimate a reduction in collaboration by 2.3% for each hour of time difference. These findings are generally in line with Chauvin et al. (2024) who estimate an overall reduction in communication of 9.4%, but find no significant effect on asynchronous communication.

Relatedness. It is important to assess the degree to which the discussed heterogeneity dimensions are interrelated in the network. A high degree of collinearity among variables that are used to tease out heterogeneous effects would lead to inability of the econometric model to distinguish the drivers of heterogeneity in the colocation effect size. I assess the relatedness of link characteristics by computing the bivariate correlation matrix of the metrics used to construct the networks for the above heterogeneity analyses. The matrix is shown as a heat map in Figure A.8. In general, the variables are not correlated to a worrying degree. In fact – apart from obviously related alternative measures for the same underlying concept like stars and forks

for quality or large firm and big tech firm – variables are only very weakly correlated with each other. This mitigates potential concerns regarding collinearity issues in the heterogeneity analyses.

A.3 Tables

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, Bureau of Economic Analysis, own calculations.

Table A.2: Sensitivity to colocation definition

Collaboration [log]	distance cutoff		
	(1) $= 0$ km	(2) < 100 km	(3) < 200 km
Colocation	2.329***	2.166***	0.866***
	(0.071)	(0.079)	(0.050)
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	0.922	0.922	0.919
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	9.26	7.73	1.38

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: Sensitivity to model flexibility

Collaboration	log				IHS			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Colocation	2.219***	2.266***	2.350***	2.204***	2.401***	2.463***	2.527***	2.388***
	(0.072)	(0.079)	(0.071)	(0.076)	(0.081)	(0.086)	(0.081)	(0.085)
Distance	-0.021***	-0.003***	-0.004***	-0.018***	-0.021***	-0.004***	-0.004***	-0.019***
	(0.002)	(0.001)	(0.001)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)
Distance squared	0.000***			0.000***	0.000***			0.000***
	(0.000)			(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	31,329	31,329	31,329	31,329
Adj. R^2	0.923	0.925	0.923	0.928	0.924	0.925	0.924	0.927
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	8.92	9.52	10.39	8.74	10.04	10.74	11.52	9.90

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. Models (1) to (4) feature the natural logarithm of collaborations between two economic areas plus one and Models (5) to (8) show the same specifications with the inverse hyperbolic sine-transformed number of links as outcomes. 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.

Table A.4: Individual-level probability models

Collaboration	(1) LPM	(2) PPML	(3) Probit
$<$ 100 km	0.00139***	0.226***	0.080***
	(0.00006)	(0.010)	(0.003)
100 – 400 km	0.00019***	0.036***	0.013***
	(0.00007)	(0.012)	(0.004)
400 – 1200 km	-0.00005	-0.008	-0.003
	(0.00004)	(0.007)	(0.003)
1200 – 2400 km	-0.00009*	-0.019**	-0.006**
	(0.00005)	(0.009)	(0.003)
2400 – 3200 km	-0.00011**	-0.020**	-0.007**
	(0.00005)	(0.009)	(0.003)
Origin FE	×	×	×
Destination FE	×	×	×
Observations	33,183,717	33,179,297	33,179,297
Users (random sample)	10,726	10,726	10,726
Sample share	0.056	0.056	0.056
(Pseudo) Adj. R^2	0.0003	0.0046	0.0046

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.5: Collaboration and time zones

Collaboration [log]	(1)	(2)	(3)	(4)
Colocation	2.329***	2.306***	2.332***	2.329***
	(0.071)	(0.071)	(0.071)	(0.071)
Distance	-0.004***	-0.001**	-0.002**	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)
Same timezone		0.082***		
		(0.010)		
Timezone difference			-0.023**	
			(0.010)	
Timezone difference (IHS)				-0.068***
				(0.012)
Origin FE	×	×	×	×
Destination FE	×	×	×	×
Observations	31,329	31,329	31,329	31,329
Adj. R^2	0.922	0.923	0.923	0.923
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	9.26	9.03	9.30	9.26

Notes: 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.6: Colocation effect for developers and inventors

Collaboration	log				IHS			
	all		connected		all		connected	
	(1) inventors	(2) developers	(3) inventors	(4) developers	(5) inventors	(6) developers	(7) inventors	(8) developers
Colocation	3.373***	2.329***	3.292***	2.478***	3.821***	2.511***	3.605***	2.571***
	(0.138)	(0.071)	(0.102)	(0.081)	(0.143)	(0.080)	(0.099)	(0.089)
Distance	-0.009***	-0.004***	-0.018***	-0.001**	-0.011***	-0.004***	-0.020***	-0.001***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
Users, multiplied	×	×	×	×	×	×	×	×
Origin FE	×	×	×	×	×	×	×	×
Destination FE	×	×	×	×	×	×	×	×
Observations	31,329	31,329	6,662	6,662	31,329	31,329	6,662	6,662
Adj. R^2	0.566	0.922	0.593	0.975	0.563	0.924	0.585	0.975
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	28.18	9.26	25.90	10.91	44.67	11.32	35.78	12.08
Relative effect size	3.04		2.37		3.95		2.96	

Notes: Table compares variations of the baseline model for the software developer to the inventor network. Model (2) is the preferred (fixed-effects) specification from Table 1, defining colocation as indicator of being in the same economic area. Models (1) to (4) use the logarithmic number of links as outcome, Models (5) to (8) feature the inverse hyperbolic sine-transformed number of links. Within these two groups, specifications are shown for inventors and software developers both on the full sample of observations and for connected economic-area pairs. The relative effect size is the ratio between estimated colocation effects from the same specification for inventors relative to software developers. Distance is scaled in 100km. 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, PatStat, Bureau of Economic Analysis, own calculations.

Table A.7: Colocation and organizations

Collaboration	baseline		link type		organization type					
					big tech		multi-est.		large	
	(1) all	(2) with info	(3) intra-org.	(4) inter-org.	(5) within	(6) involved	(7) within	(8) involved	(9) within	(10) involved
Colocation	2.329***	1.898***	1.834***	1.554***	0.122**	0.184***	1.500***	1.506***	0.463***	0.577***
	(0.071)	(0.090)	(0.126)	(0.082)	(0.054)	(0.065)	(0.125)	(0.090)	(0.092)	(0.084)
Distance	-0.004***	-0.002***	-0.001***	-0.002***	0.000	0.001	-0.001***	-0.002***	-0.000	-0.000
	(0.001)	(0.001)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.001)	(0.000)	(0.001)
Users, multiplied	×	×	×	×	×	×	×	×	×	×
Origin FE	×	×	×	×	×	×	×	×	×	×
Destination FE	×	×	×	×	×	×	×	×	×	×
Observations	31,329	31,329	31,329	31,329	31,329	31,329	31,329	31,329	31,329	31,329
Adj. R^2	0.922	0.764	0.572	0.761	0.573	0.686	0.562	0.759	0.540	0.691
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	9.26	5.67	5.26	3.73	0.13	0.20	3.48	3.51	0.59	0.78
Relative effect size	0.61		0.71		1.53		1.01		1.32	

Notes: Model (1) is the preferred (fixed-effects) specification from Table 1, defining colocation as indicator of being in the same economic area. Model (2) restricts Model (1) to links where both users provide an affiliation. Models (3) and (4) contrast the colocation effect for intra- and inter-organizational links. Model (5) estimates the colocation effect for links within the big tech firms Google, Amazon, Microsoft, Facebook, and Apple. Model (6) estimates the colocation effect for multi-establishment organizations defined as organizations with affiliated users in at least 5 different economic areas, and Model (7) for organizations with at least 200 affiliated users. Distance is scaled in 100km. 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.8: Colocation and collaboration quality

Collaboration		followers		forks		stars	
	(1) baseline	(2) $\geq$ median	(3) $<$ median	(4) $\geq$ median	(5) $<$ median	(6) ≥ 1	(7) $= 0$
Colocation	2.329***	2.033***	2.318***	2.299***	2.491***	2.013***	2.821***
	(0.071)	(0.081)	(0.078)	(0.072)	(0.121)	(0.074)	(0.109)
Distance	-0.004***	-0.003***	-0.003***	-0.003***	-0.003***	-0.003***	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Users, multiplied	×	×	×	×	×	×	×
Origin FE	×	×	×	×	×	×	×
Destination FE	×	×	×	×	×	×	×
Observations	31,329	31,329	31,329	31,329	31,329	31,329	31,329
Adj. R^2	0.922	0.805	0.828	0.855	0.664	0.850	0.741
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	9.26	6.64	9.16	8.97	11.07	6.49	15.80
Relative effect size	–	1.38		1.23		2.43	
Median	–	8		5		0	

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) to (7) estimate Model (1) on the number of links that are below (above) certain threshold values of various collaboration quality metrics. E.g., Model (2) estimates the colocation effect for links where the average number of followers of the two users is above the median number of (average) followers in all users-pairs of 8. Models (4) and (5) refer to links in projects with above- or below-median number of forks. Models (6) and (7) refer to links in projects with and without stars. Distance is scaled in 100km. 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.9: Colocation and project types

Collaboration		users		commits		age	
	(1) baseline	(2) ≥ 3	(3) < 3	(4) $\geq$ median	(5) $<$ median	(6) $\geq$ median	(7) $<$ median
Colocation	2.329***	1.964***	2.969***	2.266***	2.600***	1.999***	2.890***
	(0.071)	(0.080)	(0.120)	(0.074)	(0.116)	(0.072)	(0.116)
Distance	-0.004***	-0.003***	-0.005***	-0.003***	-0.003***	-0.003***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Users, multiplied	×	×	×	×	×	×	×
Origin FE	×	×	×	×	×	×	×
Destination FE	×	×	×	×	×	×	×
Observations	31,329	31,329	31,329	31,329	31,329	31,329	31,329
Adj. R^2	0.922	0.854	0.679	0.853	0.702	0.850	0.717
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	9.26	6.13	18.47	8.64	12.47	6.38	16.99
Relative effect size	–	0.33		0.69		0.38	
Median	–	2		15		11	

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) to (7) estimate Model (1) on the number of links that are below (above) certain threshold values of project metrics. Models (2) and (3) estimate the colocation effect links within projects that feature more than two users and two users, respectively. Models (4) and (5) refer to links within projects that feature above- (below-)median commits and Models (6) an (7) to links within projects of above- (below-)median age in months. Distance is scaled in 100km. 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.10: Colocation and user types

Collaboration		experience		Δ(experience)		programming language	
	(1) baseline	(2) ≥ median	(3) < median	(4) ≥ median	(5) < median	(6) same	(7) different
Colocation	2.329***	1.946***	2.375***	1.679***	2.492***	2.200***	2.212***
	(0.071)	(0.081)	(0.078)	(0.079)	(0.078)	(0.088)	(0.074)
Distance	-0.004***	-0.003***	-0.003***	-0.002***	-0.003***	-0.002***	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Users, multiplied	×	×	×	×	×	×	×
Origin FE	×	×	×	×	×	×	×
Destination FE	×	×	×	×	×	×	×
Observations	31,329	31,329	31,329	31,329	31,329	31,329	31,329
Adj. R^2	0.922	0.793	0.836	0.807	0.836	0.782	0.842
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	9.26	6.00	9.75	4.36	11.08	8.02	8.13
Relative effect size	–	0.62		0.39		0.99	
Median	–	11.5		7		–	

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) to (7) estimate Model (1) on the number of links that are below (above) median of user metrics. Models (2) and (3) refer to links with above- (below-)median project-level user engagement measured by the average number of commits to a project per user-pair. Models (4) and (5) refer to the average platform age of the user-pair as a measure of experience. Models (6) and (7) refer to the differential in experience between both users in a link, also measured as user platform age. Model (8) refers to links where both users feature the same (main) programming language, defined as the programming language most used by a user over all her commits. Distance is scaled in 100km. 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.11: Colocation and economic-area characteristics

Collaboration		# local users		avg. firm size	
	(1) baseline	(2) $\geq$ median	(3) Top 10	(4) $\geq$ median	(5) $\geq$ median
Colocation	2.329***	2.478***	2.430***	2.498***	2.430***
	(0.071)	(0.113)	(0.068)	(0.074)	(0.069)
Distance	-0.004***	-0.004***	-0.004***	-0.004***	-0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Colocation interactions with					
Large economic area		-0.295**			
		(0.142)			
Top 10 largest economic area			-1.978***		
			(0.446)		
Big tech firm intensity				-1.026***	
				(0.183)	
Big software firm intensity					-1.595***
					(0.386)
Observations	31,329	31,329	31,329	31,329	31,329
Adj. R^2	0.922	0.923	0.923	0.923	0.923
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	9.26	10.91	10.36	11.16	10.36
$\exp(\hat{\beta}_{\text{colocation}} + \hat{\beta}_{\text{interaction}}) - 1$	–	7.87	0.57	3.36	1.31
Relative effect size	–	1.39	18.18	3.32	7.91

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)-(5) assess the heterogeneity of the colocation effect by including interactions with local characteristics. Large economic area is an indicator for above-median number of users. Top 10 largest economic area indicates the ten largest economic areas in terms of the number of users. Big tech firm intensity is an indicator for above-median number of technology firms with more than 1,000 employees. Likewise, big software firm intensity indicates above-median number of software firms with more than 1,000 employees. Distance is scaled in 100km. 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, County Business Patterns, own calculations.

Table A.12: Colocation and strong versus weak ties

Collaboration		projects		commits			
				median		minimum	
	(1) baseline	(2) > 1	(3) $= 1$	(4) above	(5) below	(6) > 2	(7) ≤ 2
Colocation	2.329***	2.504***	2.100***	2.639***	1.382***	2.643***	1.812***
	(0.071)	(0.105)	(0.068)	(0.089)	(0.064)	(0.104)	(0.068)
Distance	-0.004***	-0.004***	-0.003***	-0.003***	-0.002***	-0.003***	-0.002***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Users, multiplied	×	×	×	×	×	×	×
Origin FE	×	×	×	×	×	×	×
Destination FE	×	×	×	×	×	×	×
Observations	31,329	31,329	31,329	31,329	31,329	31,329	31,329
Adj. R^2	0.922	0.792	0.920	0.809	0.830	0.758	0.847
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	9.26	11.23	7.16	13.00	2.98	13.05	5.12
Relative effect size	–	1.57		4.36		2.54	

Notes: Model (1) is the preferred (fixed-effects) specification from Table 1, defining colocation as indicator of being in the same economic area. Model (2) features the logarithmic number of strong ties as outcome variable, i.e., the number of inter-regional links between users with multiple joint projects. The outcome variable in Model (3) is the logarithmic number of weak ties, i.e., the number of inter-regional links between users with only one joint project. Models (4) and (5) contrast colocation in links with sporadic and intense collaboration, where sporadic collaboration is indicated by links where at least one user contributes less than two commits in all joint projects. Distance is scaled in 100km. 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.

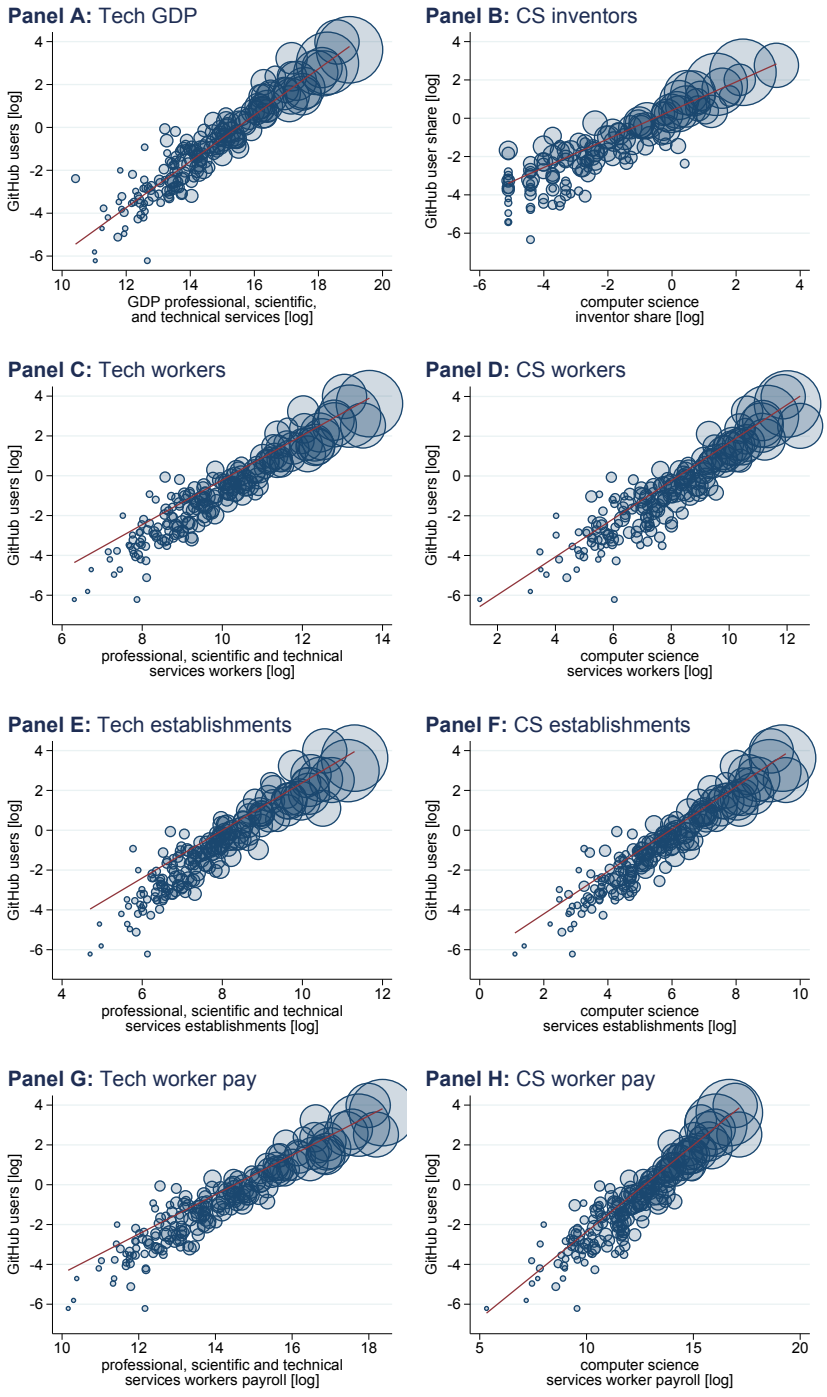
Table A.13: Colocation and collaboration intensity

Collaboration	(1)	counts (2)	counts (3)	ratios (4)	ratios (5)
	baseline	projects	commits	commits per project	commits per link
Colocation	2.329***	3.106***	4.505***	1.254***	2.029***
	(0.071)	(0.099)	(0.156)	(0.082)	(0.109)
Distance	-0.004***	-0.005***	-0.008***	-0.002***	-0.003***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Users, multiplied	×	×	×	×	×
Origin FE	×	×	×	×	×
Destination FE	×	×	×	×	×
Observations	31,329	31,329	31,329	31,329	31,329
Adj. R^2	0.922	0.907	0.852	0.555	0.547
$\exp(\hat{\beta}_{\text{colocation}}) - 1$	9.26	21.32	89.43	6.60	2.51
Relative effect size	–	2.30	9.66	–	–

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) estimate the colocation effect in the sum of projects, Model (2), and commits, Model (3), between economic-area pairs. Models (4) and (5) feature collaboration intensity measures: average number of commits per project, Model (5), and user-link, Model (6), for each economic-area pair. Distance is scaled in 100km. 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.

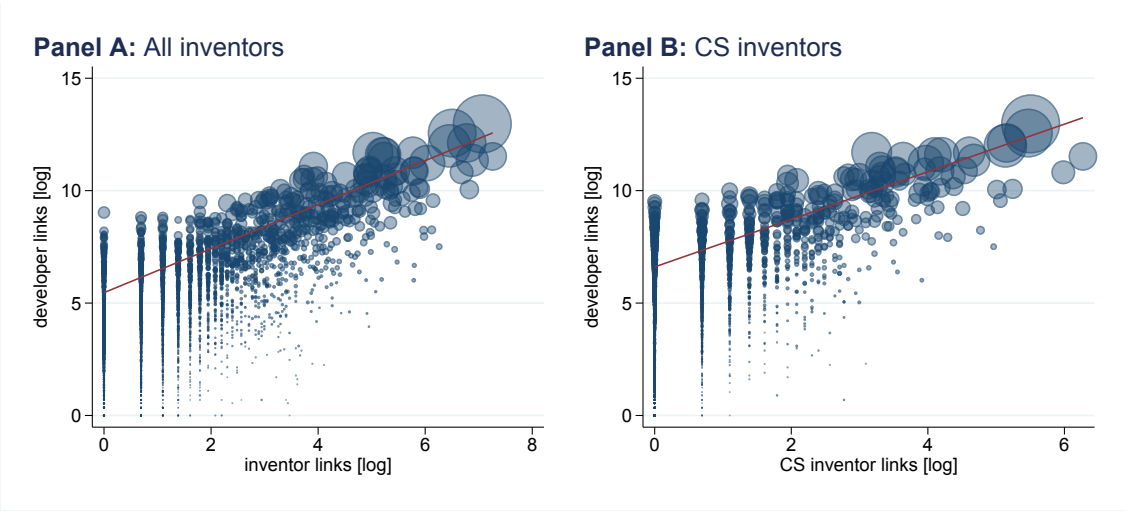
A.4 Figures

Figure A.1: Representativness



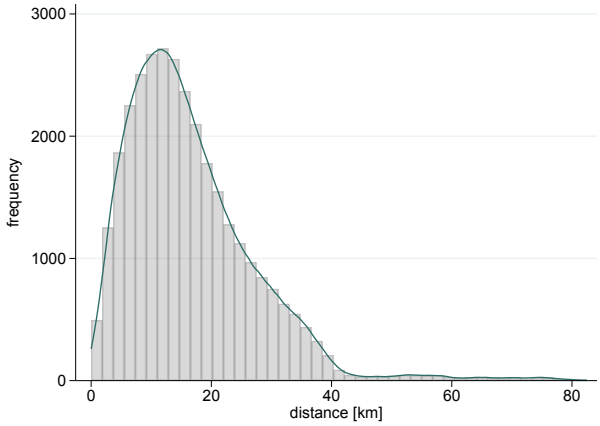
Note: Plots show the relationship between (the share of) users per economic area and economic-area level metrics related to software development after logarithmic transformation. Bubble size represents economic-area population size. Red lines are best linear fits from user-weighted log-log regressions. *Sources:* GHTorrent, Moretti (2021), Bureau of Economic Analysis, County Business Patterns, own calculations.

Figure A.2: Relation between software developer and inventor collaboration network



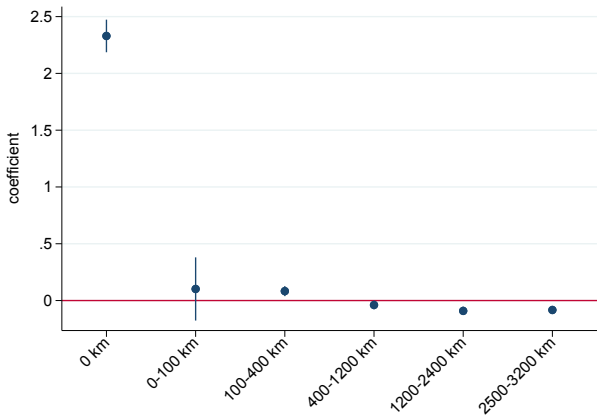
Note: Plots show the relationship between the number of inter-regional collaborations between economic areas in the software developer and inventor network. Panel A compares software developer collaborations to all collaborations in collaborative patents and Panel B to collaborative computer science patents. Collaborations are transformed logarithmically. Bubble size represents the multiplication of economic-area size in terms of users after logarithmic transformation. Red lines are best linear fits from weighted log-log regressions. *Sources:* GHTorrent, PatStat, Bureau of Economic Analysis, own calculations.

Figure A.3: Distance



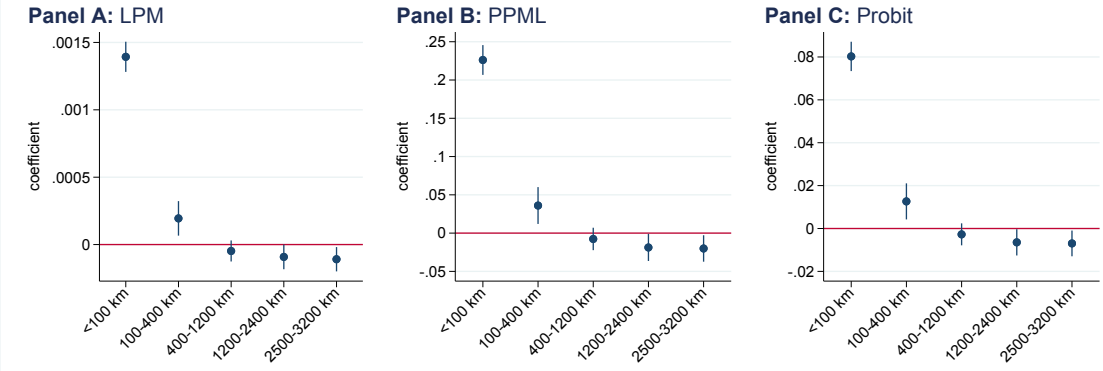
Notes: Plot shows the distribution of centroid-based geodesic distance between economic areas. The horizontal red line indicates the median distance of 1,439. The blue curve represents the Epanechnikov kernel density estimate. The right tail of the distribution starting approximately at distances greater 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.4: Non-parametric distance



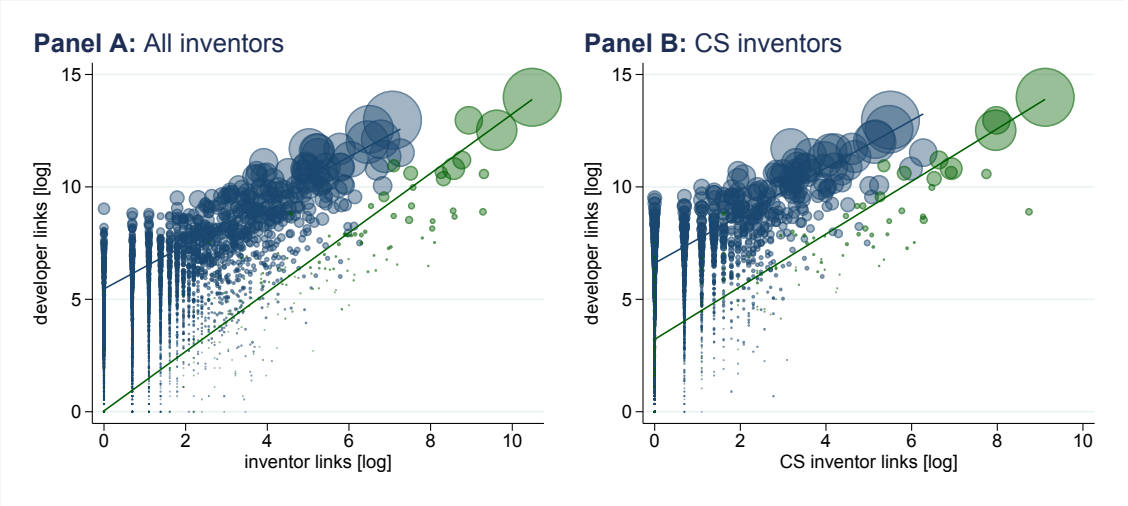
Notes: Plot shows coefficient point estimates and confidence intervals for the baseline fixed effects model specification with non-parametric distance. The indicator for distances above 3,200 km is omitted. Blue bars show 95% confidence intervals from robust standard errors. Collaborations with Anchorage, AK, and Honolulu, HI, are excluded. *Sources:* GHTorrent, own calculations.

Figure A.5: Individual-level probability models



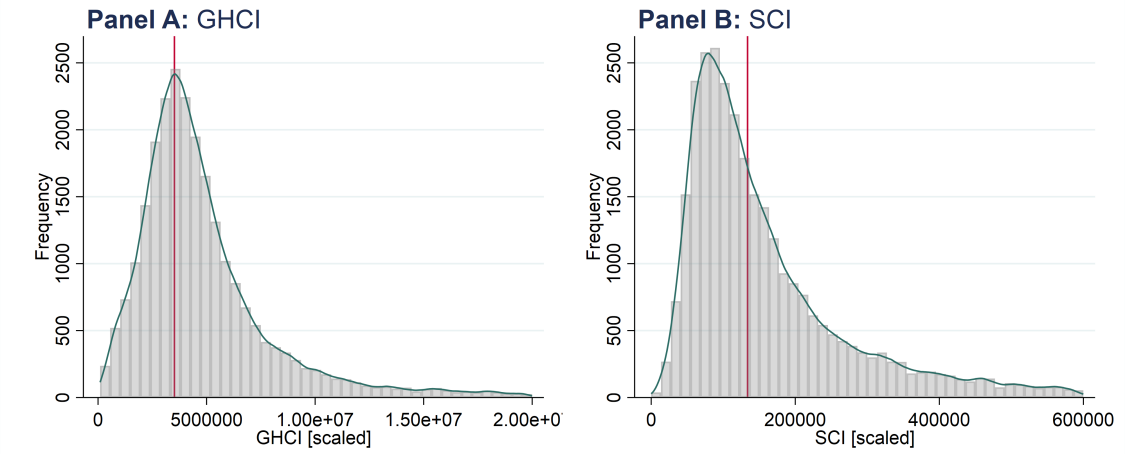
Notes: Plot shows coefficient point estimates and confidence intervals for the individual-level fixed effects model specification with non-parametric distance from Table A.4. The indicator for distances above 3,200 km is omitted. Blue bars show 95% confidence intervals from robust standard errors. Collaborations with Anchorage, AK, and Honolulu, HI, are excluded. *Sources:* GHTorrent, own calculations.

Figure A.6: Colocation effect relative to inventors



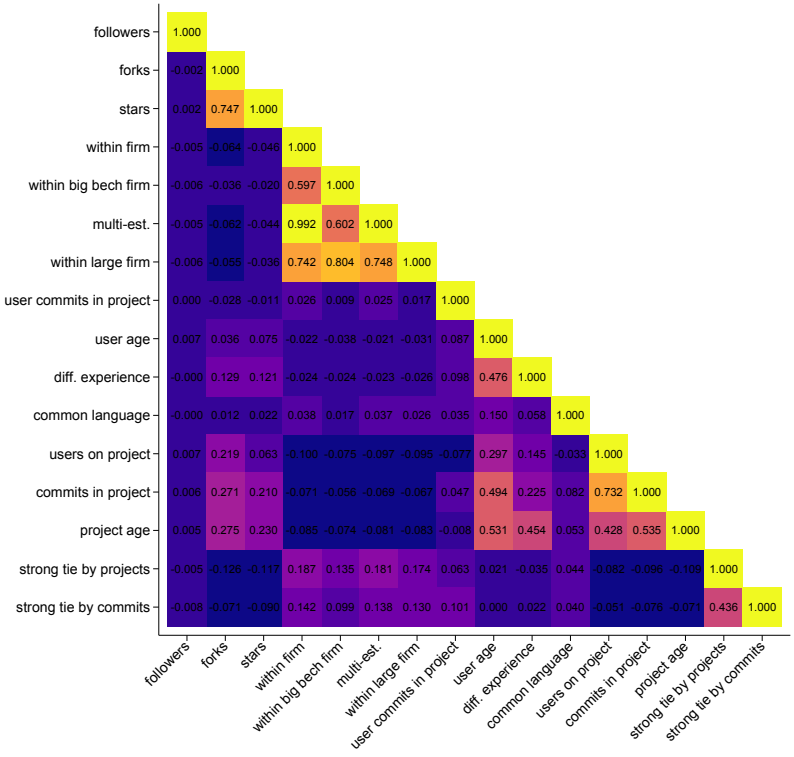
Note: Plots show the relationship between the number of collaborations between economic areas in the software developer and inventor network. Panel A compares software developer collaborations to all collaborations in collaborative patents and Panel B to collaborative computer science patents. Collaborations are transformed logarithmically. Blue bubbles depict between-economic area collaborations and green bubbles represent within-economic area collaborations. Bubble size represents the multiplication of economic-area size in terms of users after logarithmic transformation. The blue and green line are best linear fits from weighted log-log regressions for within- and between-economic area observations. *Sources:* GHTorrent, PatStat, own calculations.

Figure A.7: 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. (2018) 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. (2018), Bureau of Economic Analysis, own calculations.

Figure A.8: Relatedness of link characteristics



Note: Plots shows bivariate correlations between link characteristics for the sample where all characteristics are non-empty. Correlations are colored by their strength. *Sources:* GHTorrent, Bureau of Economic Analysis, own calculations.