RATIONALITY & COMPETITION
CRC TRR 190

Bit by Bit

Moritz Goldbeck (ifo Institute & LMU Munich)

Discussion Paper No. 422

September 05, 2023

Collaborative Research Center Transregio 190 | www.rationality-and-competition.de
Ludwig-Maximilians-Universität München | Humboldt-Universität zu Berlin
Spokesperson: Prof. Georg Weizsäcker, Ph.D., Humboldt University Berlin, 10117 Berlin, Germany
info@rationality-and-competition.de

Bit by Bit

Colocation and the Death of Distance in Software Developer Networks*

Moritz Goldbeck†

September 5, 2023
Latest Version

Abstract

Digital work settings potentially facilitate remote collaboration and thereby decrease geographic frictions 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*. Despite advanced digitization in this occupation, developers are geographically highly concentrated, with 79.8% of users clustering in only ten economic areas, and colocated developers collaborate about nine times as much as non-colocated developers. However, the colocation effect is much smaller than in less digital social or inventor networks, and apart from colocation geographic distance is of little relevance to collaboration. This suggests distance is indeed less important for collaboration in a digital work setting while other strong drivers of geographic concentration remain. Heterogeneity analyses provide insights on which types of collaboration tend to colocate: the colocation effect is smaller within larger organizations, for high-quality projects, among experienced developers, and for sporadic interactions. Overall, this results in a smaller colocation effect in larger economic areas.

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

JEL-Codes: L84, O18, O30, R32

*I thank Lena Abou El-Komboz, Dany Bahar, Raj Chetty, Thomas Fackler, Oliver Falck, Richard Freeman, Ed Glaeser, Shane Greenstein, Ricardo Hausmann, Anna Kerkhof, Bill Kerr, Frank Nagle, Giacomo De Nicola, Megan MacGarvie, 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, 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 Lara Mai, Raunak Mehrotra, Svenja Schwarz and Gustav Pirich for excellent research assistance and gratefully acknowledge public funding through DFG grant number 280092119.

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

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" argument 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; Catalini, 2018; Forman et al., 2016). Scholars proposed various explanations for this, including the importance of face-to-face interaction (Atkin et al., 2022; Battiston et al., 2021), positive industry-cluster spillovers (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 whether "distance is dying."

Knowledge work is expected to be particularly susceptible to the "death of distance" since many tasks are already digitized, as shown by high computer and internet use in related occupations (Alipour et al., 2023). Here, I look at software development as an integral and increasingly important part of the knowledge economy: software is not only a key sector on its own 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] Not only is software development a crucial and often overlooked industry, but it also offers a characteristic setting of knowledge work in general typically being a collaborative effort, which research suggests is increasingly the case in many 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" argument 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 in a highly digitized work setting at the core of the knowledge economy, i.e., software development. Drawing on detailed georeferenced network data from the largest code repository platform, *GitHub*, I analyze regional concentration and collaboration patterns of some 191 thousand U.S. software developers in public projects between 2015 and 2021. I focus on the U.S. here as a large and integrated geography 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 and team collaboration. In a first step, I provide descriptive evidence and fit gravity-type regression models to explain spatial collaboration patterns and distinguish the benefits of being colocated in the same economic area from the general relevance of increased distance. 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 the drivers of the observed spatial collaboration pattern characteristic to the digital setting in the software developer network. To this end, I leverage detailed information on the type of collaboration and individual characteristics and estimate the group-specific impact of geographic factors on collaboration depending on organizational affiliation, user and project characteristics, as well as collaboration intensity and quality.

Results show high spatial concentration with 79.8% of users clustering in only 10 of 179 U.S. economic areas. This is a stronger concentration than for computer science inventors (68.9%) and compares to only 32.2% of the population in the same economic areas. The inter-regional collaboration network exhibits a strong skewness towards large clusters, most notably the Bay Area. Binned scatter plots show collaboration is strongly associated with economic-area characteristics, especially cluster size and bilateral collaboration potential. This points to significant spillover effects in line with recent findings (Emanuel et al., 2023; Abou El-Komboz and Fackler, 2022) and suggests productivity spillovers being at least partly driven by an increase in direct collaborations (as opposed to more indirect colocation benefits). Abstracting from these cluster size effects reveals two central facts: First, there is still a large benefit from colocation in digital knowledge work. Holding economic-area characteristics constant, gravity-type regression analyses suggest colocation is on average associated with about nine times higher collaboration among software developers. Second, geographic distance is of little importance to collaboration apart from the large benefit of colocation.

Although the benefit from colocation is still large for software developers, compared to less digital networks it is much smaller: 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 (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 network where physical

proximity is essential. And while further increased geographic distance is of little relevance in the knowledge worker networks, it remains a strong and defining force for regional connectedness probabilities in the social network.

Estimating the colocation effect for spatial collaboration in different sub-groups discloses considerable heterogeneity that informs about potential drivers of the colocation premium to collaboration. Overall, there is a strong and systematic decline in the size of the colocation effect with cluster size. The largest economic areas feature a colocation effect that is more than ten times smaller than the average effect. This relationship is even better predicted by the presence of large firms that have the potential to facilitate remote collaboration across multiple establishments through their organizational structure. Granular data on the type of collaboration reveals that, indeed, 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. I further find high-quality collaboration less colocated, which points to potentially significant productivity gains from remote collaboration opportunities. 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, pointing to a trade-off between benefits from improved mentor quality and costs arising from remote mentorship.

These findings have important managerial implications, notably for the governance of knowledge worker teams, especially in the information technology sector in the context of the spatial organization of work. Most importantly, findings suggest that it is less important for collaboration in digital knowledge work to be colocated compared to less digital settings. However, heterogeneity in colocation prevalence indicates that (fully) virtual collaboration is feasible to a different degree for different 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 points to colocation still being important for intensive collaboration while non-colocated collaborations typically remain sporadic. For inexperienced workers, colocation with their teams seems to be essential. These findings have wider implications for policy making, in particular that, due to lower colocation requirements for digital collaboration, 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 in their design of policy and organization, that different types of collaboration, even within knowledge-intensive areas, might require different degrees of colocation.

The remainder of this paper is organized as follows. Section 2 discusses related literature. In Section 3, I provide a brief background on digital collaboration in software development and present the data. The empirical analysis in Section 4 first explores the role of colocation and distance for collaboration in the highly digital setting of software developer networks (Section 4.1), compares the observed spatial collaboration pattern to less digital networks (Section 4.2), and explores the drivers colocated collaboration (Section 4.3). Section 4.4 presents robustness assessments and Section 5 concludes with a discussion.

2 Related literature

Agglomeration effects and local spillovers. This work relates to the literature on geographic proximity on economic activity, which originates from the trade literature (Tinbergen, 1962; Bergstrand, 1985). Inspired by the gravity model, other fields adopted similar research designs and find geographic distance relevant, e.g., in scientific research (Catalini, 2018; Head et al., 2019), patenting (Jaffe et al., 1993; Thompson and Fox-Kean, 2005), knowledge transfer (Keller and Yeaple, 2013), and business relations (Cristea, 2011; Coscia et al., 2020; Bahar et al., 2022). Especially complex activities tend to cluster (Balland et al., 2020). Research on software development, where new ICT and digital tools are used heavily, shows strong spatial clustering in Europe (Wachs et al., 2022) and suggests increased distance matters for global collaboration, but less than for trade flows (Fackler and Laurentsyeva, 2020).[3]

While these studies provide consistent evidence for spatial clustering in a diverse set of economic activities, comprehensive insight into spatial collaboration patterns in a setting with the potential to be fully virtual is lacking. This article is the first to show comprehensive and 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.

Although distance explains geographic clustering well it is unclear to what extent physical proximity per se is a requirement for collaboration. Economic theory suggests benefits from geographic proximity arise mainly from costs for moving goods, people, and ideas (Marshall, 1920), and such costs are often but not necessarily a function of geographic distance. Empirically, studies find a high degree of localization of spillovers for productivity (Greenstone et al., 2010; Baum-Snow et al., 2020), in customer-supplier relationships (Arkolakis et al., 2023; Ellison et al., 2010), for knowledge transmission (Glaeser et al., 1992; De La Roca and Puga, 2017), and in labor markets (Moretti and Yi, 2023). Recent evidence shows strong positive spillovers from agglomeration in knowledge-intensive settings, e.g., for inventor (Moretti, 2021), firm (Nagle, 2019) and software developer productivity (Abou El-Komboz and Fackler, 2022), as well as for entrepreneurship (Wright et al., 2021). Empirical work validates that travel cost reductions due to cheap flights (Catalini et al., 2020) and new bridges (Dutta et al., 2022) lead to increased collaboration in science. At the same time, Azoulay et al. (2010) and Waldinger (2012) find physical proximity in scientific publishing less important than intellectual distance.

This study confirms that local characteristics are a key driver of collaboration in digital knowledge work while geographic distance itself is of little relevance. Especially cluster size in terms of the number of local peers explains a large part of spatial agglomeration of collaboration, confirming agglomeration benefits in software development found by Abou El-Komboz and Fackler (2022). Results further suggest more

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

opportunities for direct collaboration (as opposed to more indirect spillovers) in large clusters contribute to agglomeration effects, in line with Azoulay et al. (2010).

Geography and knowledge flows in organizations. Previous work revealed considerable challenges for remote collaboration. For example, distributed teams find it difficult to maintain mutual knowledge (Cramton, 2001), are more prone to conflict (Hinds and Bailey, 2003; Hinds and Mortensen, 2005), feature a lower sense of belonging (Fiol and O'Connor, 2005), shift the perceived ownership of knowledge from the organization to the individual (Griffith et al., 2003), and risk being divided by subgroup dynamics (Polzer et al., 2006). The literature suggests firm organization and management play an important role in addressing these challenges and facilitating collaboration over distance (Zammuto et al., 2007; Majchrzak et al., 2000). For example, Glaeser et al. (2023) find monitoring and managerial guidance lead to increased innovation, which results in an innovation premium when located closer to headquarters. For the manufacturing sector, Giroud et al. (2022) show that local productivity spillovers propagate through plant-level networks within organizations, thereby overcoming distance. Even in the context of improved ICT, Gray et al. (2015) find it beneficial to colocate R&D and manufacturing. Furthermore, the current consensus is that hybrid work organization is most effective (Bloom et al., 2022) and it has long been established that at least occasional face-to-face meetings are important for virtual teams (Maznevski and Chudoba, 2000).

While existing work focuses on the discussion of challenges for organizations in managing remote teams and tools to facilitate collaboration over distance, evidence that compares collaboration within organizations to collaboration between or outside firms is scarce. In contrast, my findings emphasize the role of large organizations in facilitating remote collaboration as opposed to collaboration outside or between organizations. Large organizations, and especially big tech firms, are systematically associated with much smaller colocation effects. At the same time, data suggests that there is still some cost associated with remote collaboration as it tends to be less intense than colocated interactions.

Remote collaboration and technology. Studies on the impact of technology on economic exchange show that improved ICT generally fosters inter-regional trade (Steinwender, 2018; Jensen, 2007), research and innovation (Agrawal and Goldfarb, 2008; Ding et al., 2010; Forman and Van Zeebroeck, 2019), and entrepreneurship (Agrawal et al., 2015). However, geographically close exchange tends to increase disproportionately, for example in research collaboration (Agrawal and Goldfarb, 2008) and bilateral trade (Akerman et al., 2022), in line with theoretical considerations that ICT and geographic proximity are complements (Gaspar and Glaeser, 1998). And although ICT helps to increase remote collaboration, it is unclear if existing technology fully eliminates the benefits of physical proximity. In non-collaborative office settings, remote work is feasible and may even increase productivity (Bloom et al., 2015; Choudhury et al., 2021). However, studies find that face-to-face is still valuable in Silicon Valley firms (Atkin et al., 2022) as well as for communication in white-collar teams (Pentland, 2012). Yang et al. (2022) show that remote collaboration of knowledge workers makes information sharing harder. Similarly, Gibbs et al. (2023) estimates a sizable productivity loss for IT professionals who work remotely which they attribute to increased commu-

nication costs. In the lab, Brucks and Levav (2022) demonstrate virtual interaction comes with a cognitive cost for creative idea generation. There is first evidence that the costs of distributed teams tend to fall over time as remote collaboration technology improves and learning effects materialize (Chen et al., 2022). Within firms, Forman and Zeebroeck (2012) show Internet adoption leads to more geographically dispersed inventor teams.

Apart from the direct effects of remote collaboration on productivity, studies point to physical proximity being central to human-capital development (Glaeser and Mare, 2001; De La Roca and Puga, 2017; Eckert et al., 2022; van der Wouden and Youn, 2023). For inventors, Akcigit et al. (2018) show interaction with successful peers is crucial for innovation. Likewise, Lee (2019) find workspace proximity facilitates individual-level exploration in an office setting in the e-commerce industry. Even among software developers, who regularly interact online and use digital tools, colocation, and online learning are complements such that for firms, a trade-off between short-term productivity gains and long-term human capital development arises (Emanuel et al., 2023).

Here I present comprehensive empirical evidence that shows collaboration is less colocated in a setting of digital knowledge work compared to less digital settings. Furthermore, by exploring colocation of certain types of collaborations I am able to provide nuanced insight into potential drivers of colocation. Evidence points to colocation being especially valuable for inexperienced workers for whom human capital development is important. And the fact that remote collaboration tends to be more high-quality and less intense is in line with higher costs associated with remote collaboration.

Social networks and connectedness. Increased data availability allows researchers to measure interpersonal connectedness in great detail and comprehensively. Bailey et al. (2018a) construct regional connectedness from *Facebook* data. Analyses of this data reveal a high degree of spatial clustering in social networks (Bailey et al., 2020b) and a strong association with travel (Bailey et al., 2020a) and trade (Bailey et al., 2021). Also drawing on *Facebook* data, Chetty et al. (2022a,b) compute social capital measures showing substantial regional variation in social connectedness between people with high and low socio-economic status.

I add to this literature by providing comprehensive insights into the professional networks of software developers, a key and increasingly important group of knowledge workers at the forefront of digital technology adoption. By comparing spatial connectedness patterns to existing comprehensively recorded human networks I show similarities and differences: while all networks exhibit spatial clustering both the functional relationship and magnitude differ widely. Connectedness in less digital social and inventor networks is much more spatially concentrated than in the highly digital software developer network and for the professional networks, there is a dichotomy between colocated and non-colocated collaboration whereas social networks exhibit a much smoother behavior with respect to geography. Further, the knowledge worker network presented here provides much richer insights regarding the nature of collaboration compared to existing professional networks that are comprehensively captured.

3 Data

Background. 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. Survey evidence generated by *GitHub* in 2021 suggests that approximately 19% of code contributions on the platform are to open-source projects (GitHub, 2021). Due to the nature of the version control system *git*, a detailed history of code changes and contributing users is available and openly visible online for public repositories. *GitHub* provides access to public user profiles and repositories via API.

Data. Data analyzed in this paper originates from *GHTorrent*, a research project by Gousios (2013) that mirrors the data publicly available via the *GitHub* API and generates a queryable relational database in irregular time intervals.[4] The resulting snapshots contain data from public user profiles and repositories as well as a detailed activity stream capturing all contributions to and events in public repositories. This paper relies on ten *GHTorrent* snapshots dated between 09/2015 and 03/2021, i.e., roughly one snapshot every seven months.[5] Overall, the data contains 44.1 million users worldwide. For this spatial analysis of software developer collaboration in the United States, the sample of *GitHub* users is selected from this data according to three criteria:

- the user reports a location that refers to a city-level location within the United States;
- the user is active in the observation period, i.e., contributes at least once in two time intervals between data snapshots;[6] and
- the user collaborates, i.e., contributes to at least one project with another in-sample user.

On their *GitHub* profile, users can indicate their location. This self-reported indication is voluntary and is neither verified nor restricted to real-world places by *GitHub*. It is thus difficult to examine the accuracy comprehensively. However, researching user profiles online that can be linked to further personal infor-

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

[5] 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.

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

mation, e.g., due to use of real name on the platform, allows to verify location from other sources such as *LinkedIn* or personal websites. Anecdotal evidence from such searches suggests that those who make a location available on *GitHub* to a large extent provide their correct location.[7] As *GitHub* also functions as a social network for software developers, users have an incentive to report their correct location for networking purposes since they are then more easily found by their local peers.

About 5.2% of users captured in the data (2.30 million) include a self-reported location in their public user profile. Thereof, 34% (778 thousand) can be georeferenced to a location within the United States.[8] This roughly corresponds to a survey conducted by *GitHub* in 2021, reporting a share of 31.5% of users being located in North America (GitHub, 2021). Of these users located in the United States, a portion of 46% (354 thousand) is active in public repositories, which I define as contributing at least once in two time intervals between subsequent data snapshots.[9] Finally, 54% of active U.S. users contribute in at least one project to which multiple users contribute in the observation period. This leaves a sample of 190,637 active, collaborating users geolocated in the United States during the observation period from 2015 to 2021. For the remainder of this paper, I refer to users and their activity in this sample.

For the purpose of regional analysis, each user is assigned to one of 179 economic areas in the United States as defined by the *Bureau of Economic Analysis* based on the self-reported geolocation on her user 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. This regional level is chosen such that it is both sufficiently detailed to study colocation and distance effects and provides an adequate level of aggregation given the number of users in each economic area. The *Bureau of Economic Analysis* economic areas define the "relevant regional markets surrounding metropolitan or micropolitan statistical areas" (Johnson and Kort, 2004). Economic areas are similar to metropolitan statistical areas (MSA) in most cases. To capture entire economic regions, economic areas tend to be larger than corresponding MSAs for big cities.

Summary statistics. In-sample users contribute to about 4.29 million *repositories*, i.e., open-source code projects on the platform. In total, they make roughly 97.3 million single code contributions to these projects, so-called *commits*. The most popular programming languages used on the platform are JavaScript, Python, as well as C and related languages (see Figure A.1). As typical for digital platforms, activity in *GitHub*'s open-source projects is highly skewed, meaning that only a fraction of users contributes the majority of content.[10] See Figure A.3 for a visual impression.

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

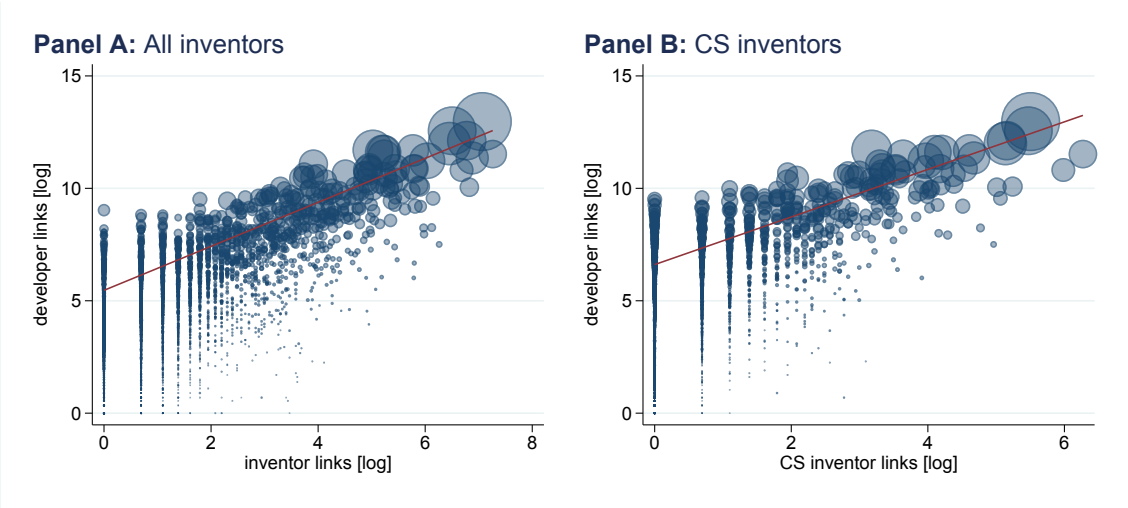
[8] This processing step also confirms above impression that most users provide correct location, as non-sense locations like, e.g., "the moon," together with other locations for which georeferencing to a country was unsuccessful, only make up 1.4% of users with non-empty location.

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

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

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

Figure 1: 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.

I define users as being linked or collaborating with each other if they contribute to at least one joint project in the observation period. There are 10.07 million links between users in the sample. Each user on average is linked to 45.2 other in-sample users (median: 4). Overall, 12.4% of links are between users in the same economic area. For the average user, 34.7% of collaborations are with other local users (median: 14.3%) and two thirds of team projects are fully colocated, meaning that all contributing in-sample users are located in the same economic area. I define links between users that have more than one joint project strong ties. 19% of links between users are strong ties. More detailed summary statistics are reported in Table A.1. To distinguish different types of collaboration I use information provided in the data on the organizational affiliation, forks, stars, and followers (see Section A.1 in the Appendix for more details).

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.[11] 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.2. 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.2). Figure 1 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.[12] 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.

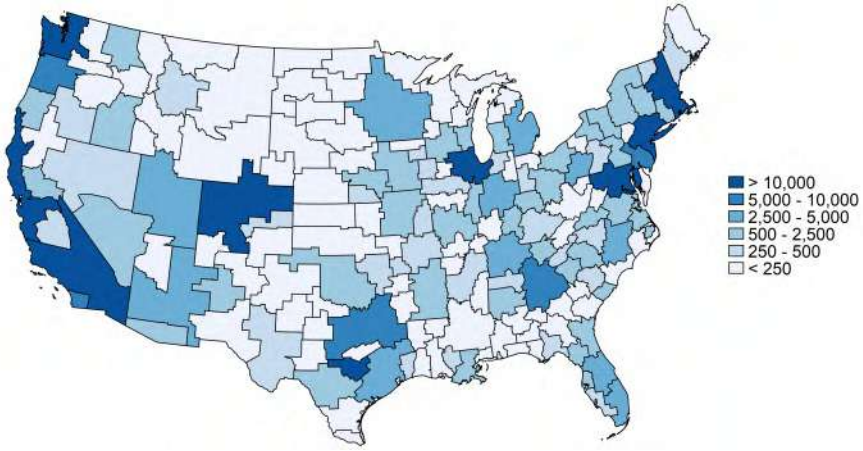
4 Empirical analysis

4.1 Main results

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

[11] For detailed information on supplementary data used here see the Appendix.
[12] For detailed information on supplementary data used here see the Appendix.

Figure 2: Geographic distribution of users



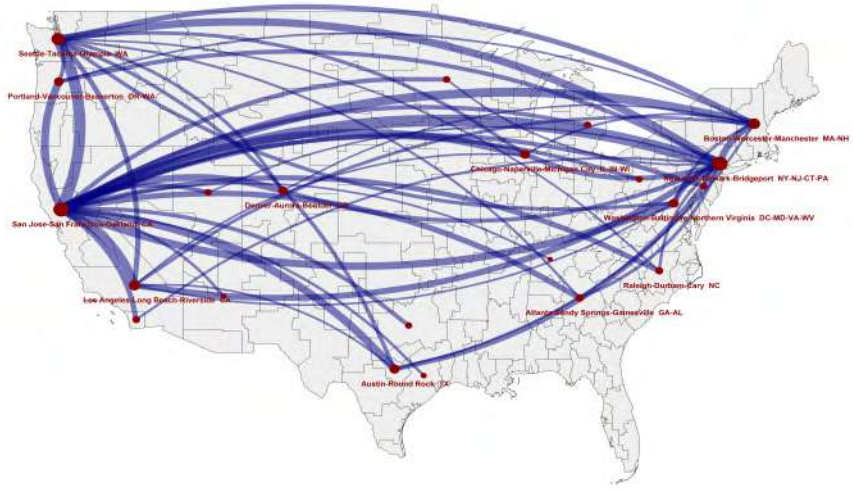
Notes: Map shows the number of (in-sample) users per economic area. The remote economic areas Anchorage, AK, and Honolulu, HI, are not shown. *Sources:* GHTorrent, own calculations.

Concentration is high even among the largest economic areas. While the largest economic area, the Bay Area, hosts over 53 thousand users, only 16.3 thousand users are located in the fifth-largest economic area containing Boston, and less than nine thousand users in the tenth-largest economic area which includes Atlanta. On average an economic area contains 1,895 users with the median economic area hosting 302 users. Normalizing these numbers by economic area population size reveals user density in the general population. Three places stand out here: San Francisco, Austin, and Seattle; all with around 0.5% (in-sample) users in terms of population. Density is less than 0.25% for all other economic areas, for most of them much lower. Collaboration – measured in terms of the number of links users in an economic area are part of relative to the total number of links – is even more concentrated at the top than users. See Figure A.5 for more complete information on the largest twenty economic areas according to these metrics.

Collaboration. Figure 3 provides an overview of the spatial structure of U.S. software developer collaboration network by mapping inter-regional links with above 20,000 collaborations. The strength of inter-regional links is indicated by the width of the blue lines, which is scaled by the logarithmic number of between-economic area user links. Naturally, central nodes correspond to the economic areas with the highest numbers of users (see Figure 2). 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 centers on 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, an average of 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 over-proportionally local by a factor of 7.8. Overall, 12.4% of all links are between users in the same economic area. For the average user, 34.7% of collaborations are with

Figure 3: Inter-regional collaboration of users



Notes: Map shows the structure of the U.S. software developer collaboration network. Important edges of the network, defined as links between economic areas above 20,000 connections, are shown in blue and scaled by the logarithm of the number of links. Economic areas shown in gray with their centroids as nodes in red, scaled by overall links to other economic areas. The remote economic areas Anchorage, AK, and Honolulu, HI, are not shown. *Sources:* GHTorrent, own calculations.

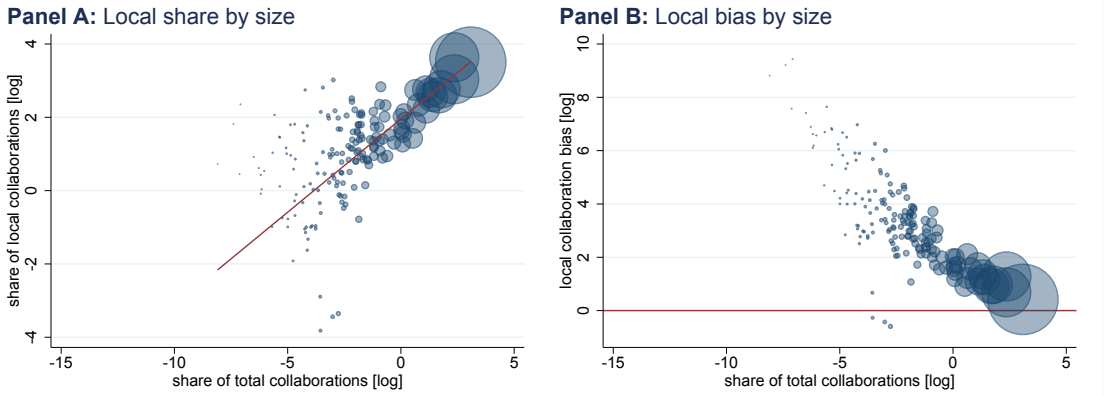
other local users (median: 14.3%), and two-thirds of team projects are fully colocated, meaning that all contributing in-sample users are located in the same economic area. The ten largest economic areas in terms of users are involved in 67.9% of cross-economic area collaborations, a number with relatively little variation across economic areas.[13] Note that this is less than their combined user share of around 80% implying an disproportionally high share of local collaboration relative to other economic areas.

The larger an economic area, measured by total collaboration share, the more of its users' collaborations are typically local. This strong relationship can be intuitively explained by increased opportunity for collaboration in a larger pool of users. However, smaller economic areas with respect to their size disproportionately collaborate more with other local users. This is shown by a strong negative relationship of economic area size and collaboration relative to a hypothetical situation with random sampling, i.e., where links occur with equal probability irrespective of geography. These findings, illustrated by Figure 4, point to high relevance of being colocated for collaboration.

Cluster size, colocation, and distance. To assess the role of cluster size, colocation, and distance in spatial collaboration patterns, I construct binned scatter plots. Panel A of Figure 5 shows a binned scatter plot for the median number of links between economic areas depending on geographic distance, with one point for each percentile of bilateral collaboration counts. Geographic distance in all specifications is the centroid-

[13] See Figure A.6 for a distribution plot.

Figure 4: (Local) collaboration and distance



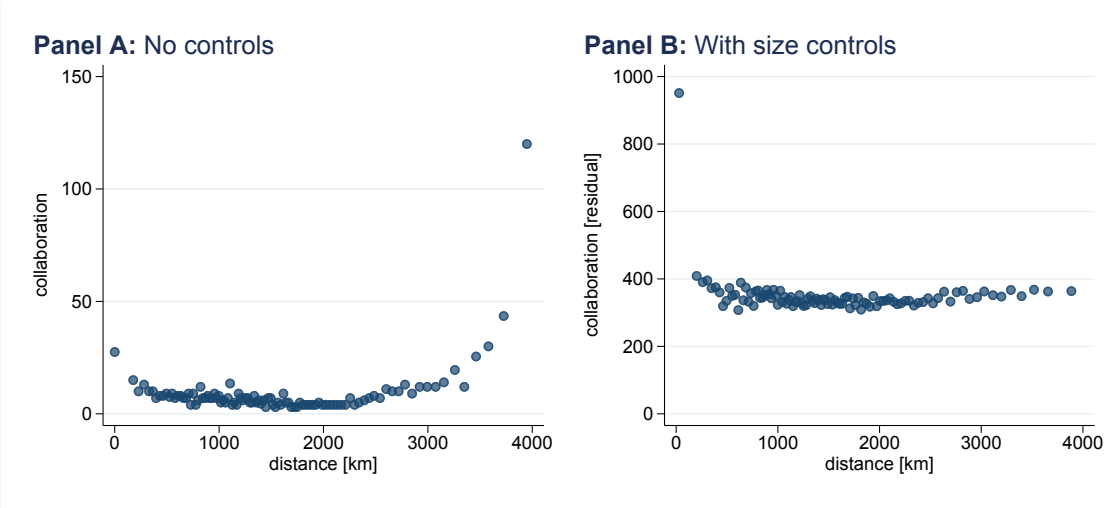
Notes: Plots depict localization patterns in the software developer network. Panel A shows the relationship between the share of collaborations of an economic area's users in all collaborations. The red line represents the best linear fit weighted by total collaboration share as economic-area size measure. Panel B shows the deviation of the economic area user collaboration share from the benchmark of being equal to the percentage share in all collaborations. The horizontal red line (=0) represents this "flat-world" benchmark. Economic areas above the benchmark line feature a higher local collaboration share than their share in total collaborations, economic areas below the benchmark line have a lower share of local collaborations than their share in total collaborations. Bubble size indicates the collaborations of economic area users. *Sources:* GHTorrent, own calculations.

based geodesic distance between economic areas; Figure A.7 plots the distance distribution. The graph shows a U-shaped relationship with a stronger increase in collaborations on the right. This pattern is driven by collaboration between the large economic areas on opposite coasts, which reemphasizes that cluster size is a major driver of collaboration.

To disentangle the effect of cluster size and distance, I construct another binned scatter plot (Panel B) after controlling for a set of variables measuring user size of each economic area pair: the number of users and users squared (to allow for nonlinear effects) for the two economic areas, respectively, and the number of users multiplied for each economic-area pair as a representation of bilateral collaboration potential. Factoring out cluster-size effects yields a collaboration pattern that is essentially flat over the whole distance range, with the notable exception being in the first distance percentile, which captures colocation, for which (residual) collaborations are much higher.[14] Excluding the first percentile, residual medians range between 308 and 409 with a mean of 343. Being colocated (i.e., in the first distance percentile) increases median collaboration by a factor of 2.8 relative to the mean of other percentiles to a (residual) collaboration median of 951, conditional on user size controls. This suggests that, for region pairs with similar cluster size, being colocated is associated with almost three times more collaborations at the median.

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

Figure 5: Collaboration and distance



Notes: Figure shows binned scatter plots of the median number of collaborations and the geographic distance between economic-area pairs. The number of bins is 100, i.e., each point represents one percentile of economic-area pairs. Panel A plots the binned scatter without controls. Panel B plots the binned scatter after controlling for the following variables: users and users squared for both economic areas, respectively, and the multiplication of users of each economic-area pair. Means are added back to residuals before plotting. Within-economic area collaborations as well as Honolulu, HI, and Anchorage, AK, economic areas are excluded. *Sources:* GHTorrent, own calculations.

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

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

where collaborations are explained by a colocation indicator marking collaboration between users in the same economic area, $\mathbb{1}\{\text{coloc}_{i,j}\}$, a distance term, and origin and destination economic-area characteristics.[15] In all specifications I include the continuous centroid-based distance, $\text{dist}_{i,j}$. As control variables, I either include origin and destination economic-area characteristics, $\mathbf{X_i}$ and $\mathbf{X_j}$, or origin and destination economic-area fixed effects. Explicit controls include the number of users, GDP, and population. To control for collaboration potential between two economic areas, I further add the multiplication of origin and destination users, $\mathbf{X_{i,j}}$.

The main results confirm collaboration is strongly positively associated with being colocated. Estimates in Table 1 are remarkably stable overall specifications. The effect size for colocation is large and statistically highly significant, suggesting colocated users collaborate 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

[15] 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.

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.

characteristics, results suggest 0.1% to 0.6% fewer collaborations 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. In line with the literature, the large colocation effect points to direct collaboration with other locals as an important driver of local spillover effects in agglomerations while the importance of other cluster-size controls indicates it is not the only explanation.

Results confirm that economic-area characteristics 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 explained variation increases to 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 further improves model fit to 92%. This implies that around 90% of the variation in regional collaboration is explained by economic-area characteristics, especially cluster size.

4.2 Benchmarks

I am interested in whether the spatial collaboration pattern exhibits less concentration in a digital work setting like software development. As spatial clustering is typical for all human networks, I compare spatial collaboration patterns among software developers to two less digital human networks: the (computer science) inventor collaboration network and general social networks. 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. And although there are other differences than their degree of digitization 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.

4.2.1 Inventor networks

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 that is mostly characterized by non-routine cognitive tasks. Third, both typically work in an office setting with high computer use intensity. Hence, I put the colocation 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.

Inventor collaboration network. I combine data from *PatStat* from 2015 to 2021 with inventor geolocations from the Seliger et al. (2019) and select inventors of collaborative patents located in the U.S. With this information, I define an inventor collaboration link, similar to the definition of software developer collaboration, as having filed at least one joint patent in this period. To get a sample that is as similar as possible to software developers, I select inventors of computer science patents.[16] I arrive at a sample of around 17,000 U.S. inventors that filed a collaborative computer-science patent in this time period.

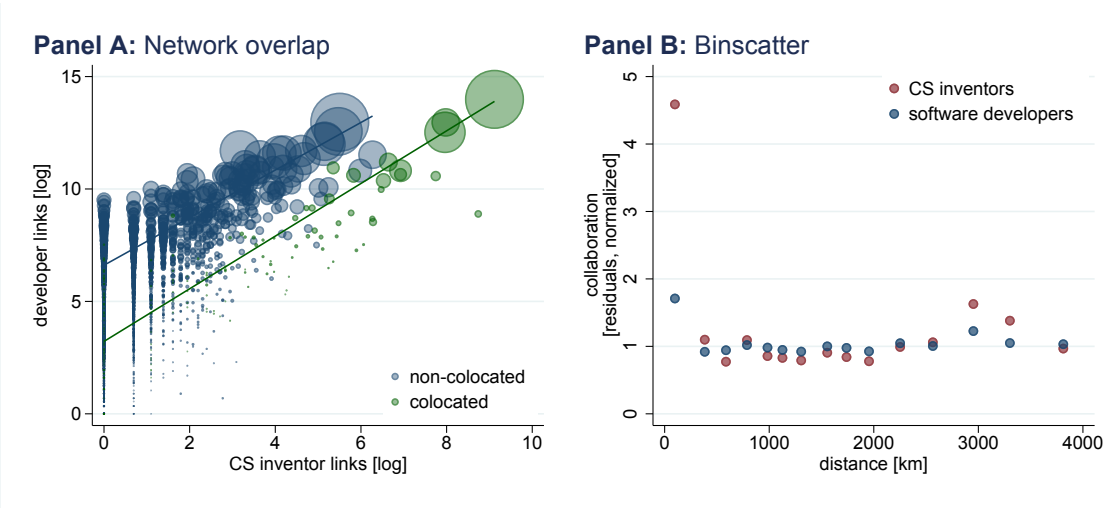
Network relatedness. Panel A of Figure 6 plots the relation between software developer and computer-science inventor networks and differentiates between (blue) and within (green) economic-area collaborations. Marker size represents a measure of economic-area size. There is a strong linear relationship between the two networks. This high inter-regional network overlap means that software developers and inventors exhibit similar inter-regional collaboration patterns.[17] This is an indication that computer science inventors indeed are a viable comparison group for software developers.

Colocation and distance. There is a parallel shift to the right of the green observations in Panel A of Figure 6, representing within-economic area (i.e., colocated) collaborations. This parallel shift in logarithmic values means that, while exhibiting a comparable pattern otherwise, inventor collaborations are systematically more colocated than collaborations in the software developer network. Parallelism also implies that this logarithmic effect is relatively homogeneous across economic areas.

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

[17] Figure A.11 shows a similar plot for all inventors, a larger sample of around 76,000 individuals.

Figure 6: 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. 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. Panel B shows binned scatter plots of the median number of collaborations and the geographic distance between economic-area pairs for both computer-science inventors (red) and software developers (blue). The number of bins is 15. Plots show residuals after controlling for the following variables: users and users squared for both economic areas, respectively, and the multiplication of users of each economic-area pair. 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.

To quantify the difference in colocation effect size between the two networks, Panel B of Figure 6 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 implies the colocation effect is about 2.7 times larger in the computer-science inventor network relative to the software developer network. Regression analyses in Table A.5 confirm this descriptive finding and also point to a two to three times larger colocation effect for inventors, who are about 26 to 28 times more likely to collaborate locally.

Intuitively, a larger colocation effect for inventors of computer science patents compared to software developers is explained by the differences between the two groups. Inventors' work results in a patent (filing) and therefore always claims novelty and, as a result, requires more creativity and innovation in collaboration

processes. 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. Software consists of program code and thus software development tends to be, by nature, more codified than inventing. 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.

4.2.2 Social networks

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 here. A comparison of spatial collaboration patterns in software developer and social networks can inform on differences between (mostly) work-related digital collaboration networks and face-to-face intensive general social networks.

Connectedness indices. To study social networks, I use data on regional connectedness from *Facebook*. Connections on *Facebook* map to a large extent to real-world friendship, family and acquaintanceship ties. As such, observed regional network data constructed form active users on *Facebook* are an adequate representation of real-world social networks.[18] Bailey et al. (2018b) 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 by

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

Importantly, the index is independent of region size and 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).[19] Figure A.12 shows histograms of scaled GHCI and SCI.

Regional network overlap. Interestingly, the two regional connectedness indices are essentially orthogonal to each other, with a low Pearson's correlation of 0.0248 which is not statistically significantly distinguishable from zero. This is also shown by Panel D of Figure 7; a data example for the economic area containing Los Angeles in Figure A.14 provides an illustration. While the (weighted) number of collaborations on *GitHub* is strongly associated with large clusters, this relationship vanishes for the GHCI since it is constructed analogous to the SCI and, therefore, is independent of economic-area size. This shows that software developer and general friendship networks measured through size-independent indices such as GHCI and

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

[19] For details on index construction, and aggregation see the Appendix.

SCI feature no significant regional overlap.[20] Intuitively this is explained by general friendships typically being much more tied to one's geographic center of life.

Comparing spatial decay. Data confirms the presence of a strong colocation effect in both networks. Figure 7 plots raw data from scaled GHCI (Panel A) and SCI (Panel B) after logarithmic transformation. A large colocation effect is already clearly visible in the raw data, represented by the sharp upward shift of the (logarithmic) distribution at a distance of zero for both indices. Apart from the colocation effect, GHCI is essentially independent of distance, in line with the previous findings. In contrast, the SCI features strong and decreasing spatial clustering as depicted by the continued decrease over the whole distance range. The decrease in social connectedness with increasing distance is particularly strong for distances smaller than 500km.

For a model-based comparison of the relationship of the indices to geographic distance, I fit fractional polynomial regressions to flexibly model the relationship in the data.[21] Panel C of Figure 7 graphs the predicted relationships and their fit to the underlying data. The fitted curve in blue represents the relationship between the scaled GHCI and geographic distance while the fitted curve in red shows the same relationship for the scaled SCI. For both indices, there is a clearly visible colocation effect, represented by a discontinuity at a distance of zero. Comparing predicted index values at a distance of zero to the smallest non-zero distance allows me to quantify the colocation effect. The quantification yields an 11.2-fold increase in relative connectedness probability for GHCI. This is larger but comparable to the colocation effect estimated above, which includes more controls. For SCI, the colocation effect is 41.4, i.e., 3.7 times larger than for GHCI. Given further strong spacial decay in SCI and not for GHCI this multiple represents a conservative estimate.

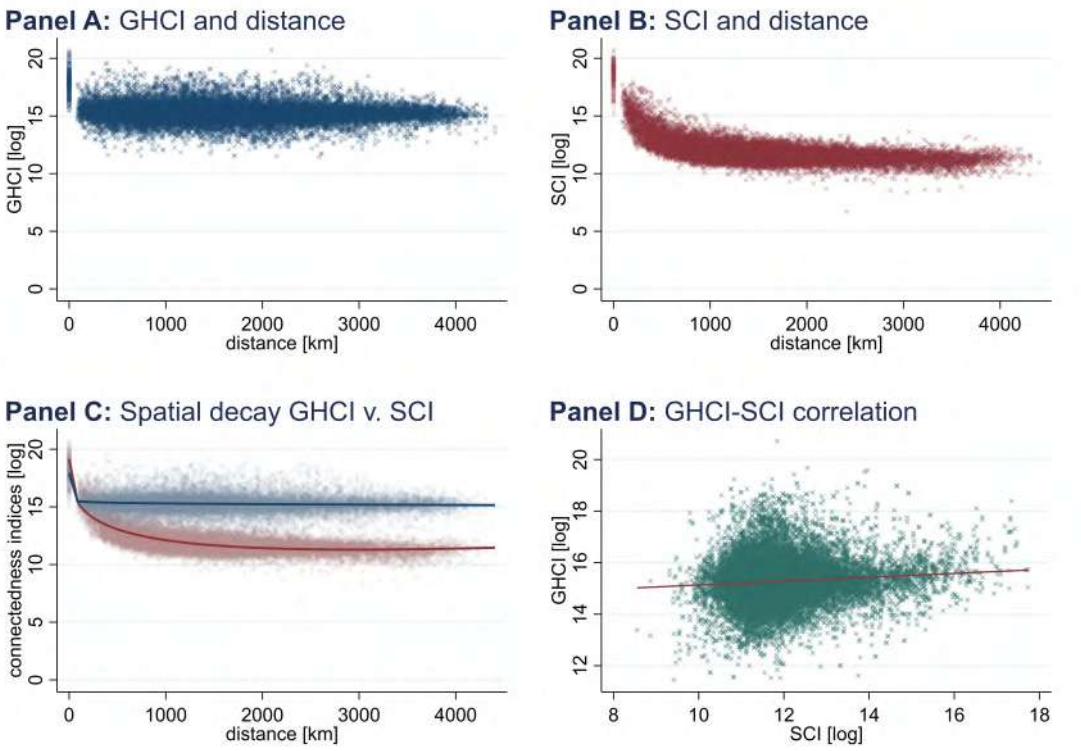
Spatial decay of the relative probability of a connection is present in both indices. It is, however, much more pronounced for predicted SCI and barely visible for the GHCI; Figure A.13 plots the predicted absolute and logarithmic index values with and without the colocation effect on different scales. The data shows that software developer connectedness remains at a much higher (relatively stable) level with increasing distance as compared to social connectedness, which strongly and continuously decreases in distance.

This evidence suggests that connectedness is generally associated with geographic factors in the social network compared to knowledge worker networks. While the colocation effect is larger, as well, looking at colocation alone would be misleading since, additionally, there is a strong and continued spatial decay in connectedness for social networks that is not present in knowledge worker networks. I interpret these findings as evidence that even though the colocation effect in knowledge work is large, it is relatively small when compared to non-digital general social networks.

[20] SCI data is constructed so that it is impossible to tease out the underlying inter-regional network. As a result, network overlap before accounting for region size similar to Panel A in Figure 6 cannot be analyzed here.

[21] See Appendix for detailed information on the fractional polynomial model used.

Figure 7: Relative collaboration probability and distance



Note: Upper Panels show scattered values of scaled GHCI (Panel A) and scaled SCI (Panel B) after logarithmic transformation. Both indices are scaled between 1 and 1,000,000,000. Scaled SCI from Bailey et al. (2018b) is mean-aggregated from county-county level weighted by multiplied populations of each county-pair and rescaled between 1 and 1,000,000,000. Panel C shows the predicted relationship between scaled GHCI (blue) and scaled SCI (red) indices and distance as estimated by a fractional polynomial regression. Logarithmic values of scaled GHCI and SCI are depicted by blue and red markers, respectively. Panels A to C show connected economic-area pair observations. Panel D shows the correlation between scaled GHCI and SCI after logarithmic transformation with within-economic-area collaborations excluded. *Sources:* GHTorrent, Bailey et al. (2018b), U.S. Census Bureau, own calculations.

4.3 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 as well as their affiliation to organizations to measure and classify collaborations along the following dimensions: organizational affiliation, quality, user and project types, and collaboration intensity. This allows me to study which factors are systematically related to a stronger or weaker colocation effect and, hence, to gain further insights into the drivers of and mechanisms behind the observed overall colocation effect.

The descriptive findings in Figure 4 already suggest that the colocation effect might be particularly strong in smaller economic areas and weaker in large hubs. Regression analyses that include interaction terms of

the colocation indicator and economic-area characteristics presented in Table A.10 confirm this descriptive finding. The colocation effect is 28% smaller in economic areas with an above-median number of users compared to a below-median number of users and only 94% smaller in the 10 largest economic areas compared to the rest. There are several potential explanations that lead to this effect heterogeneity at the aggregate economic area level.

Organizations. One potential channel through which this heterogeneity might occur is large organizations (Giroud et al., 2022), i.e. in large economic areas there are also larger firms with multiple establishments that are able to facilitate remote collaboration. For a first indication of this, I run model specifications that interact the colocation indicator with the number of local technology or software firms with above 1,000 employees from *County Business Patterns*. Indeed, the colocation effect is 70% smaller in economic areas with an above-median number of technology firms and 87% smaller in economic areas with an above-median number of software firms. Thus, economic-area characteristics support this view of large firms as facilitators of remote collaboration.

To investigate this channel more directly, I draw on user-indicated affiliation in the data. Around 30% of in-sample users provide their affiliation to an organization. Constructing the economic-area collaboration network from only links between users that both indicate their affiliation and estimating the baseline model specification yields a colocation effect of 5.67, meaning that links of users with affiliation information are 39% less colocated compared to the baseline. This indicates that the sample of users that provide their affiliation generally exhibits a collaboration pattern that is less local. To investigate the role of organizations in facilitating remote collaboration, I distinguish within- and between-firm links within the sample of users that provide their affiliation information. I also classify organizations into groups using the number of affiliated users, big tech firm affiliation, and the number of economic areas of affiliated users. For each of these indicators, I construct two economic-area-level networks according to a decision criterion. For example, I compute the collaboration network for intra- and inter-organizational links at the economic area level. The resulting estimates of the colocation effect from the baseline model specification shown in Panel A of Table 2 are 5.26 for the network of intra-organizational links and 3.73 for the inter-organizational network. This suggests that links within organizations are actually more colocated by 41%.

However, many firms are relatively small and thus have little scope to facilitate remote collaboration.[22] Therefore, it is more appropriate to compare inter- and intra-organizational links of users affiliated with large firms in particular. Defining large organizations as firms with more than 200 affiliated users, I find generally smaller but significant colocation effects of 0.59 for within-large firm collaborations 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. Results are shown in Panel A of Table 2. Similarly, looking at only users affiliated with one of the big tech firms (Amazon, Google,

[22]The organization size distribution is plotted in Figure A.4 in the Appendix.

Apple, Microsoft, or Facebook) yields within-firm collaborations 35% less colocated compared to between-firm links with involvement of a big tech firm user. Generally, big tech firm users exhibit even smaller colocation effects. 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 and computing the respective inter- and intra-organizational collaboration networks yields no differences in the estimated colocation effect but a generally small colocation effect of around 3.5. Overall, these findings provide direct evidence that in particular the largest organizations seem to be successful in facilitating remote collaboration which is in line with the more indirect effects derived from economic-area characteristics in Table A.10. Detailed regression results are presented in Table A.6 in the Appendix.

Quality. Colocated and non-colocated collaboration potentially systematically differs in quality. Theoretically, there are two opposing forces at play. On the one hand, if high-quality projects require more creative and innovative collaboration and, therefore, are more intensive in face-to-face interaction, the colocation effect is expected to be larger for high-quality collaboration. On the other hand, if remote collaboration is more costly because face-to-face interaction is still cognitively easier (Yang et al., 2022; Brucks and Levav, 2022), remote links would tend to form only when there are large expected benefits (i.e., high-quality projects) suggesting a weaker colocation effect for high-quality projects.

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. Using a similar approach as for organizational affiliation to directly measure link quality, I construct economic-area collaboration networks for links between user pairs with an average number of followers that lies above or below the median compared to all links and compare the colocation effect estimates. The results shown in Panel B in Table 2 suggest the colocation effect is 28% smaller for high-quality links with above-median followers. A second measure of quality on *GitHub* is *forks*. Users can fork (public) projects on the platform, i.e., copy the current version to another repository. This is done in cases where the original project is useful in other projects and, therefore, indicates user interest and usefulness in the community. Using the same method as before, I compute two collaboration networks: one for user pairs that have at least one joint project with an above-median number of forks and one for links where users only have joint projects with a below-median number of forks. Using forks as a quality measure, high-quality collaborations are less colocated by 19%. As the last quality measure on the platform, I use *stars*. Users can award stars to repositories on *GitHub* to bookmark them and find the project more easily via a list of starred projects. Hence, stars on a project can be interpreted as an indication of interest in the project by the developer community and thus a sign of project quality. Most projects do not receive any stars so this measure is a quite strong sign of quality. Therefore, I construct collaboration networks for links where at least one joint project has received a star and links where none of the joint projects received a star. In line with the previous results, high-quality collaborations feature a smaller colocation effect. But with a 59% smaller colocation effect, this effect is even larger using this measure. All in all, the data provide support for the view that the team formation cost effect dominates the

Table 2: 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 estimated colocation effects from models similar to the baseline Model (6) in Table 1. The models are estimated using different outcome variables, i.e., the number of links between economic areas, according to various heterogeneity dimensions. Where applicable, relative effects shown refer to effect size ratios between two related models that count collaborations above and below a threshold value of a variable of interest. Relative to the baseline effect is the ratio 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.

face-to-face requirement for high-quality projects. Detailed regression results are presented in Table A.7 in the Appendix.

User type. Another dimension along which the colocation effect might differ is user characteristics. Apart from self-indicated location and affiliation, there are no additional characteristics of users available in the user profile data. However, users' activity data contains useful information that helps to distinguish user types. First, I calculate each user's tenure on the platform from the month of her first commit. Experience with digital collaboration on the platform might lead to learning effects as users get more and more familiar with collaborating remotely. As a result, the colocation effect is potentially smaller for more experienced users. I investigate this hypothesis by computing networks for above- and below-median experience, measured by the average tenure for each user pair. The results in Panel C of Table 2 confirm the hypothesized prediction and suggest the colocation effect is smaller among experienced users by 38%.

Second, links are often formed between users with different experience levels. Compared to links between equally experienced peers, these links are especially beneficial for both the experienced and the inexperienced user: the experienced user gains from the assistance of the inexperienced user while the inexperienced user profits from the other users' experience by observing how to run a project on the platform. If it is true that these links are more valuable to users (Akcigit et al., 2018), they might also be more willing to incur the remote collaboration cost. Thus, remote collaboration is expected to be more prevalent for links with higher experience differentials between users. I test for this by computing this differential and comparing estimated colocation effects for links with above- and below-median experience differentials. In fact, collaboration between users with an above-median experience differential colocate less by 61% as shown in Panel C of Table 2.

Lastly, software developers often specialize in certain programming languages and potentially benefit from division of labor in joint projects where different programming languages are needed. Therefore, links between users with different skills in terms of their programming languages might be especially valuable and hence remote collaboration costs might be less relevant for these links, leading to a lower colocation effect in cross-language collaborations. For each project, the data indicates the programming language a user most often committs in. I define a user's main programming language as the language that most often occurs as the programming language of a project and use this information to identify if collaborating users feature the same main programming language. I then estimate the colocation effect for the network of users with a shared main programming language versus the collaboration network of users with different main programming languages. Results suggest that the colocation effect does not differ significantly in the two networks. Detailed regression results are presented in Table A.9 in the Appendix.

Collaboration intensity. Another dimension along which it is likely that the colocation effect varies is collaboration intensity. There is a vast literature originating from Granovetter (1973) that discusses the role of strong versus weak ties. In this literature, weak ties have been identified as especially valuable in social

networks for information transmission (Rajkumar et al., 2022) and especially to gain new non-redundant information (Yang et al., 2022). If there are costs associated with remote collaboration but at the same time out-of-network links are disproportionately valuable, a natural solution for developers is to engage in remote collaboration, but not as intensively as for more easily to sustain local collaborations, i.e., through weak ties.

A first approach to assess this hypothesis is to use measures of collaboration intensity rather than the number of links between economic areas as outcome variable. Table A.12 presents the regression results from baseline model specification for the number of project and commit links between economic areas as well as the intensity measures commits per project link and commits per user link. Results show that both project links and commit links are colocated to a greater extent than user links. Project links feature a colocation effect that is 2.3 times larger than for user links and commit links are even 9.7 times more colocated than user links. Consequently, collaboration intensity as measured by commits per project is 6.6 times higher locally than non-locally. Measured as commits per link, collaboration intensity is still 2.5 times higher for colocated collaboration.

An even more direct way to study heterogeneity with respect to collaboration intensity is to compute link-level collaboration intensity measures and generate economic-area networks for different collaboration intensity levels. Panel D of Table 2 presents the results for three different metrics of collaboration intensity. First, I use the number of joint projects to calculate a collaboration network for strong and weak links, where weak links are defined as users who collaborate on only one project. I find a 57% stronger colocation effect for strong ties. Second, I distinguish collaboration intensity within joint projects by the average number of commits in joint projects of a user pair and compute networks for above- and below-median project commits. Also here results show that more intense collaborations are more local, but to an even higher degree of 4.4 times. Lastly, I define weak ties via a minimum number of commits. The idea behind this definition is to capture sporadic contributions to other (open-source) projects that do not represent any in-depth collaboration or interaction. Specifically, I label a link as a weak tie when, in all joint projects, a user does not commit more than twice. In line with the other definitions, I find a 2.5 times higher colocation effect for strong ties. Detailed regression results are presented in Table A.11 in the Appendix.

These results suggest that not only do users collaborate much more locally, but also that these local collaborations typically are much more intense than non-colocated collaborations. In fact, colocated and non-colocated collaborations might be of quite different nature in the sense that non-colocated collaboration is of much more sporadic nature, pointing rather towards occasional contributions to other (open-source) projects than to core project team membership.

Project type. Colocation effect size is likely to differ across project types, especially between small and large teams or projects. There are multiple reasons for this presumption. First, larger projects might be more visible and more useful to a broader developer community because they attract a lot of attention and often provide crucial public goods to the community. Second, it might be easier to contribute to large-scale

software development effort that has the organizational mechanisms and contacts in place to allow other users to contribute easily. Third, teams on projects that require a large number of contributing developers might expand their search pool for new team members geographically.

I assess this by constructing networks for large and small projects in terms of users, commits, and project duration. Results are presented in Panel E of Table 2 and support the above hypotheses. Links in projects with below-median team size are much more local than larger teams; the colocation effect for collaborations in small teams is 77% smaller. Similarly, smaller projects in terms of commits exhibit a 31% smaller colocation effect. Longer-running projects are also colocated to a lower degree. They feature a colocation effect that is 72% smaller than for younger projects. Detailed regression results are presented in Table A.8 in the Appendix. These results confirm that large and long-running projects are organized more spatially distributed while small and shorter-running projects are more likely to be colocated.

Relatedness. It is important to assess the degree to which the discussed 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.15. 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.

4.4 Robustness

I run further analyses to assess the sensitivity of the results to changes in the model specification. I start by checking the sensitivity to alternative specifications of 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.

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.8 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 coefficient 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 5. 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).

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.9).

This study follows a cross-sectional approach for multiple reasons. First, a cross-sectional approach makes it possible to obtain the necessary sample size for robust estimation and extract a meaningful and stable network representation. Second, during the observation period from 2015 to 2021 *GitHub* experienced high activity and user growth, and thus changes in the composition of users likely confound dynamic analyses. And third, there are no major events during the observation period that led to aggregate level shifts in platform usage. As a result, I expect to see only gradual dynamic changes, if any, in the colocation effect. As an indication of this, Figure A.10 plots colocation estimates from the baseline model for each time interval. Since sample size reduction leads to more unconnected economic-area pairs, I estimate dynamics for both all and only connected economic-area pairs. In general, results show a quite stable pattern over time. If anything, the colocation effect slightly decreases over time, driven in large parts by the extensive margin, i.e., more connected economic-area pairs. While this intuitively makes sense as a result of the general trend

towards remote (office) work, it is unclear if these patterns represent true dynamics of the colocation effect or rather compositional changes or differences in sample size.

Much of the variation in collaboration across economic areas is explained by economic-area characteristics. In the preferred model I opt for origin and destination fixed effects as well as the multiplication of the number of users in origin and destination as a representation for bilateral collaboration potential. To address potential concerns that other bilateral characteristics drive the colocation and distance effects, I increase model flexibility with respect to such factors by including multiplicative GDP and population as well as squared terms for users, GDP, and population in various constellations. Results are reported in Table A.3. Model fit does not improve significantly when adding these additional control variables. Effects for distance and colocation are comparable in magnitude and precision. Some specifications yield a slightly larger colocation effect while others lead to a slightly smaller effect. I thus conclude that the more parsimonious, preferred specification represents an adequate choice.

The fact that various ways to estimate an effect size for the colocation effect by use of both descriptive and regression analysis yield similar results is generally reassuring. To further validate the robustness of these estimates, I use an alternative to the logarithmic transformation of the outcome variable, the inverse hyperbolic sine (IHS) transformation. IHS transformation avoids the potentially concerning handling of unconnected economic-area pairs that might lead to underestimation of the colocation effect size. Table A.3 reports regression results for various model specifications, contrasting for each specifications the results with log- versus IHS-transformed number of links. The effects are very similar across all comparisons with IHS-transformed estimates being systematically slightly higher. For the main specification, I opt for the more conservative estimates from the models with a log-transformed outcome.

5 Discussion and 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. Even in collaboration networks of software developers, a group with large remote collaboration potential that operates within a highly digital work setting, data shows strong spatial concentration in a few large clusters consistent with strong agglomeration effects. While, indeed, cluster size is strongly associated with collaboration, results emphasize an additional significant positive effect of colocation for collaboration: colocated users collaborate about nine times as much as non-colocated users.

At the same time, however, there is evidence in line with the long-standing prediction that geographic frictions are less relevant in digital work settings. First, apart from the colocation effect I find strong evidence of further increased distance being only of limited relevance for software developer collaboration. Second, the size of the colocation effect is actually relatively small when compared to less digital networks; both social networks and computer science inventor networks exhibit colocation effects more than twice as large.

These findings suggest the relevance of geographic distance for collaboration is indeed subdued in digital knowledge work.

Heterogeneity analyses reveal large differences in the colocation effect for different types of software developer collaboration. Notably, the colocation effect is much smaller within large organizations and in economic areas with a high presence of large technology and software firms. Further, remote collaboration is typically of higher quality and more sporadic and collaboration of inexperienced users is more colocated than for their experienced peers while links between inexperienced and experienced developers are less likely to be colocated. Larger and longer-running projects are more distributed. Overall, this implies the colocation effect is larger in smaller economic areas and smaller in large hubs.

The broad scope and descriptive nature characterizing the contribution of this analysis have limitations. The colocation effect is smaller among software developers compared to less digital settings, but it is unclear to what extent this is due to digitization and ICT use as opposed to other differences between the settings. Likewise, 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.

This study has two main managerial and policy implications. First, colocation is associated with a sizable increase in collaboration even in a digital work setting with corresponding downsides to (fully) remote work whenever collaboration is important. Second, however, the collaboration premium from colocation varies widely depending on the setting's characteristics such as organizational affiliation, collaboration intensity, as well as user and project types. Both innovation policy makers and managers should take this into account when designing incentive structures for knowledge worker teams with respect to colocation. A wider implication for regional and labor market policy is that advanced digitization and ICT potentially attenuate strong agglomeration effects in high-skilled labor markets.

References

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

Agrawal, Ajay and Avi Goldfarb, "Restructuring Research: Communication Costs and the Democratization of University Innovation," *American Economic Review*, 2008, *98* (4), 1578–1590.

_ , Christian Catalini, and Avi Goldfarb, "Crowdfunding: Geography, Social Networks, and the Timing of Investment Decisions," *Journal of Economics & Management Strategy*, 2015, *24* (2), 253–274.

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

Akerman, Anders, Edwin Leuven, and Magne Mogstad, "Information Frictions, Internet, and the Relationship between Distance and Trade," *American Economic Journal: Applied Economics*, 2022, *14* (1), 133–163.

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

Alipour, Jean-Victor, Oliver Falck, and Simone Schüller, "Germany's Capacity to Work from Home," *European Economic Review*, 2023, *151*, 104354.

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.

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

Bahar, Dany, Prithwiraj Choudhury, Do Yoon Kim, and Wesley W. Koo, "Innovation on Wings: Nonstop Flights and Firm Innovation in the Global Context," *Management Science*, 2022.

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

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

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

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

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

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

Balland, Pierre-Alexandre, Cristian Jara-Figueroa, Sergio G Petralia, Mathieu P. A. Steijn, David L. Rigby, and César A. Hidalgo, "Complex Economic Activities Concentrate in Large Cities," *Nature Human Behaviour*, 2020, *4* (3), 248–254.

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

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

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

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

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

Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying, "Does Working from Home Work? Evidence from a Chinese Experiment," *The Quarterly Journal of Economics*, 2015, *130* (1), 165–218.

_ , **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.

_ , Christian Fons-Rosen, and Patrick Gaulé, "How Do Travel Costs Shape Collaboration?," *Management Science*, 2020, *66* (8), 3340–3360.

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.

Chen, Chinchih, Carl Benedikt Frey, and Giorgio Presidente, "Disrupting Science," *Working Paper*, 2022.

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

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

Choudhury, Prithwiraj, Cirrus Foroughi, and Barbara Larson, "Work-from-Anywhere: The Productivity Effects of Geographic Flexibility," *Strategic Management Journal*, 2021, *42* (4), 655–683.

Coscia, Michele, Frank M. H. Neffke, and Ricardo Hausmann, "Knowledge Diffusion in the Network of International Business Travel," *Nature Human Behaviour*, 2020, *4* (10), 1011–1020.

Cramton, Catherine Durnell, "The Mutual Knowledge Problem and Its Consequences for Dispersed Collaboration," *Organization Science*, 2001, *12* (3), 346–371.

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

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

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

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

Ding, Waverly W., Sharon G. Levin, Paula E. Stephan, and Anne E. Winkler, "The Impact of Information Technology on Academic Scientists' Productivity and Collaboration Patterns," *Management Science*, 2010, *56* (9), 1439–1461.

Dingel, Jonathan I. and Brent Neiman, "How Many Jobs Can Be Done at Home?," *Journal of Public Economics*, 2020, *189*, 104235.

Dutta, Sunasir, Daniel Erian Armanios, and Jaison D. Desai, "Beyond Spatial Proximity: The Impact of Enhanced Spatial Connectedness from New Bridges on Entrepreneurship," *Organization Science*, 2022, *33* (4), 1620–1644.

Eckert, Fabian, Mads Hejlesen, and Conor Walsh, "The Return to Big-City Experience: Evidence from Refugees in Denmark," *Journal of Urban Economics*, 2022, *130*, 103454.

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

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

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

Fiol, C. Marlene and Edward J. O'Connor, "Identification in Face-to-face, Hybrid, and Pure Virtual Teams: Untangling the Contradictions," *Organization Science*, 2005, *16* (1), 19–32.

Forman, Chris and Nicolas van Zeebroeck, "From Wires to Partners: How the Internet has Fostered R&D Collaborations within Firms," *Management Science*, 2012, *58* (8), 1549–1568.

_ and Nicolas Van Zeebroeck, "Digital Technology Adoption and Knowledge Flows within Firms: Can the Internet Overcome Geographic and Technological Distance?," *Research Policy*, 2019, *48* (8), 103697.

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

Gaspar, Jess and Edward L. Glaeser, "Information Technology and the Future of Cities," *Journal of Urban Economics*, 1998, *43* (1), 136–156.

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.

Glaeser, Chloe Kim, Stephen Glaeser, and Eva Labro, "Proximity and the Management of Innovation," *Management Science*, 2023, *69* (5), 3080–3099.

Glaeser, Edward L. and David C. Mare, "Cities and Skills," *Journal of Labor Economics*, 2001, *19* (2), 316–342.

_ , Hedi D. Kallal, Jose A. Scheinkman, and Andrei Shleifer, "Growth in Cities," *Journal of Political Economy*, 1992, *100* (6), 1126–1152.

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

Granovetter, Mark S., "The Strength of Weak Ties," *American Journal of Sociology*, 1973, *78* (6), 1360–1380.

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.

Griffith, Terri L., John E. Sawyer, and Margaret A. Neale, "Virtualness and Knowledge in Teams: Managing the Love Triangle of Organizations, Individuals, and Information Technology," *MIS Quarterly*, 2003, pp. 265–287.

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.

Hinds, Pamela J. and Diane E. Bailey, "Out of Sight, Out of Sync: Understanding Conflict in Distributed Teams," *Organization Science*, 2003, *14* (6), 615–632.

_ **and Mark Mortensen**, "Understanding Conflict in Geographically Distributed Teams: The Moderating Effects of Shared Identity, Shared Context, and Spontaneous Communication," *Organization Science*, 2005, *16* (3), 290–307.

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.

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

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

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

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

Lee, Sunkee, "Learning-by-Moving: Can Reconfiguring Spatial Proximity Between Organizational Members Promote Individual-level Exploration?," *Organization Science*, 2019, *30* (3), 467–488.

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

Majchrzak, Ann, Ronald E. Rice, Arvind Malhotra, Nelson King, and Sulin Ba, "Technology Adaptation: The Case of a Computer-supported Inter-organizational Virtual Team," *MIS Quarterly*, 2000, pp. 569–600.

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.

Marshall, Alfred, *Principles of Economics*, MacMillan, 1920.

Maznevski, Martha L. and Katherine M. Chudoba, "Bridging Space Over Time: Global Virtual Team Dynamics and Effectiveness," *Organization Science*, 2000, *11* (5), 473–492.

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.

Pentland, Alex Sandy, "The New Science of Building Great Teams," *Harvard Business Review*, 2012, *90* (4), 60–69.

Polzer, Jeffrey T., C. Brad Crisp, Sirkka L. Jarvenpaa, and Jerry W. Kim, "Extending the Faultline Model to Geographically Dispersed Teams: How Colocated Subgroups Can Impair Group Functioning," *Academy of Management Journal*, 2006, *49* (4), 679–692.

Rajkumar, Karthik, Guillaume Saint-Jacques, Iavor Bojinov, Erik Brynjolfsson, and Sinan Aral, "A Causal Test of the Strength of Weak Ties," *Science*, 2022, *377* (6612), 1304–1310.

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*, John Wiley & Sons, 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.

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

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

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

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

van der Wouden, Frank and Hyejin Youn, "The Impact of Geographical Distance on Learning through Collaboration," *Research Policy*, 2023, *52* (2), 104698.

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

Waldinger, Fabian, "Peer Effects in Science: Evidence from the Dismissal of Scientists in Nazi Germany," *The Review of Economic Studies*, 2012, *79* (2), 838–861.

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

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

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

Zammuto, Raymond F., Terri L. Griffith, Ann Majchrzak, Deborah J. Dougherty, and Samer Faraj, "Information Technology and the Changing Fabric of Organization," *Organization Science*, 2007, *18* (5), 749–762.

A Appendix

A.1 Supplementary information

Organizations. Similar to locations, users can indicate their affiliation on *GitHub*. To analyze within- and between-organization collaboration patterns, I select links where both users self-report their affiliation. There are 1,095,141 links where both users report an affiliation, reducing the sample to 57,616 U.S. users (30% of the total sample of 190,637 U.S. users).[23] Fuzzy matching is combined with manual data cleaning to harmonize the reported affiliations. This yields 37,997 distinct organizations with an average number of 6.1 affiliated users, but about 44% of organizations are represented through only one user in the data.[24] Big tech firms are identified as Amazon, Google, Microsoft, Apple, or Facebook. 8.3% of users are affiliated with big tech firms. I define large organizations as organizations with more than 200 affiliated users. There are 65 large organizations and 18.9% of users are affiliated with a large firm. For the purpose of this analysis, I define multi-establishment organizations as organizations with in-sample users in five or more economic areas. There are 7,248 multi-establishment organizations with an average of 12.9 locations. 53.3% of users are affiliated with a multi-establishment organization.

Quality. As measures for collaboration quality I use information in the data on followers, forks, and stars. Users on *GitHub* can follow each other so that the number of followers a user has is an indicator for her popularity among other users on the platform. I calculate the average number of followers in each collaboration (user-pair) as a measure of popularity of these contributors. The median user-pair average number of followers is 8. Repositories on *GitHub* can be *forked*, i.e., copied into other projects. This allows amending and extending code from other projects without altering the original code when having no write access to open development branches in the original repository. Forked code is either re-used and extended in other projects or further developed to propose integration into the original repository. Therefore, forks can be seen as indicator for the usefulness of a project to other users. I calculate the number of forks in each project as a project quality measure. The median number of forks is 5. Repositories can also be awarded *stars* by users. Starring on *GitHub* essentially is a bookmarking functionality. Users can access a list of all projects they have starred to more easily find them and *GitHub* recommends similar projects to users based on this list. Thus, receiving stars is an indicator that other users find a project interesting. Only 38.0% of projects are awarded a star from at least one other user.

Project types. I compute various metrics as project characteristics. First, team size is calculated as the number of (in-sample) users contributing to a project in the observation period. Median team size is two; note that this is also the minimum number of users by the way I constructed the sample. Second, I calculate the sum of commits to a project as a measure of both project complexity and size. The median number of

[23] Interestingly, almost all links with affiliation (in total 1,095,380) are links where both users report their affiliation.

[24] See Figure A.4 for the size distribution.

commits in a project is 15. Lastly, project age is defined here as the number of months since the month of the first commit in a project. This number features a median of 11 months.

User types. Measures of user-pair characteristics are derived from user activity data. First, I count the average number of commits to a project in the observation period for each user-pair. To get a measure of average user engagement, I take the mean of this number across all joint projects. For the median user-pair, each user commits on average three times to a joint project. As a measure of user age and experience, I calculate tenuere on the platform as the time in months since a users' first commit. For each user-pair, I average this number. The median user-pairs' average tenure on *GitHub* is 11.5 months. From this measure, I derive for each user-pair the difference in experience in months. The median user-pair has an experience difference of 7 months. Lastly, since the data provides the programming language most used in each project for each user, I identify the most-used programming language for all users by aggregation across projects and then mark user-pairs where both users feature the same main programming language in at least one joint project. In 27.3% of user-pairs both users code the same (main) programming language in at least one joint project.

Strong and weak ties. To measure collaboration intensity at the link level, I use two different measures to distinguish strong and weak ties between users. As first method, I define a link between two users as strong if they contribute to more than one joint project in the observation period. According to this definition, 19% of links between users are strong ties. To get at the collaboration intensity within joint projects, I use a second method where I define a link as weak if in all joint projects at least one of the users commits twice or less. According to this definition, 74% of links between users are weak ties.

Collaboration intensity. At the economic-area pair level, I calculate various measures for collaboration (intensity) next to the number of user links. I define two measures of overall collaboration between economic-area pairs: First, I count project-level links, i.e., user pairs with multiple joint projects are counted according to the number of joint projects. Second, I use the sum of commits in each user pair and then aggregate this number to economic-area pairs. Further, I define two measures of collaboration intensity between economic-are pairs: Fist, I measure collaboration intensity per project as the ratio of overall commits per economic-area pair relative to the number of projects between two economic areas. Second, I calculate a similar ratio for each economic-area pair using the average number of commits per user-pair.

Connectedness indices. GHCI and SCI indices are calculated using Equation 2. SCI data on the county-county level is taken from Bailey et al. (2018b)[25] and aggregated to economic-area level using the methodology suggested in Bailey et al. (2021):

$$\text{SCI}_{i,j} = \sum_{r_i \in R(i)} \sum_{r_j \in R(j)} \text{PopShare}_{r_i} * \text{PopShare}_{r_j} * \text{SCI}_{r_i,r_j} \tag{3}$$

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

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 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

$$\text{I} \rightarrow \frac{\text{I} - \min(\text{I})}{\max(\text{I}) - \min(\text{I})} * [\text{S}_{\max} - \text{S}_{\min}] + \text{S}_{\min} \tag{4}$$

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

Index aggregation. Here I reproduce the derivation of Equation 3 used to aggregate the index to economic-area level from Bailey et al. (2021):

$$\begin{aligned}
\text{SCI}_{i,j} &= \frac{\text{links}_{i,j}}{\text{pop}_i * \text{pop}_j} \\
&= \frac{\sum_{r_i \in R(i)} \sum_{r_j \in R(j)} \text{links}_{r_i,r_j}}{\sum_{r_i \in R(i)} \text{pop}_{r_i} * \sum_{r_j \in R(j)} \text{pop}_{r_j}} \\
&= \sum_{r_i \in R(i)} \sum_{r_j \in R(j)} \frac{\text{pop}_{r_i}}{\sum_{r_i \in R(i)} \text{pop}_{r_i}} \frac{\text{pop}_{r_j}}{\sum_{r_j \in R(i)} \text{pop}_{r_j}} \frac{\text{links}_{r_i,r_j}}{\text{pop}_{r_i} * \text{pop}_{r_j}} \\
&= \sum_{r_i \in R(i)} \sum_{r_j \in R(j)} \text{PopShare}_{r_i} * \text{PopShare}_{r_j} * \text{SCI}_{r_i,r_j}
\end{aligned} \tag{5}$$

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 PopShare_{r_i}.

Fractional polynomials. For the purpose of estimating a smoothed yet flexible 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 $p_i \in \{2, 1, 0.5, 0, 0.5, 1, 2, 3\}$ suggested in Royston and Sauerbrei (2008) by

$$x^{(p_1,p_2,\ldots,p_m)\prime}\beta = \beta_0 + \beta_1 x^{(p_1)} + \beta_2 x^{(p_2)} + \ldots + \beta_m x^{(p_m)} \tag{6}$$

where $x^{(0)} = \ln x$ and each repeated power multiplies with another $\ln x$.

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 "Pro-

fessional, 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)[26] and take averages for the years 2014 to 2020.[27] From the *U.S. Census Bureau* I use county-level population estimates and apply the same aggregation procedure.[28] 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 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).[29] Inventor latitude and longitude are assigned to economic areas using the economic area shape file by the *Bureau of Transportation Statistics*.[30] 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.

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

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

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

[29] Retrieved at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/OTTBDX.

[30] Retrieved at https://maps.princeton.edu/catalog/harvard-ntadbea.

A.2 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: 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.6: 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.7: 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.8: 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.9: 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.10: 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.11: 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.

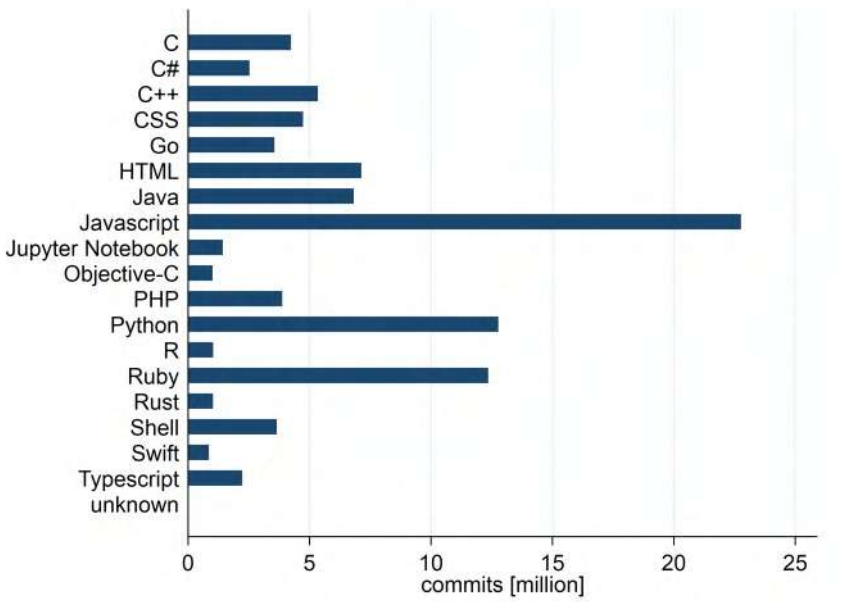
Table A.12: Colocation and collaboration intensity

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.

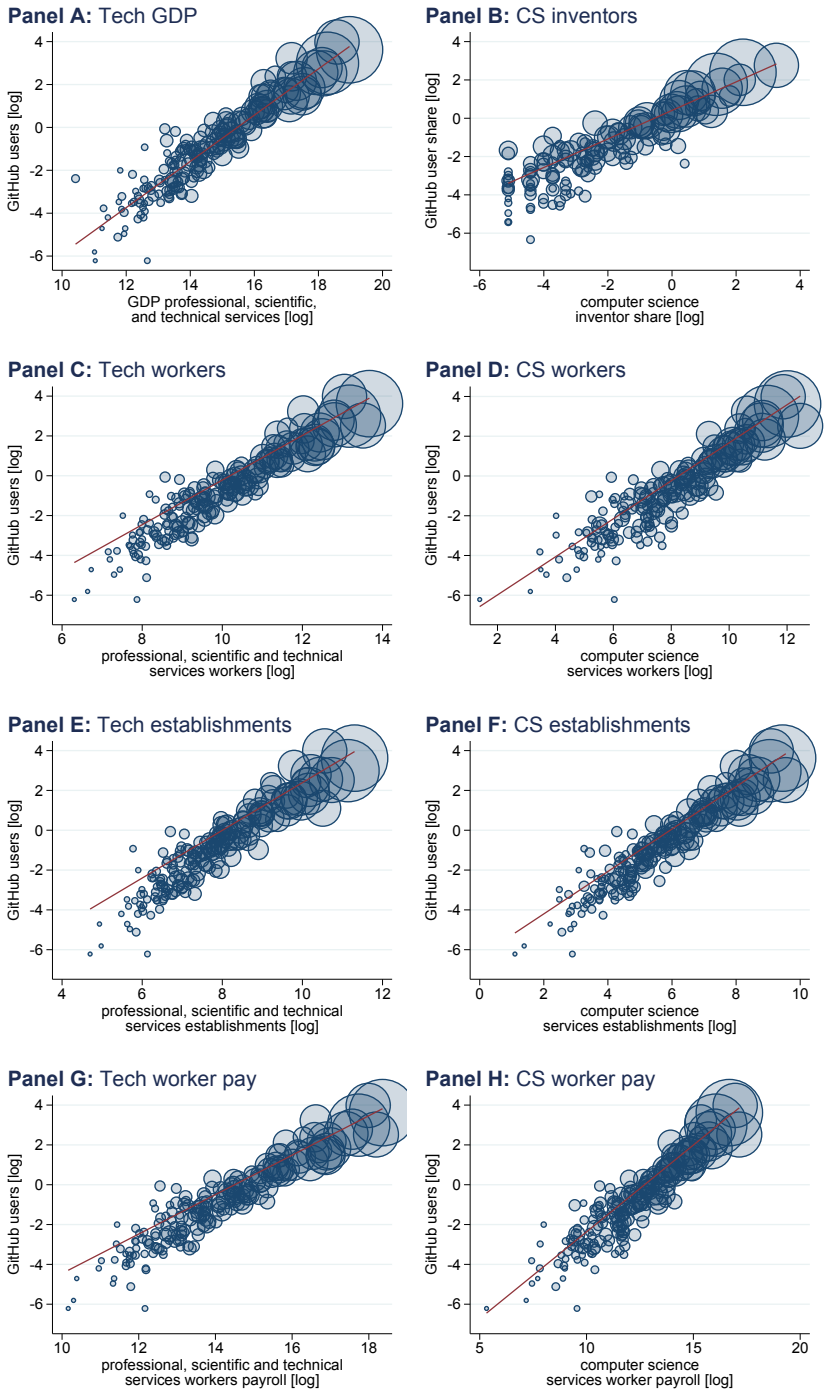
A.3 Figures

Figure A.1: Programming languages



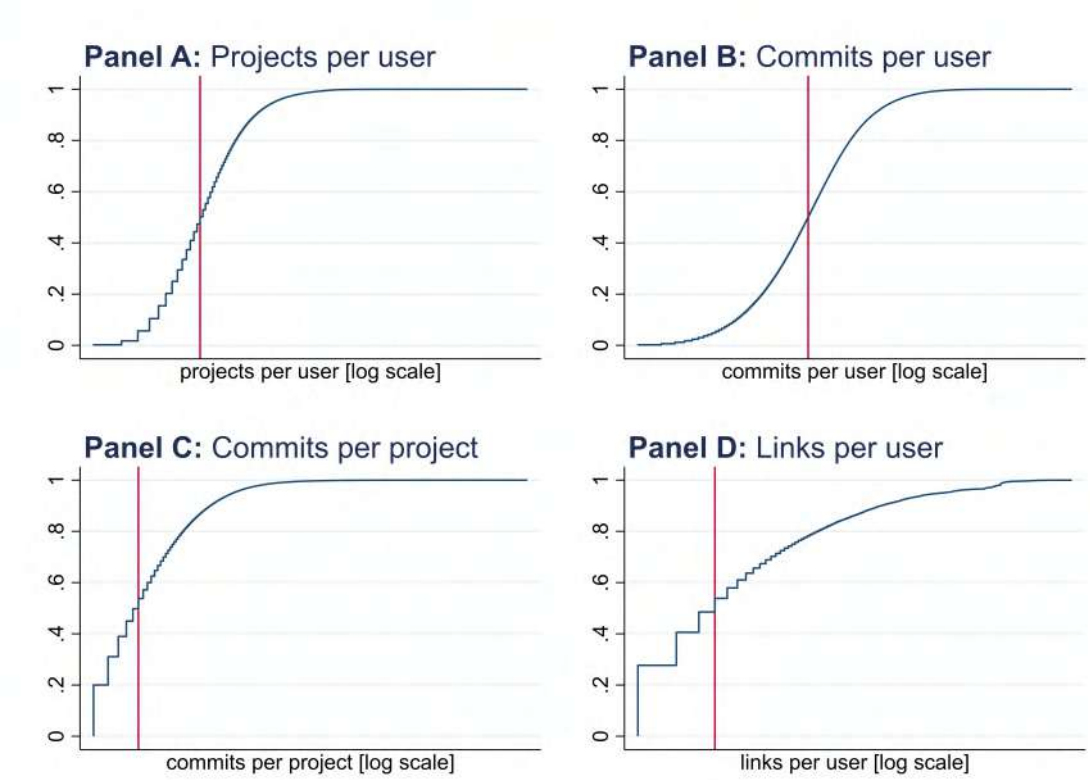
Note: Bars show the number of *commits* contributed to open-source projects by active, collaborating users in the United States in the observation period for each programming language. Unknown refers to *commits* that are not assigned to a programming language in the data. *Sources:* GHTorrent, own calculations.

Figure A.2: Representativness



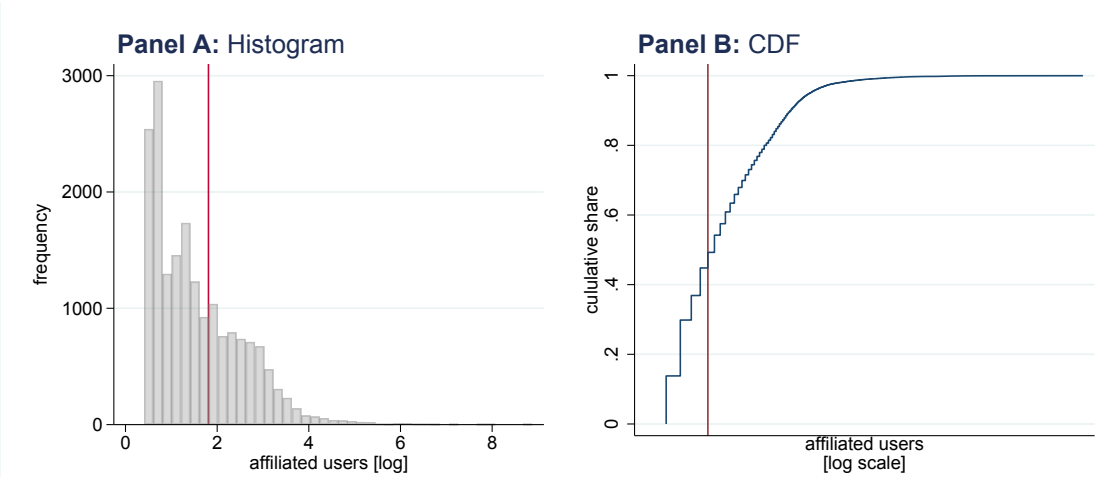
Note: Plots show the relationship between (the share of) users per economic area and economic-area level metrics related to software development after logarithmic transformation. Bubble size represents economic-area population size. Red lines are best linear fits from user-weighted log-log regressions.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

Figure A.3: CDFs of user activity



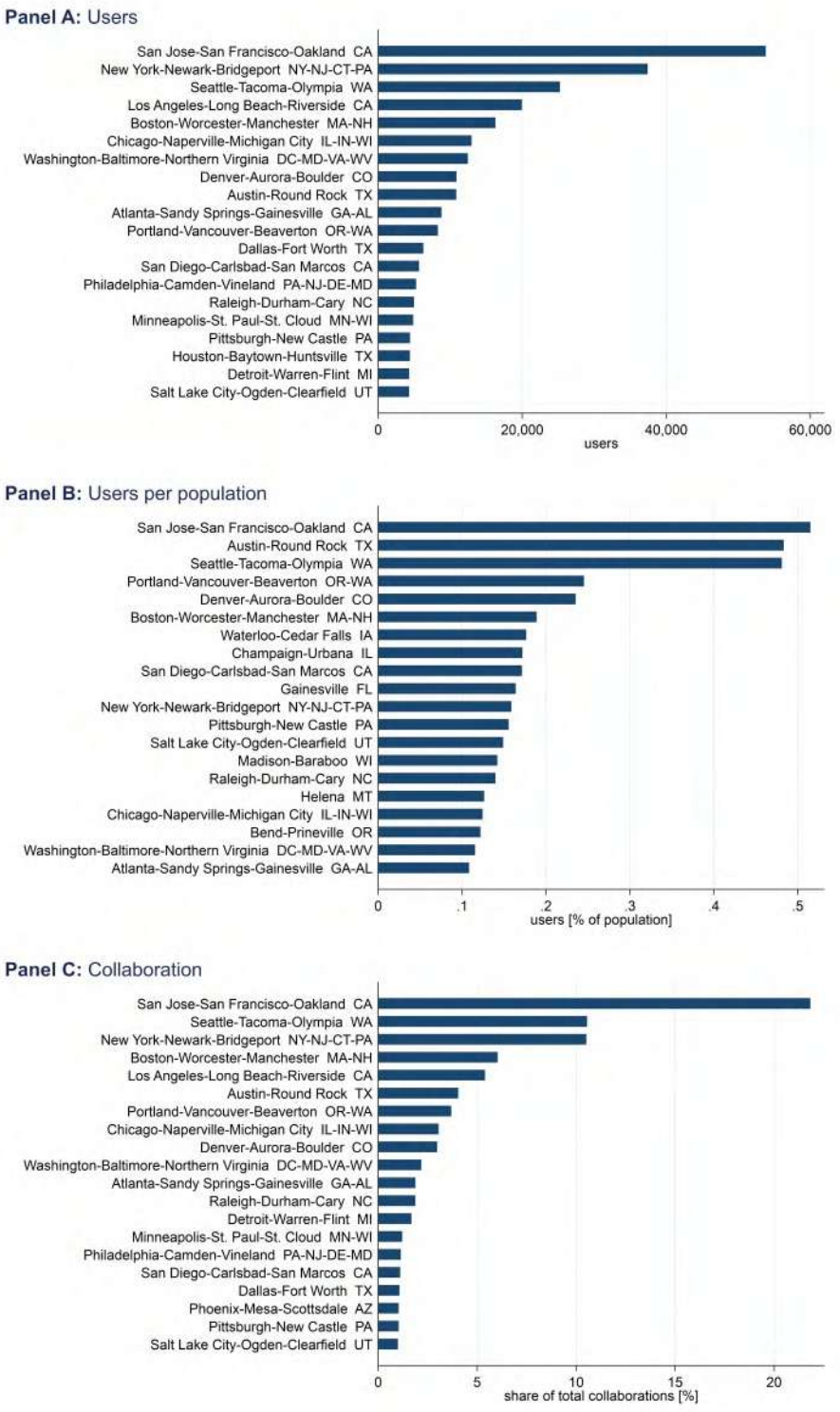
Note: Plots show cumulative density functions for different user activity metrics. Vertical red lines represent median values of each metric (i.e., projects per user: 14; *commits* per user: 156; *commits* per project: 7; links per user: 4). All x-axes are scaled logarithmically. The graph for *commits*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

Figure A.4: Organization size



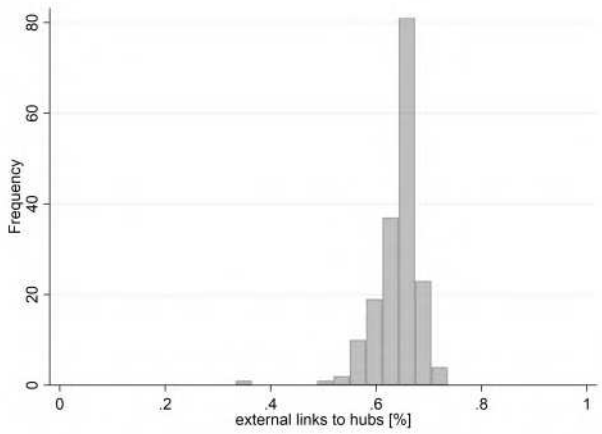
Notes: Plot shows the distribution of organization size as measured by number of affiliated users. Panel A shows a histogram and Panel B a cumulative distribution function. The horizontal red line indicates mean (6.1; histogram) and median (3.5; CDF) affiliated users. Organizations with only one affiliated user are excluded. *Sources:* GHTorrent, Bureau of Economic Analysis, own calculations.

Figure A.5: Concentration at the top



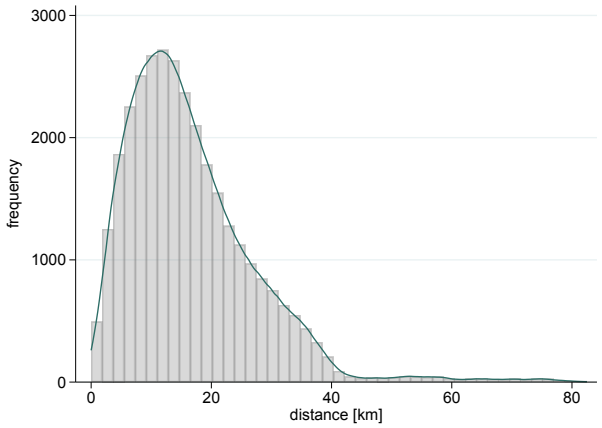
Notes: Plots show the values of different user and activity concentration metrics for the twenty largest economic areas in terms of respective metrics. *Sources:* GHTorrent, own calculations.

Figure A.6: Collaboration with hubs



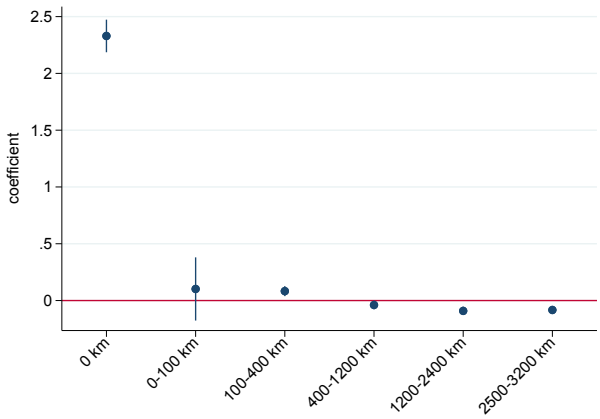
Notes: Plot shows the distribution of collaboration shares of each economic area with hubs, defined as the ten largest economic areas in terms of users. *Sources:* GHTorrent, Bureau of Economic Analysis, own calculations.

Figure A.7: Distance



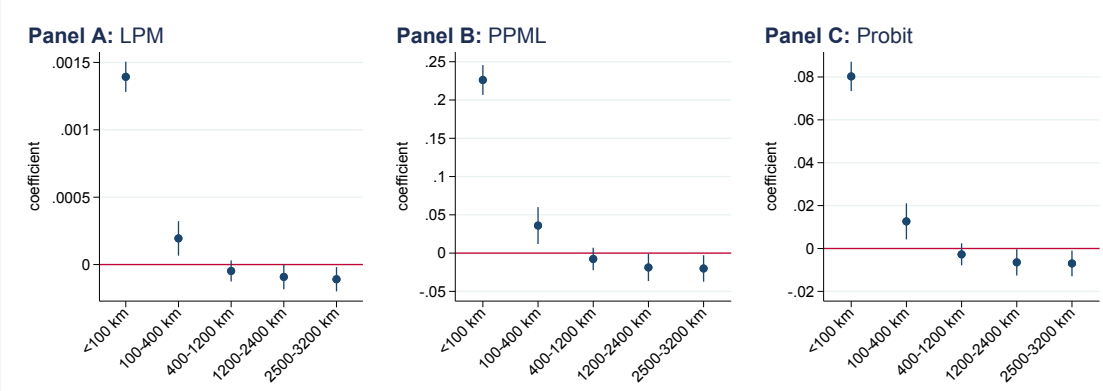
Notes: Plot shows the distribution of centroid-based geodesic distance between economic areas. The horizontal red line indicates the median distance of 1,439. The blue curve represents the Epanechnikov kernel density estimate. The right tail of the distribution starting approximately at distances greater 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.8: Non-parametric distance



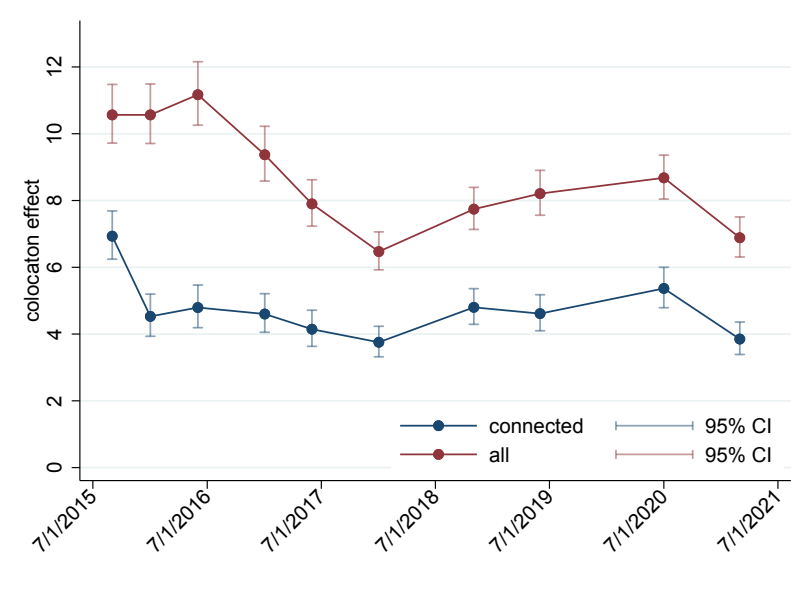
Notes: Plot shows coefficient point estimates and confidence intervals for the baseline fixed effects model specification with non-parametric distance. The indicator for distances above 3,200 km is omitted. Blue bars show 95% confidence intervals from robust standard errors. Collaborations with Anchorage,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

Figure A.9: Individual-level probability models



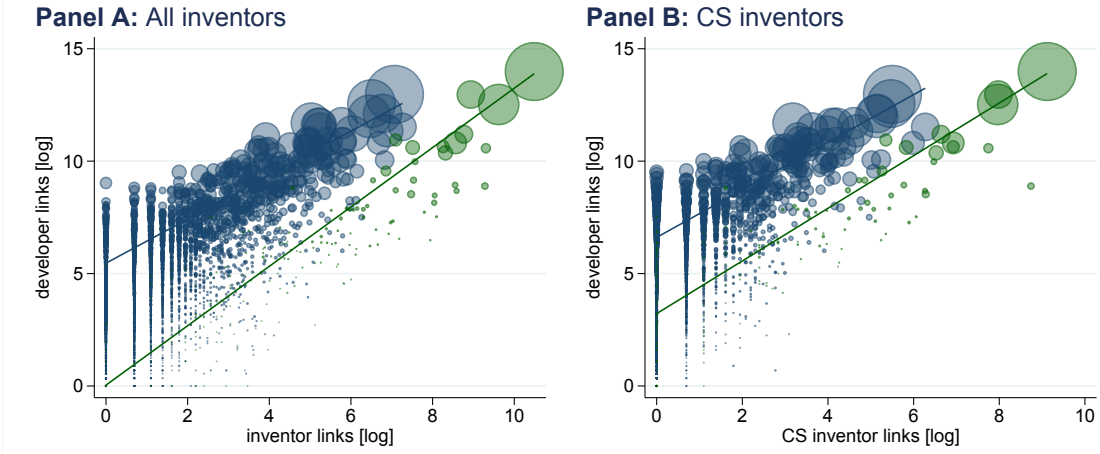
Notes: Plot shows coefficient point estimates and confidence intervals for the individual-level fixed effects model specification with non-parametric distance from Table A.4. The indicator for distances above 3,200 km is omitted. Blue bars show 95% confidence intervals from robust standard errors. Collaborations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

Figure A.10: Colocation dynamics



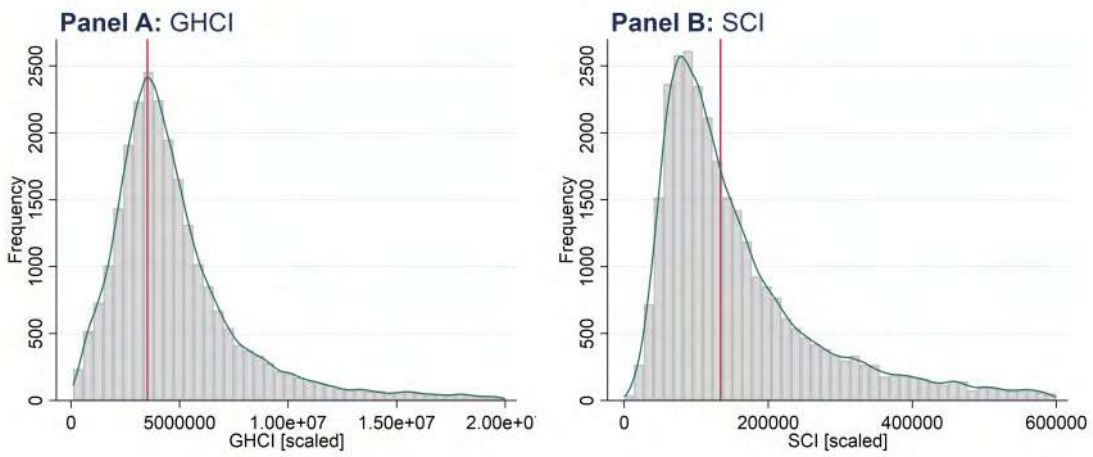
Notes: Plot shows coefficient point estimates and confidence intervals for the baseline fixed effects model specification with non-parametric distance. The indicator for distances above 3,200 km is omitted. Blue bars show 95% confidence intervals from robust standard errors. Collaborations with Anchorage,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

Figure A.11: Colocation effect relative to inventors



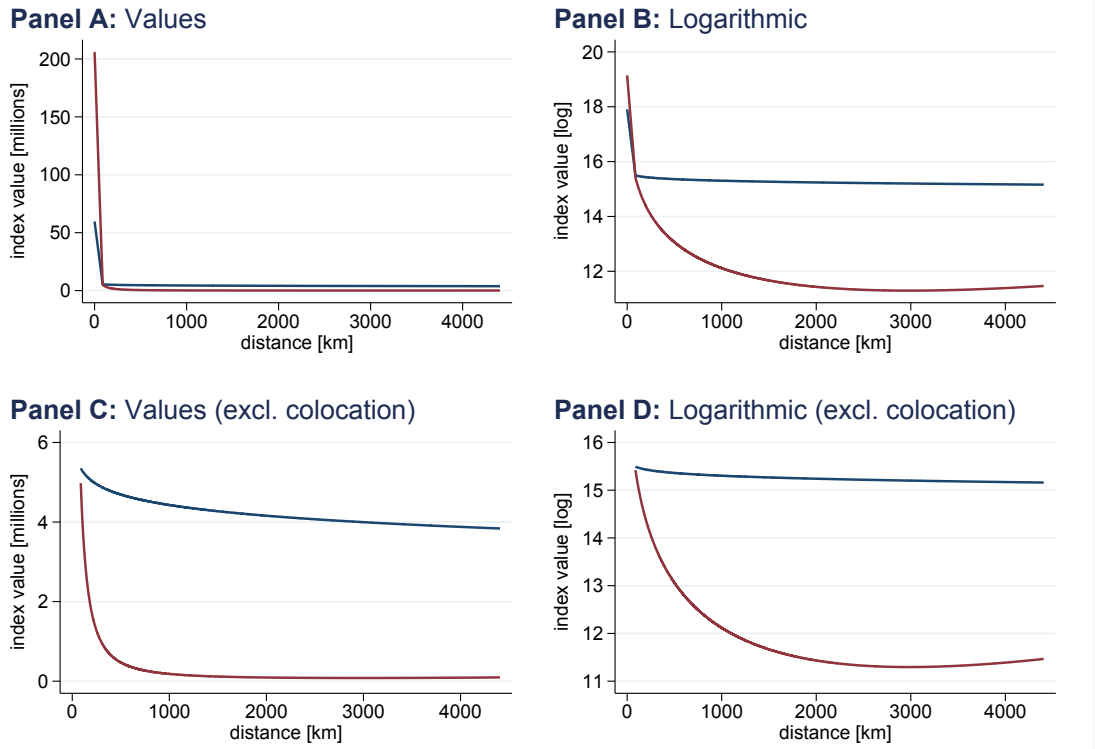
Note: Plots show the relationship between the number of collaborations between economic areas in the software developer and inventor network. Panel A compares software developer collaborations to all collaborations in collaborative patents and Panel B to collaborative computer science patents. Collaborations 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.12: Histograms of scaled GHCI and SCI



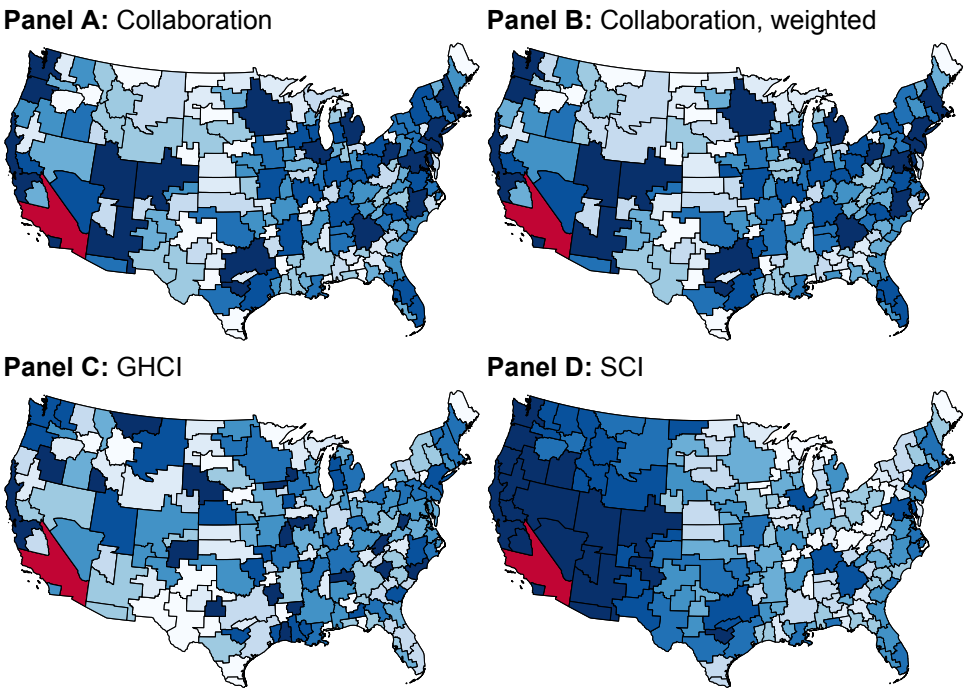
Note: Plots show the distribution of scaled GHCI and SCI regional connectedness indices. The horizontal red lines indicate medians of 133,753 for the GHCI and 3,518,538 for the SCI. The blue curves represent the Epanechnikov kernel density estimates. Both indices are scaled between 1 and 1,000,000,000.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

Figure A.13: Spatial decay



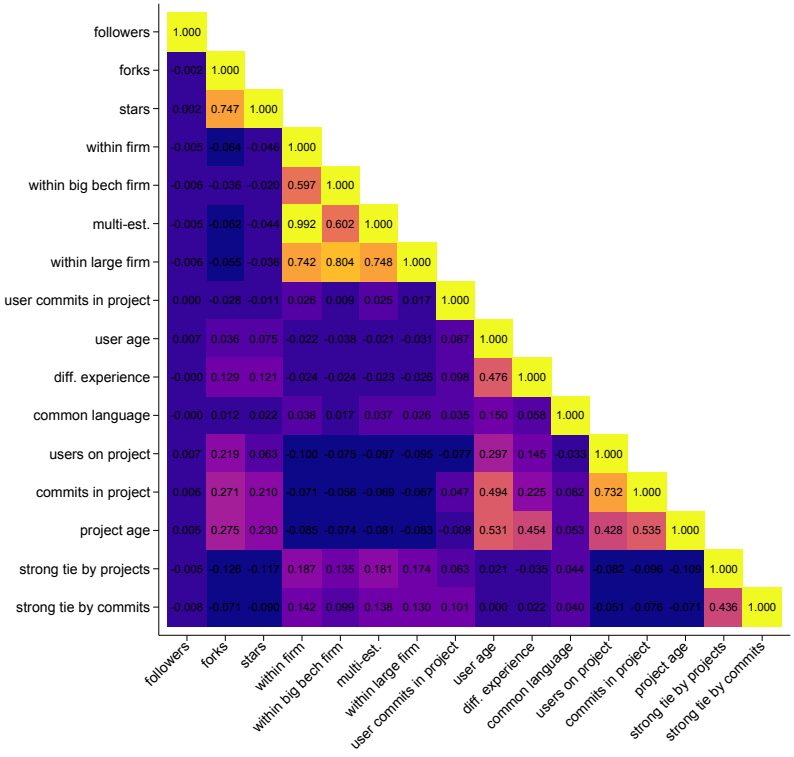
Note: Plot shows spatial decay as predicted per fractional polynomial model with (Panels A and B) and without (Panels C and D) the colocation effect and in values (Panels A and C) and logarithmically (Panels B and D). *Sources:* GHTorrent, Bailey et al. (2018b), Bureau of Economic Analysis, own calculations.

Figure A.14: Data example for Los Angeles-Long Beach-Riverside, CA



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

Figure A.15: 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.