Goldbeck, Moritz

Research Report

Essays in the economics of digital transformation

ifo Beiträge zur Wirtschaftsforschung, No. 106

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

Suggested Citation: Goldbeck, Moritz (2024) : Essays in the economics of digital transformation, ifo Beiträge zur Wirtschaftsforschung, No. 106, ISBN 978-3-95942-134-8, ifo Institut - Leibniz-Institut für Wirtschaftsforschung an der Universität München, München

This Version is available at:
https://hdl.handle.net/10419/300804

ifo
BEITRÄGE
zur Wirtschaftsforschung

106
2024

Essays in the Economics of Digital Transformation

Moritz Goldbeck

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

ifo
BEITRÄGE
zur Wirtschaftsforschung

106
2024

Essays in the Economics of Digital Transformation

Moritz Goldbeck

Herausgeber der Reihe: Clemens Fuest
Schriftleitung: Chang Woon Nam

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

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie; detaillierte bibliografische
Daten sind im Internet über
https://dnb.d-nb.de
abrufbar.

ISBN Nr. 978-3-95942-134-8

Druck: Pinsker Druck und Medien GmbH, Mainburg

ifo Institut im Internet:
https://www.ifo.de

Essays in the Economics of Digital Transformation

Inaugural-Dissertation
zur Erlangung des Grades
Doctor oeconomiae publicae (Dr. oec. publ.)

eingereicht an der
Ludwig-Maximilians-Universität München
2024

vorgelegt von

Moritz Goldbeck

Referent: Prof. Dr. Oliver Falck
Korreferentin: Prof. Dr. Lisandra Flach
Promotionsabschlussberatung: 10. Juli 2024

Datum der mündlichen Prüfung:	01.07.2024
Namen der Berichterstatter:	Prof. Dr. Oliver Falck
	Prof. Dr. Lisandra Flach
	Prof. Claudia Steinwender, Ph.D.

Preface

The digital transformation profoundly impacts our economies. Advances in information and communication technologies (ICT) that were deemed impossible just decades ago keep bringing down the costs of collecting, storing, and transmitting information exponentially (see, e.g., Leiserson et al., 2020; Mack, 2011). The impact and breadth of technological progress is immense and not only spurred the invention of countless new products – some of which quickly evolved into entire industries – but also revolutionized the way our economies are organized (Greenstein, 2015; Bresnahan et al., 2002). The macroeconomic importance of technology for continued economic progress is hard to underestimate, and the extent to which new knowledge and information contribute to the creation of economic and societal welfare has never been as high as today.[1] At the same time, despite unprecedented efforts to innovate and develop new technologies, knowledge creation and productivity growth are slowing (see, e.g., Goldin et al., 2024; Park et al., 2023; Bloom et al., 2020; Gordon, 2017). It is therefore paramount to study and understand the knowledge economy (see, e.g., Drucker, 1969; Mokyr, 2002; Powell and Snellman, 2004; Antràs et al., 2006) as a distinct yet interrelated part of the economy in order to unlock the full potential of the digital transformation to the benefit of our societies (Goldfarb and Tucker, 2019).

The rise of the knowledge economy and the digital transformation are inextricably linked to each other and, at the same time, embedded in the wider economy where they simultaneously affect production, consumption, and labor markets while being intertwined with key economic developments. A fundamental principle of this complex transformation is that ICTs enable the globalization of economic activity by reducing the costs of coordination and information transmission over distance (see, e.g., Helpman, 2009; Freund and Weinhold, 2004). This relaxes geographic frictions that have constrained economic activity to a particular location for most of history. Consequently, the digital transformation is closely linked to the geographic reorganization of the world economy that shapes the information age (Forman et al., 2018). In previous decades, digital technologies facilitated the emergence of complex global production networks by fostering coordination between suppliers, firms, and their customers (Baldwin,

[1] In line with canonical macroeconomic models (Aghion and Howitt, 2008; Romer, 1990; Arrow, 1962), ample empirical evidence shows that the share of economic growth attributable to technological progress increases (see, e.g., Helpman, 2009; Baily, 2002; Hall and Jones, 1999), and that improvements in ICT play a key role (see, e.g., Draca et al., 2007; Lipsey et al., 2005; Baily and Lawrence, 2001; Jorgenson and Stiroh, 1999).

2017). This triggered tremendous economic gains from exploiting comparative advantage and fueled the concentration of the knowledge economy in advanced economies. As a result, goods production is geographically highly fragmented, while knowledge work remains largely localized and regions fiercely compete to attract global talent to work for local firms (Glennon, 2024; Kerr, 2020).

This might change with the current wave of technology-induced globalization. For a long time, leading scholars have predicted the 'death of distance' in knowledge work enabled by ICT (e.g., Cairncross, 1997; Baldwin, 2017; Baldwin and Dingel, 2022), but evidence supporting this hypothesis remains scant (Glaeser and Ponzetto, 2010). In recent years, however, digital technologies have developed further, and prominent new digital products have emerged that specifically target the needs of white-collar office workers. For example, cloud-based software enables knowledge worker teams to streamline and synchronize computer-based projects online in real-time, such that anyone with an internet connection can access materials from anywhere in the world (Choudhury, 2020; Choudhury et al., 2021). Digital technology not only provides knowledge workers with online access to project material but also offers sufficiently close substitutes for face-to-face meetings through tools for online virtual interaction, if used appropriately (Karl et al., 2022; Samuel, 2015). A variety of digital products and services for both asynchronous – e.g., chat rooms, forums, e-mail, or project feeds – and synchronous – e.g., virtual audio and video meetings – communication is available and widely adopted by knowledge workers. Additionally, office closures during the pandemic acted as a catalyst for the widespread uptake and acceptance of digital tools for virtual interaction, which further increased remote collaboration and teleworking (OECD, 2021).

These and related technological and cultural developments significantly decrease the barriers to remote work and collaboration, making it more likely than ever that geographic distance becomes less important in knowledge work. First empirical evidence already points in this direction (Emanuel et al., 2023; Chen et al., 2022). Still, it is essential to develop a thorough understanding of whether and how digital technologies affect the geographic organization of knowledge work since these developments will not only impact productivity and economic growth but also, e.g., regional development, labor markets, office workplaces, and cities. Additionally, with continuous progress at the technological frontier – think, for example, of real-time capable team assistants based on artificial intelligence (Dell'Acqua et al., 2023) or on-the-fly audio language translation (Megasis Network, 2023) – it is already foreseeable that geographic barriers for collaboration in knowledge work will continue to fade (cf. Baldwin and Dingel, 2022). Thus, we should learn from observing and studying the early-stage and limited

application of such technologies in the knowledge economy today, as findings might provide valuable insights for the broader roll-out and potential future developments. Producing knowledge on these phenomena allows us to design effective, evidence-based policies that shape these developments to the benefit of our economies and societies.

* * *

In this dissertation, I empirically study economic impacts of digital transformation in four essays. Within the context of digital transformation, all essays are linked through their connection to the shift of labor markets towards knowledge work and how this affects economic geography or vice versa. In the *first essay*, I show local information and communication technology infrastructure, the basic precondition for connectivity, enables places to grow their economies. Importantly, growth effects are associated with significant structural change towards manufacturing employment in connected regions. The *second* and *third essay* explore the role of geographic distance for knowledge worker collaboration and differentiate between colocation, distance, and border effects while considering links to social connectedness and culture. Finally, the *fourth essay* emphasizes the role of online platforms in allowing knowledge workers to signal skill in the labor market from anywhere. Although united by the common theme briefly outlined here, each essay addresses a specific research question in digital economics and features a unique setting. Therefore this dissertation is structured into four self-contained chapters that can be read independently.

The *first chapter* explores the effects of internet infrastructure provision on local economic growth. Existing literature shows that broadband internet infrastructure, the basic precondition to participate in the digital economy, fosters economic growth in countries with a workforce that is highly educated (e.g., Czernich et al., 2011; Akerman et al., 2015). In this chapter, together with Valentin Lindlacher, I show this finding extends to remote areas in developing countries with a large informal sector and an low-skilled workforce. Further, already low-speed internet connectivity accessed predominantly in cybercafés spurred a significant increase in local economic activity. The accompanying structural transformation of labor markets in connected regions towards manufacturing employment suggests digital technology facilitates economic growth by enabling higher-productivity work in the formal sector. These results document that internet availability, as the basic precondition for digital transformation, is economically beneficial in various economic contexts. The digital transformation critically depends on the presence or absence of physical infrastructure and, therefore, is inherently intertwined with geography. As general-purpose technology and a

fundamental building block for other technologies, internet availability shapes local labor markets and fosters economic growth.

In the *second chapter*, I shift focus from the basic precondition for digitization to its frontier and investigate the capability of digital technology to bridge regions by enabling remote collaboration. In particular, I explore to what extent geographic distance still matters for collaboration in one of the most digitized settings of the knowledge economy: software development. I generally observe software developers to be highly concentrated in space and collaboration is about nine times higher for colocated compared to non-colocated developers. However, besides colocation, increased geographic distance is not significantly associated with less collaboration, unlike in less digital social networks. Furthermore, despite colocation being associated with higher collaboration rates, the colocation effect is much smaller than for inventors or in the social network. Especially collaboration within large organizations and weak ties are more distributed in space. These descriptive findings suggest the relevance of geographic distance and colocation is subdued in a digital work setting. As a result, digital tools potentially integrate the digital economy's workforce geographically and allow knowledge workers to contribute and participate in projects from anywhere, although still not to the same extent as colocated workers.

The *third chapter* continues to study knowledge worker collaboration across space and, in particular, revolves around international collaboration in software development. It addresses the question of whether cross-border collaboration of knowledge workers is subject to a border effect, i.e., a reduction of collaboration across international borders. Border effects are widely studied in economics and have been found sizable for various economic outcomes, most prominently goods trade (e.g., Santamaría et al., 2023b; McCallum, 1995). For digital knowledge work like software development, however, typical drivers of the border effect such as transport costs do not apply, and hence, their absence could further facilitate globalization in the knowledge economy. Data on open-source software developer collaborations reveals a sizable border effect that is, however, about five to six times smaller as in trade. Further results in this chapter, based on joint work with Lena Abou El-Komboz, demonstrate that the remaining border effect is entirely explained by cultural factors such as a shared language or overlapping interests. These findings are in line with the results from the second chapter and point toward reduced barriers to collaboration in digital work environments. Specifically, the results emphasize the importance of cultural factors for international collaboration in digital knowledge work, where geographical and technological constraints are less prevalent.

Lastly, the *fourth chapter* focuses on knowledge worker labor markets and specifically investigates if software developers signal skill through activity on public online platforms. If valued by employers, skill signals on such platforms can be obtained independent of a developer's geographic location, in contrast to traditional signals such as formal degrees from prestigious schools. In this sense, online activity on public open-source software platforms is a less geographically discriminatory signal of skill and could potentially relax geographic frictions in knowledge worker labor markets. An important economic phenomenon in this setting is that such signaling activity generates significant positive externalities from open-source software production. Generating open-source software is an increasingly important and highly valuable public good (see, e.g., Korkmaz et al., 2024; Hoffmann et al., 2024). Together with Lena Abou El-Komboz, in this chapter I find that private labor market incentives in the form of software developers' career concerns indeed spur significant activity on an open-source software platform, although the activity is less targeted towards the community and more towards activity with high labor market value. This suggests digital technology enables developers to successfully participate in the labor market independent of their geographic location while, at the same time, generating societal value.

Overall, this dissertation documents inclusionary effects of the digital transformation. Digital technology connects regions, economies, and most importantly: people. While local connectivity is the precondition to participate in the benefits of the digital economy, internet-based digital tools facilitate communication and information flows so that knowledge workers can interact and collaborate more seamlessly. As a result, participation in the digital economy becomes easier for people located anywhere, making the economic landscape less geographically discriminating. At the same time, results show that geographic frictions have, by far, not vanished yet – even in highly digitized settings. Naturally, this dissertation does not aim to offer a holistic assessment of the net economic and societal welfare effects of digital transformation. Other research regards negative effects of the digital transformation and points out mitigation strategies. Here, I explicitly focus on the potential of digital transformation to bring along economic benefits in the knowledge economy. In this spirit, this dissertation highlights relevant settings in the digital economy that emphasize economic benefits of digital transformation through technology that connects people across space.

* * *

On a meta-note, digital transformation impacts not only our economies but also the way they are studied. I therefore want to highlight two game-changing benefits of digital transformation

relevant to economic research, including this dissertation. First, digital technology opens unprecedented methodological possibilities for social science research (Angrist et al., 2017). Most importantly, computing power capable of analyzing large-scale datasets has become accessible to researchers (Lazer et al., 2020). This sparked the creation of software packages and empirical tools for data analysis and empirical research (Muenchen, 2012), bringing down the fixed costs of conducting empirical economic research with observational data. For example, this dissertation benefits from large-scale regression analysis implementations with high-dimensional fixed effects (e.g., Correia, 2019; Berge et al., 2023), efficient memory storage techniques (e.g., Morgan, 2022), and natural language processing algorithms (e.g., Wickham et al., 2024).

Second, digitization produces vast amounts of data, and although only a fraction of this data is accessible to researchers, this produces nearly endless and previously unheard-of opportunities to study societal and economic phenomena empirically (Varian, 2014). For example, in this dissertation I use calibrated data on nighttime luminosity captured by satellites (Li et al., 2020), allowing me to proxy economic activity at a high spatial resolution anywhere in the world. In other projects, data from a large online code repository platform provides access to detailed activity streams of new software developers worldwide (Gousios, 2013), making it possible to study the production process of software in great detail and observe knowledge workers' spatial and temporal activity patterns at granular scale. Moreover, extensive data on online behaviour collected by Obradovich et al. (2022) allows me to observe overlaps in shared interests between social groups along hundreds of thousands of dimensions, representing a novel bottom-up measure of cultural proximity with unprecedented granularity and depth.

As a researcher, I am amazed and at the same time humbled by these possibilities for science and evidence-based policy consulting enabled by digital transformation. In this spirit, I attempt to distill both new and relevant knowledge from available data using state-of-the-art empirical methods in this dissertation.

Keywords: digitization; software; knowledge work; culture; language; ICT; development; nighttime light; Africa; growth; cybercafé; geography; networks; knowledge economy; colocation; digital platforms; signaling; open source; job search

JEL-No: F66; H40; J24; J30; J61; L17; L84; L86; O18; O30; O31; O33; O36; R11; R32

Acknowledgments

This dissertation benefited immensely from diverse contributions by many people who created the intellectually inspiring and pleasant environment needed to thrive. Foremost, I sincerely thank my dissertation supervisor Oliver Falck; for his invaluable advice, genuine and unwavering encouragement throughout the past years, and for granting me the freedom to explore, fail, learn, and grow both as a researcher and personally. I am further grateful to Lisandra Flach for co-supervising this dissertation and to Claudia Steinwender for valuable suggestions and serving on my committee.

It was my pleasure to work on joint research projects with my awesome co-authors and colleagues Lena Abou El-Komboz and Valentin Lindlacher. I learned a great deal from working with them. Further, my colleagues at the ifo Center of Industrial Organization and New Technologies – Victor, Christina, Simon, Fabian, Nikola, Valentin, Mo, Anna, Thomas, and Sebastian – deserve special thanks for creating the open and inspiring team environment I greatly benefited from and for countless stimulating discussions. At ifo Institute, I had the privilege of advising and working with several bright and motivated students, and owe thanks to Raunak Mehrotra, Svenja Schwarz, and Gustav Pirich for infallible research assistance.

I sincerely thank Ricardo Hausmann for inviting me to his lab at Harvard University; Christina Langer, Ljubica Nedelkoska, and Muhammed Yıldırım for making this stay possible; the CESifo Young Ambassador Program for financial support; and the Growth Lab team for an unforgettable and inspiring experience. Moreover, I gratefully acknowledge public funding of my research and am especially thankful for the permanent funding of ifo Institute through the Leibniz Association. I also appreciate project-specific public funding through the German Research Foundation and the Bavarian Institute for Digital Transformation at the Bavarian Academy of Sciences. Colleagues at the ifo Center for International Economics introduced me to economic research during my time as a student research assistant, for which I want to explicitly thank Alexander Sandkamp, Erdal Yalçin, and Gabriel Felbermayr.

Without the support and resources available at ifo Institute, much of the empirical work in this dissertation would have been infeasible. I am especially grateful to Sebastian Wichert, Valentin Reich, and Heike Mittelmeier at the LMU-ifo Economics & Business Data Center as well as Kumar Subramani and Elke Sindram at ifo Institute's information technology department.

Lastly, I thank all colleagues at ifo Institute and the Munich Graduate School of Economics at the Department of Economics of Ludwig Maximilian University as well as countless seminar and conference participants who contributed to this dissertation, be it through administrative support or constructive and thoughtful comments.

March 2024

Moritz Goldbeck

Contents

List of Figures

List of Tables

1 Digital Infrastructure and Local Economic Growth: Early Internet in Sub-Saharan Africa

We study if low-speed internet availability fosters local economic growth in rural areas of developing countries by analyzing remote towns in Sub-Saharan Africa. We measure local economic growth of each town by tracking nighttime light emissions. In a difference-in-differences setting, we exploit exogenous countrywide shocks to internet availability induced by submarine cable arrivals in the 2000s and use the rollout of national inter-regional fiber cables to identify towns incidentally connected early. We find that internet availability induces economic growth. Compared to a control group of similar but later connected towns, connected towns experience 11 percent higher light intensity, which translates to 3.3 percentage points higher annual economic growth in the years after internet connection. Additional results suggest this is mainly driven by per-capita productivity growth and not by migration into connected towns. The effect is stronger in towns with better access to regional markets and internet availability is associated with a shift from agriculture to manufacturing in regional employment.[1]

Keywords: ICT; development; nighttime light; Africa; growth; cybercafé
JEL-No: O33; O18; R11

[1] This chapter is based on joint work with Valentin Lindlacher. We thank Vojtech Bartos, Mathias Bühler, Oliver Falck, Thomas Fackler, Jonas Hjort, Anna Kerkhof, Tobias Korn, Markus Ludwig, Niklas Potrafke, Helmut Rainer, David Roodman, Marcelo Sant'Anna, Claudia Steinwender, Maria Waldinger, and Kathrin Wernsdorf for valuable comments and suggestions as well as seminar participants at the ifo Instiute, University of Munich, TU Dresden, 14th RGS Doctoral Conference, the 3rd International Workshop on Market Studies and Spatial Economics, the 10th UEA European Meeting, the EBE CRC Summer School, the 9th ECINEQ Meeting, the UEA PhD Student Workshop, the ifo Dresden Workshop on Regional Economics, Economic Research South Africa Seminar, the VfS Annual Conference, the AFREN Doctoral Workshop, and the CESifo Area Conference on the Economics of Digitization. We thank Nicolas Göller for excellent research assistance.

1.1 Introduction

In the last decades, the provision of digital infrastructure enabled widespread access and adoption of the internet in most parts of the world. Evidence shows positive effects of broadband internet availability on individual-level economic performance (see, e.g., Akerman et al., 2015) and country-level economic growth (see, e.g., Czernich et al., 2011) for developed countries. Hopes are high that internet access fosters regional economic growth in the developing world as well (World Bank, 2016). For example, in Sub-Saharan Africa (SSA), where impulses for economic growth are needed urgently to fight poverty, governments, public-private partnerships, and private consortia alike invest large amounts of money in internet infrastructure projects. However, provision of internet access is complex and costly due to a lack of legacy infrastructure such as fixed-line telephony networks (see, e.g., Williams, 2010).[2] Until 2020, SSA countries invested more than 28 billion USD into their national internet backbone (Hamilton Research, 2020).[3] Despite these enormous digital infrastructure investments, a growth effect of internet in SSA is not assured. Low population density apart from the mega-cities, missing hardware, financial constraints, and a lower willingness to pay lead to low adoption rates (World Bank, 2016). At the same time, the potential of the internet seems particularly high in SSA since alternative ICT is largely absent (ITU, 2019). Given the large investment requirements and unclear economic benefits, it is crucial to understand how internet availability affects regional economic development in SSA, especially in rural areas where provision is particularly costly.

In this paper, we examine whether internet availability causes local economic growth in remote areas of Sub-Saharan Africa and, as a result, contributes to rural development. In contrast to the existing literature, we focus on the extensive margin of internet provision in a developing-country setting featuring low literacy rates and agrarian, labor-intensive economies. Specifically, we study remote towns during the initial introduction of the internet in SSA through the first wave of internet-enabled submarine cables from 1999 until the mid-2000s, enabling low-speed internet connectivity (0.5-2 Mbps). At the time, people accessed the internet predominantly in cybercafés, small community-based internet centers that provide local communities with internet access using minimal infrastructure (see, e.g., Southwood, 2022). We track economic activity at the town level in response to plausibly exogenous shocks in local internet availability. To assess potential mechanisms, we decompose growth of towns

[2] Ngari and Petrack (2019) estimates that laying down one kilometer of fiber-optic cable in SSA costs between USD 15,000 and 30,000.

[3] Facebook announced an effort to build a new internet-enabled submarine cable (SMC) to Africa for one billion USD in 2020 (Rascouet et al., 2020; Anderson and O'Connor, 2020). China plans to invest more than 60 billion USD in Africa's digital infrastructure as part of its 'Belt and Road' initiative (Invesco, 2019).

into spatial expansion (extensive margin) and density of economic activity (intensive margin) and interpret these components as pointing more towards population or productivity growth, respectively. We corroborate this analysis with an assessment of changes in local population density. In addition, we investigate changes in regional employment shares to study structural transformation associated with internet availability.

Our baseline sample captures the evolution of 210 remote towns in 10 SSA countries provided with (international) internet bandwidth between 1999 and the mid-2000s and a pre-existing national backbone outside larger cities. We tap two main data sources. First, we measure local economic growth, the key outcome of interest, using nighttime light (NTL) intensity captured by satellite, a well-established proxy introduced by Henderson et al. (2011) at the country level and validated by Storeygard (2016) on the city level for SSA. We compute yearly economic activity of each town by assigning NTLs to individual agglomerations via built-up areas from *Africapolis*. Second, we use data on the rollout of national internet infrastructure backbones from Hamilton Research (2020) to measure internet infrastructure availability in each town. The data comprises a comprehensive record of the locations of internet access points in SSA. Because data only starts in 2009, we conduct an extensive review of national backbone deployment projects to assign construction years to access points. This enables us to study the early- and mid-2000s when the first wave of sub-marine cable arrivals brought the internet to SSA for the first time at noticeable scale.

To identify the causal effect of internet availability on local economic growth, we exploit quasi-random variation in the timing of country-wide internet access induced by the arrival of the first wave of sub-marine cables (SMCs) in SSA. This approach was established by Hjort and Poulsen (2019), who exploit an internet speed upgrade induced by the second wave of SMCs with higher capacities. In a difference-in-differences framework, we additionally exploit the national backbone expansion to define comparable treatment and control towns. National backbone expansions aim to connect political and economic centers (Williams, 2010). Importantly, towns located on-route between such 'nodal cities' typically receive access points. We assign treatment status to towns that were connected to the national backbone when the internet became available country-wide, while the control group consists of similarly-sized towns getting internet connection only later. In a fixed effects model with town and country-year fixed effects, we then compare economic growth of towns with backbone access at the time when internet becomes available country-wide for the first time to a control group of similar towns getting access only later. Our key identifying assumption is that treatment and control group towns would have evolved similarly in the absence of treatment. This

assumption cannot be tested for, but we estimate a dynamic event-study specification of our model to show that there are no differences in pre-treatment trends of economic activity between treatment and control group towns.

We find that connection to the internet through an access point on average leads to a 11 percent increase in NTL emission of towns in rural SSA in the years after country-wide connection compared to a control group of similar towns not connected through an access point at that time. Applying the established light-to-GDP elasticity from Henderson et al. (2012), this translates into about 3.3 percentage points higher economic growth. We then decompose this overall effect into measures for intensive- and extensive-margin growth and find higher statistical significance for intensive-margin growth, suggesting an increase of per-capita productivity. Together with the fact that we do not find effects on population growth, this points towards economic development rather than a spatial redistribution of economic activity. Further, we find this effect accompanied by a shift in regional employment shares. In regions with connected towns, manufacturing employment shares increase by 1.3 percentage points relative to regions getting connected later. This is consistent with the literature on industry-bias of ICT towards high value added sectors (see, e.g., Baumol, 1967; Ngai and Pissarides, 2007). Heterogeneity analyses with respect to different measures of market access suggest stronger effects in towns better integrated into regional markets, in line with existing works that establish complementarity between ICT and trade (see, e.g., Baldwin, 2019; Steinwender, 2018).

To ensure that our results are indeed driven by internet availability, in addition to town and country-year fixed effects, we control for the rollout of mobile GSM coverage.[4] Our model further takes into account potential changes in the importance of geographic factors over time. Apart from absent pre-trends, placebo tests corroborate that the effect is tied to the unique structure of the exogenous variation we exploit. It is therefore unlikely that treatment is confounded by parallel infrastructure rollouts. Nevertheless, we assess this possibility more directly using georeferenced survey data on electricity availability and find no evidence in support of parallel expansion of electricity grids. We assess robustness of our results to alternative model specifications, in particular regarding the composition of the control group and measurement approaches. Finally, we estimate less demanding variations of our model to assess robustness on larger sample sizes and external validity.

This study makes three main contributions. First, our unique settings allows us study the

[4] During our observation period, all countries only had basic (GSM) mobile coverage which enables calls and SMS messaging, but not surfing the web. Importantly, 3G coverage, and therefore mobile internet, was unavailable.

causal effect of internet availability on local economic growth in a sample of remote towns in rural SSA. We are the first to show and quantify significant effects for rural SSA during the period when internet first became available in these areas, which feature labor-intensive, agrarian economies. We show that internet availability has an effect on economic growth beyond political and economic centers and thus contributes to rural development in developing countries. Second, while most studies are concerned with broadband internet, we focus on low-speed connectivity. As alternative means to access the internet in SSA were non-existent or prohibitively expensive, especially in rural areas, our shock truly measures the extensive margin of internet, from virtually no connectivity to speeds between 0.5 and 2 Mbps enabling basic functionality like e-mail and web browsing. Third, people in rural SSA predominantly access the internet via cybercafés during the 2000s before mobile internet spread to rural areas from 2010 onward. We contribute by examining growth effects of the internet in the context of these community-based and cost-efficient institutions, which are overlooked in the literature with its heavy focus on mobile internet.

The remainder of this paper is organized as follows. First, we discuss related literature in Section 1.2. Section 1.3 introduces the data. In Section 1.4, we present our empirical strategy. Results are provided in Section 1.5 and Section 1.6 concludes with a discussion.

1.2 Related literature

Internet and economic growth. We contribute to three main strands of literature. First, we add to the broad literature assessing the impact of the internet on economic growth. For developed countries, the effect of digital infrastructure and especially (broadband) internet has been assessed widely. For example, Czernich et al. (2011) identify an effect of broadband infrastructure on annual per-capita growth in OECD countries. Bertschek and Niebel (2016) find a firm-level productivity effect of mobile internet in Germany. For the US, Kolko (2012) finds a positive relationship between broadband expansion and a host of local economic outcomes such as population growth, employment, and wages. For developing countries, Hjort and Tian (2021) survey the evolving literature on internet and growth. Much of this literature is focused on mobile internet, as mobile phones are the main technology through which individuals access the internet in developing countries at least since 2010 (see, e.g., Rodríguez-Castelán et al., 2021; Williams et al., 2011; Aker and Mbiti, 2010). Several recent studies examined the effect of mobile internet availability in developing countries in the 2010s and quite consistently found an increase in consumption and a reduction in poverty, e.g., in Nigeria (Bahia et al., 2020), Senegal (Masaki et al., 2020), and Tanzania (Bahia et al., 2021).

Focusing on mobile internet *use*, Roessler et al. (2021) show smartphone use increased per-capita household consumption significantly. In contrast, Suri and Bhattacharya (2022) find no impact on a wide range of economic outcomes including employment and consumption in a RCT distributing free phone data in Kenya. Haftu (2019) observe an effect of mobile phones but not for internet availability on per-capita income at the country level. Similarly, Rotondi et al. (2020) find an effect of mobile phone coverage and ownership on rural development in developing countries. At the country level, Thompson and Garbacz (2011) finds stronger effects of mobile internet in low-income countries, but no effects of fixed-line broadband. Evidence on the channels through which economic outcomes in developing countries are affected by the internet remains scant. Generally, the broader literature suggests internet advances economic growth by reducing information frictions, improving the management of supplies, increasing the productive efficiency of firms, and reducing transportation costs (see, e.g., Aker, 2010; Hjort and Tian, 2021). In Brazil, Barbosa et al. (2021) find organizational firm restructuring and employment losses in response to broadband availability. Hjort and Poulsen (2019) study the employment effects of large increases of available international bandwidth around 2010 in SSA and find a skill-biased and net positive employment effect at the individual level. In a comment, however, Roodman (2024) questions the validity of these results and ascribes them to geocoding and measurement errors.

Our study is the first to causally investigate growth effects of early, low-speed connectivity when the internet first became available in rural SSA. This contrasts with the literature's focus on mobile internet after 2010, which leads to the previously prevalent institution of cybercafés being largely overlooked. Cybercafés are important institutions that introduced the internet to most individuals in SSA during the 2000s (Southwood, 2022). As cybercafés do not require individual-level hardware, they are extremely cost-efficient and serve entire local communities with minimal infrastructure. It is important to understand such community-based modes of technology access as well as their economic effects in more detail. Especially in remote areas or where legacy infrastructure is lacking, their scalability and cost-efficiency is a crucial feature to achieve widespread adoption quickly. This work emphasizes that internet infrastructure availability in a setting where cybercafés are the predominant access technology enhances local economic growth in remote areas of developing countries.

ICT and market integration. A growing literature investigates the effects of information and communication technologies on market integration. ICT facilitates the integration of markets by improving communication and information flows. Reserach shows that, by reducing information frictions, ICT enhances, e.g., the efficiency of labor markets (Autor et al., 2015) and

fosters trade (Leuven et al., 2021; Steinwender, 2018; Freund and Weinhold, 2004). Generally, ICT is found to exhibit a skill- and sector-bias and therefore affects industries and occupations differently (see, e.g., Michaels et al., 2014; Ngai and Pissarides, 2007; Autor et al., 2006; Baumol, 1967) and likely also has differential effects on trade (Grossman and Rossi-Hansberg, 2008). For developing countries, the literature on the role of ICT for market efficiency is still slim. Baldwin and Forslid (2023) argue digital technologies enable developing countries to pursue a services-led growth model by exploiting comparative advantages in previously untradable sectors. In their seminal paper, Jensen (2007) shows how price dispersion drops in response to mobile phone adoption in rural India around 2000. Aker (2010) confirms this effect of the introduction of mobile phones on prices between 2001 and 2006 in Niger. For Chinese firms, Fernandes et al. (2019) observe increased exporting in response to internet availability.

This paper shows even remote towns in rural SSA benefit from internet availability. We add to this literature not only with our focus on remote areas in developing countries but also by explicitly analyzing low-speed, community-based internet connectivity. With the notable exception of Jensen (2007), the literature neglects the important era when ICT technologies became first available in the developing world. In line with existing literature studying developed economies, our analyses suggest a complementary between (regional) trade and ICT even in a setting with agrarian and labor-intensive economies and low literacy rates.

Regional development, geography, and infrastructure. There is a large body of related literature on the effect of infrastructure provision on regional development. Infrastructure provision is typically much less profitable and at the same time more expensive in rural areas (see, e.g., Chaurey and Le, 2022). There is an established literature for developing countries for non-digital infrastructure, most importantly transportation (see e.g., Asher and Novosad, 2020; Banerjee et al., 2020; Aggarwal, 2018; Donaldson, 2018; Jedwab et al., 2017; Ghani et al., 2016; Storeygard, 2016; Faber, 2014) and electricity (see e.g., Lee et al., 2020; Burlig and Preonas, 2016; Chakravorty et al., 2014; Grogan and Sadanand, 2013; Rud, 2012; Dinkelman, 2011). Although not in all settings, this literature largely finds infrastructure beneficial for regional development. For digital infrastructure, the literature on regional development predominantly considers developed countries (see, e.g., Briglauer et al., 2019). Although the regional digital divide is discussed widely (see, e.g., Lagakos, 2020; Fukui et al., 2019; Buys et al., 2009), only few studies investigate settlements outside of the large cities in more rural and remote areas (e.g., Hjort and Poulsen, 2019). Rotondi et al. (2020) acknowledge the potential of mobile phones for rural development in poor countries. A more active strand of literature assesses regional inequality in developing countries as with rapid urbanization

(OECD, 2020) rural areas fall behind economically. Economic productivity is typically higher in urban areas for several reasons including thick labor markets, knowledge spillovers, and low transportation costs (see, e.g., Curiel et al., 2017; Albouy, 2016; Clark et al., 2002; Deller et al., 2008). While studies mostly compare the economic progress in mega-cities versus secondary cities, with inconclusive findings regarding inequality trends (e.g., Bluhm and Krause, 2022; Christiaensen and Kanbur, 2017; Fetzer et al., 2016; Christiaensen and Todo, 2014), studies on rural agglomerations are lacking. Notably, Henderson et al. (2012) indicates that the hinterland grows faster than coastal areas and primate cities do not grow faster than their hinterland.

Our work contributes by showing that connectivity effectively contributes to narrowing the digital divide in remote towns in rural areas of developing countries. Although we cannot speak to the relative development with respect to secondary and primate cities, we observe unconnected remote towns falling further behind compared to their incidentally connected counterparts. We further corroborate findings in existing works of positive effects of ICT infrastructure in rural SSA and show that individual-level effects sum up to significant aggregate effects on economic growth at the local level of towns.

1.3 Data

To assess the impact of internet availability on local economic growth, we combine data on economic growth and internet infrastructure at the level of towns in Sub-Saharan African (SSA) countries.[5]

1.3.1 Local economic growth

For SSA countries, comprehensive sub-national or even city-level records of economic activity is lacking, especially panel data is unavailable. Therefore, we use night-time light (NTL) emissions as a proxy for economic activity. NTL data is available worldwide from 1992 until 2013 from the *U.S. Air Force Defense Meteorological Satellite Program*'s *Operational Linescan System* (DMSP-OLS). The instruments of DMSP-OLS satellites measure light intensity on an integer scale from 0 to 63 with pixels covering 30 arc-second grid cells, an area of .86 square kilometers at the equator. In most years, at least two satellites are deployed to capture NTL; DMSP-OLS data averages measurements and reports yearly composites. The

[5] We define Sub-Saharan Africa as the mainland of the African continent without the Northern African countries, Algeria, Egypt, Libya, Morocco, Tunisia, and Western Sahara. We exclude South Africa as economically more developed country due to lack of comparability.

remote sensing community acknowledges the usefulness of NTL data to measure economic activity (see, e.g., Levin et al., 2020; Levin and Duke, 2012), but emphasizes the importance to correct DMSP-OLS composites for various sources of measurement error such as saturation (Ma et al., 2014) and atmospheric light (Määttä and Lessmann, 2019; Wei et al., 2014). Recently, shortcomings of the raw data like the lack of calibration are increasingly recognized in economics (see, e.g., Roodman, 2024; Gibson et al., 2021). We use the harmonized version of the yearly DMSP-OLS composites from Li et al. (2020), who extract only light emitted by human settlements by excluding night lights from aurora, fires, gas flares, boats, and other temporal lights unrelated to human settlements and make the data temporally consistent via an exhaustive inter-calibration procedure.

NTL data is an established proxy for local economic growth (see, e.g., Asher et al., 2021; Bluhm and McCord, 2022), especially where official statistics are lacking or unreliable (Donaldson and Storeygard, 2016; Nordhaus and Chen, 2015; Chen and Nordhaus, 2011). In general, NTL emission by human settlements represents mostly outdoor use of light typically associated with human consumption or production activities, which is, in turn, closely related to income and GDP (Levin et al., 2020). However, this relationship is complex, indirect, and noisy; and by using it we abstract from many issues such as public versus private light emissions, tracing specific sources of light, or classifying light emission of settlements into consumption versus investment activities. Yet, there is an empirically well-established relationship between NTL and economic growth. In the economic literature, Henderson et al. (2012) demonstrate the (linear) relationship between GDP growth and NTL growth at the country level and subsequent studies (see, e.g., Määttä et al., 2022; Storeygard, 2016; Chen and Nordhaus, 2011) validate that this also holds at the sub-national, grid, or city level. Bluhm and McCord (2022) find NTL data more suited to capture changes in GDP at lower baseline levels of GDP and population densities, and Mellander et al. (2015) shows NTLs tend to slightly overestimate economic growth in large urban areas and underestimate growth in rural areas. Other concerns regarding NTL data like blurring and top-coding are concentrated in cities and metropolitan areas (see, e.g., Gibson et al., 2021; Bluhm and Krause, 2022). NTL data therefore is especially well-suited for our analysis, targeting mid-sized towns in remote areas of SSA.

The key advantage of NTL data is its geographic specificity. To measure local economic growth at the town level, we map NTL data to human settlements using built-up areas from *Africapolis* (OECD, 2020).[6] This database contains the geographical delineation of 7,496 SSA towns and cities with more than 10,000 inhabitants in 2015. By integrating small towns into the data and

[6] *Africapolis:* `https://africapolis.org`, accessed on 01/05/2023.

combining satellite imagery with various census and administrative sources, *Africapolis* data is the first to provide comprehensive geographic information on the agglomeration landscape in SSA. The median size of an *Africapolis* agglomeration in 2015 is about 21,000 inhabitants and around 90% of towns feature less than 100,000 inhabitants. In 2000, agglomerations were considerably smaller with a median population of about 10,000, and about 90% of agglomerations inhabited by less than 45,000 people.

1.3.2 Internet infrastructure

We measure internet availability across time at the town level by combining two data sources. Our first source is *Africa Bandwidth Maps*, a database maintained by *Hamilton Research* and sourced directly from network operators.[7] The database contains a comprehensive record of internet access points and their locations on the African continent and covers the period from 2009 until today, updated yearly. The data represents a detailed record of national fiber-optic internet infrastructure rollout in SSA. Before construction of such cables, internet access in SSA was extremely limited and prohibitively expensive (see, e.g., LeBlanc and Shrum, 2017; Williams, 2010; Gitta and Ikoja-Odongo, 2003).[8] Consequently, national internet backbone access constituted the first viable and affordable way to go online for the vast majority of SSA people, especially in rural areas (see, e.g., Kitimbo, 2023; LeBlanc and Shrum, 2017).

An internet access point is a node in the (usually fiber-optic) backbone of a nation's internet network. From the access points, internet users in the surrounding area are reached by local (wired or wireless) 'last mile' infrastructure. Although geodata on fiber cable lines is available and often used (e.g. in Hjort and Poulsen, 2019), drawing on the exact locations of access points on these lines is a superior measure as local networks branch out from there and have a limited reach (Roodman, 2024). At the time, users in rural Sub-Saharan Africa predominantly accessed the internet via cybercafés (see, e.g., Williams et al., 2012; Southwood, 2022). Cybercafés (or: internet cafés) are community-based centers with wired internet access typically in the form of small shops or rooms with computers (LeBlanc and Shrum, 2017). Internet access was offered at pre-paid hourly rates in cybercafés (Southwood, 2022). In the 2000s, cybercafés usually were the only way to access the internet in rural SSA (Williams et al., 2012), and people not only used cybercafés for communication and entertainment but also for professional purposes such as maintaining business contacts and managing the delivery of goods and supplies (see, e.g. Gitta and Ikoja-Odongo, 2003; Mbarika et al., 2004). We provide

[7] *Africa Bandwidth Maps:* `http://www.africabandwidthmaps.com`, accessed on 04/11/2023.
[8] Technologies used prior to national backbone access were either satellite- (e.g., VSAT) or telephony-based via narrowband dial-up modems (Williams, 2010; Nyezi, 2012).

further background on last-mile transmission technologies and cybercafés in Sub-Saharan Africa in the 2000s in Section A.1.2 in the Appendix.

We leverage the *Africa Bandwidth Maps* data on access points to measure local internet infrastructure availability. Using the geolocation information, we compute the geographic distance between each *Africapolis* town and the nearest access point to define towns as within-reach when being located within a distance of 10 kilometers to an access point.[9] Note that this measure of internet infrastructure availability does not ensure local adoption at the town level as we do not directly observe the presence of cybercafés nor other means of local end-user uptake. Therefore, similar to other studies exploiting local internet infrastructure availability, our results are best interpreted as intention-to-treat effect (ITT). The ITT effect is typically of particular interest when estimating aggregate effects as it takes into account adoption rates. In addition, ITT effects in this institutional setting might be particularly strong. The predominant access mode through cybercafés at the time did not require individual hardware adoption and nearby presence of a cybercafé is highly likely in locations with internet availability (Williams et al., 2012). Therefore, cybercafés have the potential to serve entire local communities with internet access efficiently (Southwood, 2022).

We infer the construction date of access points from the first year they show up in the data. For access points already present in the first data year, 2009, we conduct an extensive review of internet backbone deployment projects for each SSA country to determine their construction date, going back until the late 1990s. Although it is not always possible to determine the exact year of construction, we are able to determine which access points were constructed until the year the countrywide internet connection was established, which is sufficient information for our analysis.[10] Figure A.9 maps of all 2,708 access points and their construction year.[11] We provide a brief overview of each countries' national backbone expansion in Table A.2, and Section A.1.1 details a country example as well as further background information on national backbone rollouts in SSA.

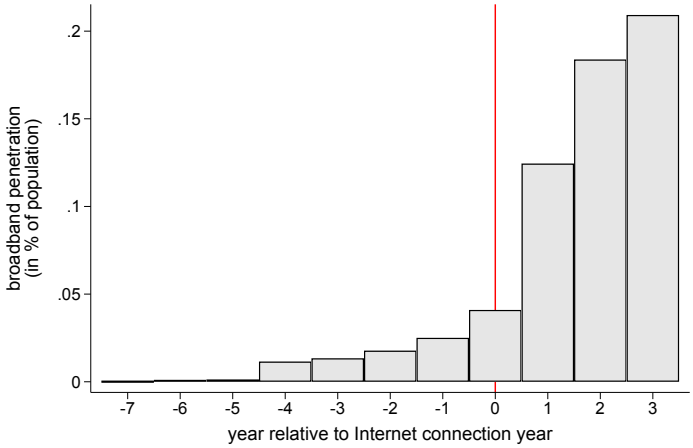
Our second data source is the *Submarine Cable Map* by *TeleGeography*, a comprehensive

[9] According to the literature (see, e.g. Ngari and Petrack, 2019) as well as interviews with industry experts, this is an appropriate (average) distance. Robustness checks with alternative distance cut-offs supports this information (Table A.20 and Table A.19).

[10] Documentation of our review of deployment projects, including a source register, is provided in Table A.23 in Section A.4.

[11] About half of them were constructed after 2013 and larger cities are typically served by more than one access point, usually for bandwidth reasons. This implies that, for example, in 2019, although 189 new access points were constructed, only 27 cities and towns were first connected. In total, around 900 *Africapolis* cities and towns are within-reach of an access point in the most recent year of our data (2020).

Figure 1.1: SMC connection and internet adoption



Note: International capacity is calculated from SMC capacities assigned to each country by using population-weighted shares. Adoption rates are calculated relative to the establishing year of internet connection in each country and then aggregated taking the weighted mean. Weights are population size in 2000. *Sources:* Submarine Cable Map, ITU, WDI.

collection of information on global submarine cables.[12] Submarine cables are fiber-optic cables for large-scale international data transmission over long distances and form the backbone of the international internet infrastructure. SMC construction typically is a joint effort of governments, private investors, and/or multinational organizations (Williams, 2010). The arrival of the first wave of internet-enabled SMCs in SSA countries from 1999 until the mid-2000s first brought internet connection to Sub-Saharan Africa at noticeable scale. The largest SMC from the initial wave is SAT-3 which started operating in 2001 and featured landing points in nine West African countries.[13] Before SMC arrival, the number of SSA internet users was tiny, with only 0.2 million users in 1998, while in 2002 already 3.2 million people used the internet (Southwood, 2022). With the first wave of SMCs, international bandwidth constraints that previously kept prices high relaxed considerably. Figure 1.1 shows that internet adoption rates jump in SSA countries after SMC arrival, although still remaining at relatively low levels.[14] SMCs

[12] *Submarine Cable Map:* `https://www.submarinecablemap.com`, accessed on 04/11/2023.

[13] SSA countries connected by SAT-3 are Angola, Benin, Cameroon, Côte d'Ivoire, Gabon, Ghana, Nigeria, Senegal, and South Africa. The cable originates in Sesimbra, Portugal, and Chipiona, Spain, and routes via the Canary Islands in Alta Vista. Building costs for SAT-3 are estimated at USD 650 million (Southwood, 2022).

[14] See Table A.1 for country-specific connection years. Before the first SMCs landed on SSA shores, the only way to connect to the internet on the continent was via satellite or telephony cables. Telephony cables are unavailable in the vast majority of SSA, especially in rural areas. While being largely unconstrained by geography and local infrastructure, satellite connection is costly and allows only for narrow bandwidths. South Africa, which we do not study here, was connected in 1993 through an internet-enabled SMC (SAT-2) that preceded an old co-axial telephone cable from 1968 (SAT-1).

of the first wave provided capacities for internet at basic speeds, i.e., connections featuring around 0.5 to 2 Mbps (Hjort and Poulsen, 2019; Agyeman, 2007). Between 2009 and 2012, these SMCs were proceeded by the next generation of SMCs with much higher capacities enabling higher-speed internet connectivity.[15] Landings of SMCs are often described as transformative moments for SSA countries (see, e.g., Graham et al., 2015).

For our empirical analysis, we use the date on which first-wave submarine cables connecting SSA counties start operating, the so-called *ready-for-service* (RFS) date as well as information on the exact landing point in each SSA country from the *Submarine Cable Map*. The RFS year of the first SMC in a country marks the year in which international internet connection was established. Connection to the international internet network is crucial for SSA countries since, especially at the time under study, the vast majority of web pages and applications used in SSA are hosted on servers located in North America or Europe, and thus almost all African internet traffic is routed inter-continentally (Kende and Rose, 2015; Chavula et al., 2015).[16] We geolocate the landing points and relate each of them to an *Africapolis* agglomeration. For countries that established international internet connection through a neighboring country (mostly landlocked countries), the date at which a border access point was established marks the connection year.

We exploit RFS dates as differences in the timing of SMC arrival introduce quasi-random and country-wide variation in internet availability. Hjort and Poulsen (2019) introduced this shock in the economic literature. Three features of this setting come together that are important for the identification strategy in this paper. First, the need of SSA internet traffic to be routed intercontinentally. Second, the fact that each SSA country has a single national backbone network with roughly equal (technically feasible) speed irrespective of the distance to the SMC landing point. This implies that each SSA country has a specific and country-wide treatment date – the year of SMC arrival. Third, the order in which SSA countries are reached by SMCs is geographically determined. This generates quasi-random variation in the timing of internet availability across SSA countries.

1.3.3 Supplementary data

To take into account simultaneous expansion of other digital infrastructure, we draw on mobile coverage data from *Collins Bartholomew*. Their *Mobile Coverage Maps* provide information on the availability of mobile signal and differentiate between the cellular technologies GSM

[15] Country-specific years of this 'speed upgrade' are reported in Table A.1.
[16] This is true even for 'local' content like websites of SSA businesses and organizations as hosting infrastructure such as data centers within SSA is lacking, especially at the time we study.

(2G), UMTS (3G), and LTE (4G). During our observation period, GSM (2G) mobile signal became available in SSA countries and none of the countries in our sample started rolling out internet-enabled UMTS technology. From the yearly shape files provided in the data, we compute, for each town in our sample, the share of its built-up and 2 kilometer buffer area covered with GSM signal in each year. Typically, this town areas are either fully covered or no signal is available, i.e., the resulting value is either 0 or 1. While not enabling mobile internet, GSM signal implies the availability of basic communication functionalities such as making calls or sending short text messages.

We further tap time-varying geographic data on local population density from *Gridded Population of the World* (GPW) provided by the *NASA Socioeconomic Data and Applications Center* (SEDAC). GPW data models the distribution of human population counts and densities on a continuous global raster surface. This data offers the same spacial resolution as the DMPS-OLS NTL data (30 arc-second grid cells), but comes only in a time resolution of five-year intervals. We proxy town-level population similarly to economic activity by aggregating pixels within buffered built-up areas and applying the natural logarithm.

Data on employment by industry originates from census data in the *IPUMS International* database.[17] We aggregate this household-level data to the sub-national regional level (Admin-2) and caluclate employment shares by industry, i.e., agriculture, manufacturing, and services. Censuses are carried out roughly every ten years and at different points in time for different countries. For details on census years by SSA country, see Table A.8 in the Appendix.

We obtain additional geographic information from various sources. From *OpenStreetMap* (OSM), we source information on the status as national or regional capital and link it to *Africapolis* towns.[18] To assign the status as economic center to a town, we use population information in the year 2000 from *Africapolis*. Furthermore, we use OSM to collect the location of financial, health, and educational infrastructure, as well as rivers. We obtain information on other transportation infrastructure from *Natural Earth Data* (roads and railroads) and the *World Port Index* (shipping ports).[19] *Africapolis* provides information on each town's altitude and population density. In addition, we source data on terrain ruggedness in 30 arc-second resolution from Nunn and Puga (2012).[20]

[17] *IPUMS International:* `https://international.ipums.org/international/`, accessed on 04/12/2023.
[18] *OpenStreetMap*: `https://www.openstreetmap.org/`.
[19] *Natural Earth Data*: `https://www.naturalearthdata.com/`, accessed on 04/12/2023; *World Port Index*: `https://msi.nga.mil/Publications/WPI`, accessed on 04/12/2023.
[20] Nunn and Puga (2012) data: `https://diegopuga.org/data/rugged`, accessed on 04/12/2023.

1.3.4 Combining the data

We are interested in the development of remote towns in SSA countries in response to an exogenous shock in internet availability. To this end, we track NTL emissions of each town over time by assigning DMSP-OLS NTL pixels to *Africapolis* towns and measure internet availability in each of these towns via access points from *Africa Bandwidth Maps* and SMC arrival dates from the *Submarine Cable Map*. As we focus on incidentally connected remote towns, 'nodal cities' – national and regional capitals as well as economic centers – are excluded. Specifically, we define economic centers as cities with more than 50,000 inhabitants in 2015 according to *Africapolis*.[21]

Our subjects of interest are remote towns in Sub-Saharan Africa. During our observation period, by far not all remote towns receive internet access points. We thus define our comparison group using two criteria. First, we select remote towns, i.e., non-nodal cities, for which an internet access point becomes available until the end of our data period in 2020. Second, we include only towns that remain unconnected until the end of our five-year post period, so that there are no compositional changes in treatment and control group during the observation period, which would confound our analysis. As a result, there is a trade-off between the length of our observation period and comparability of treated and control towns. To make sure our specification is appropriate, we show parallel trends and robustness to changes in the specification with respect to observation period definition (see Section 1.5.2).

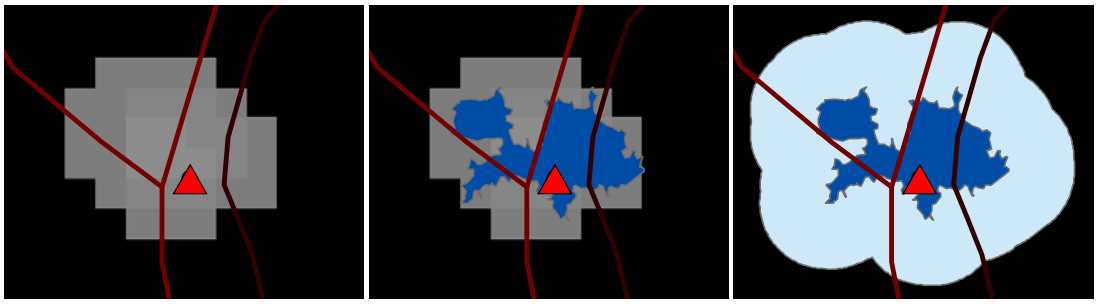
NTLs are the best available measure to track economic growth of remote towns in SSA countries for two main reasons. First, NTLs provide the necessary geographic resolution to measure local economic growth of each town. Second, remote towns lie far enough away from each other to clearly separate lights emitted by nearby towns. Panel (a) of Figure 1.2 shows Dassa-Zoumè in Benin in 2004, a typical town for our sample with around 19,000 inhabitants in 2000, according to *Africapolis* estimates. The contiguous area of gray pixels represent NTL emissions of Dassa-Zoumè and can clearly be attributed to the town, with lighter gray pixels indicating stronger light emissions. Roads leading through Dassa-Zoumè are depicted as red lines and railroads in dark red.

We require a town to emit NTL in each year of observation avoid measurement error due to background noise in the data (Chen and Nordhaus, 2011). This ensures that the data captured represents an appropriate proxy for economic growth at the town level, but comes at the expense of losing the smallest towns. With this measurement method, we are able to trace

[21] Robustness tests with respect to this choice are presented in Table A.16 and Table A.17.

the economic growth of *Africapolis* towns with on average around 16,000 inhabitants in 2000 and a distribution ranging from 10,000 to 50,000 excluding nodal cities. Figure A.13 displays the density distribution for towns in our sample. An additional advantage of the stable light emission requirement is that included towns likely have electricity connection over the whole observation period (Falchetta et al., 2020; Dugoua et al., 2018), precluding electricity grid expansion as a confounding factor in our analysis. Nevertheless, we perform robustness analyses with respect to this requirement in Table A.11 and Table A.12.

Figure 1.2: Data example Dassa-Zoumè, Benin (2004)



Note: Panels (a) through (c) show a data example for Dassa-Zoumè, Benin, in 2004. Panel (a) shows NTL emissions for the year 2004, three years after the SMC connection year of Benin. Light intensity is shown by lighter grays. The red triangle indicates an internet access point is present in 2004 (built in 2001, in this case). Red linkes represent major roads and dark red lines railways. Panel (b) additionally shows Dassa-Zoumè's built-up area in dark blue. Panel (c) adds a 2 kilometer buffer around the built-up area in light blue. *Sources:* Li et al. (2020), Africapolis, Open Street Map, Natural Earth Data, Africa Bandwidth Map.

DMSP-OLS NTL emitted by human settlements blurs out to adjacent pixels, so NTL extend beyond towns actual geographic expansion, measured by their respective *Africapolis* built-up areas. Panel (b) of Figure 1.2 shows this for the town Dassa-Zoumè in Benin in 2004. The NTLs (gray) extend out of the towns' built-up area (blue). This phenomenon is known as 'blurring' or 'overglow' (Abrahams et al., 2018). We account for NTL blurring by extending the built-up area by a buffer area of 2 kilometers in order to capture all NTLs emitted by a town. As illustrated for Dassa-Zoumè by Panel (c) of Figure 1.2, this allows us to include all relevant NTL pixels.[22]

For each town-year, we measure NTL emissions by summing up the light intensities of pixels within a town's area as defined above. This method of local NTL aggregation was proposed and validated by Storeygard (2016) and accounts for both increased light intensity and geographical expansion. Changes in NTL emissions over time are a measure of economic growth as shown in Henderson et al. (2012) and Storeygard (2016). Specifically, Henderson

[22] For robustness, we also show the results for a specification without a buffer as well.

et al. (2012) observe a stable linear relationship between changes in NTL and GDP growth both in a worldwide sample of countries and for low- and middle-income countries in particular, with an estimated light-to-GDP elasticity of around 0.28. This implies that a 10% increase in NTL from one year to the next translates to a 2.8% increase in GDP year-on-year.

In addition to this composite NTL measure, we derive two other measures from NTL. First, we compute the average light intensity of all pixels in a town's area as an indication for per-capita GDP growth (intensive margin). Second, we calculate the sum of all lit pixels in a town's area as a measure of population growth through spatial expansion (extensive margin). Although noisy and imperfect, these measures provide suggestive evidence on the underlying source of economic growth. As an alternative to the NTL-based measure of intensive-margin growth, we separately analyze changes in population via high-resolution grids from the *GPW* database.

Lastly, for each town in each year we compute if there is an internet access point within-reach, i.e., within 10 kilometer distance. According to available information (see, e.g., Ngari and Petrack, 2019) and interviews with telecoms experts, this is an appropriate average reach of technologies at access points used at the time. If a town is located less than 10 kilometers away from an access point, we record a town as connected to the national backbone. In Dassa-Zoumè, for example, there is an access point within-reach in 2004 (built in 2001), marked by red triangles in Figure 1.2. Together with the information of the date when a SMC first arrives in the respective country in which a town is located, we know when internet first became available in each town.

1.3.5 Descriptive statistics

Our analysis is focused on mid-sized, remote towns. With our measurement technique, we identify 510 agglomerations in 10 SSA countries emitting NTLs each year. Thereof, 70 agglomerations (13.7%) are classified as nodal cities and 118 towns (23.1%) are still unconnected to an internet access point at the end of our data period in 2020. 112 towns (21.9%) received access to the national backbone during the five-year post-period after SMC arrival and are excluded in our main specification as their treatment confounds the control group. Thus, our main sample contains 210 towns in 10 SSA countries, where there are both treated and control towns, with yearly NTL emission in the observation period and eventually receiving access to the national backbone in their country. This represents 41% of agglomerations detected via NTL and 18.9% of all *Africapolis* towns in the studied countries.[23]

[23] The *Africapolis* data records a total of 1,113 agglomerations with less than 50,000 inhabitants in 2000 in countries with both treated and control towns.

In our sample, 97 towns (46.2%) were already connected to the national backbone via an access point prior to country connection via SMC or a neighboring country and therefore form the treatment group. The remaining 113 towns constitute the control group and receive access to the national backbone, too, but after the five-year post period. Table A.3 reports summary statistics for our sample.

Our identification builds on a comparison of remote towns receiving connection prior to SMC arrival due to their location on-route between nodal cities, and remote towns connected to the national backbone only later. On average, treated towns have 16,595 and control towns 16,314 inhabitants. Treated and control group towns are not only almost identical in their average population but also in their population distribution (Figure A.13). In addition, we show our comparison captures similar towns by analyzing the expansion of national backbones that connects more cities and towns over time. Panel (a) of Figure A.2 plots the average population size in each year relative to the country connection year for towns in our sample as well as nodal cities. Nodal cities connected earlier are much larger and average population size declines quickly at first and more slowly after about five years post-connection. This shows that national backbone expansions prioritize larger nodal cities. Panel (b) of Figure A.2 focuses on treated and control towns and shows that there is no clear association of population size and connection timing for control towns. Average population size of control towns lies between 11,000 and 19,000, with no clear time trend relative to the country connection years. This points to the absence of selection into treatment and supports the notion of incidental connection of on-route towns.

1.4 Empirical strategy

Internet availability is not randomly assigned to locations. Our identification strategy aims to break the correlation between internet availability and unobserved determinants of local economic growth by exploiting two sources of exogenous variation: the staggered rollout of, first, the national internet infrastructure and, second, international sub-marine internet cables. This generates quasi-random spatial and temporal variation in internet availability conditional on town and country-year fixed effects as well as geography controls.

Our baseline fixed-effects panel data regression model to estimate the relationship between

internet availability and local economic growth is a difference-in-differences specification:

$$y_{ic(i)t} = \beta_0 + \beta_1(\text{connection}_{c(i)t} \times \text{access}_i) + \beta_2 \text{GSM}_{it} + {}_3(\mathbf{X}_i' \times \text{connection}_{c(i)t}) + \alpha_{ic(i)} + \alpha_{c(i)t} + \varepsilon_{ic(i)t}, \tag{1.1}$$

where $y_{c(i)t}$ is economic growth of town i in country $c(i)$ in calendar year t as proxied by nighttime light (NTL) intensity. Internet is available in town i in calendar year t if two conditions hold simultaneously: the country has a sub-marine cable connection and the town has access to the national backbone. The variable $\text{connection}_{c(i)t}$ indicates if *country* $c(i)$ has internet connection in calendar year t via a sub-marine cable, and access_i is an indicator if *town* i has internet connection, defined as being located within 10 kilometers (geodesic) distance to an access point at the time of SMC arrival in country $c(i)$. Consequently, the interaction term $\text{connection}_{c(i)t} \times \text{access}_i$ indicates internet availability in town i in country $c(i)$ in calendar year t.[24] The coefficient of interest is β_1 and captures the effect of internet availability on local economic growth.

This specification mimics a hypothetical situation where internet availability is randomly assigned to towns. The model essentially compares 'treated' towns that are connected to the national backbone at the time of SMC arrival to other ('control') towns that receive connection to the national backbone at a later point in time. We argue that this exploits two types of exogenous variation. First, we use exogenous variation in internet availability at the country level from the quasi-randomness in the timing of SMC arrival. SMCs arriving in SSA countries at the time under study come from Europe and typically feature one landing point in each SSA country they passed. Thus, SMC arrival time is mainly geographically determined (Hjort and Poulsen, 2019). Together with separate national backbones in each country, this generates temporal variation in country-wide internet availability: at the ready-for-service date, internet becomes available in all locations within a country that are connected to the terrestrial backbone network.

The second source of exogenous variation in internet availability comes from the rollout of national backbones, during which remote towns typically receive an access point only when they lie on the route between nodal cities. The routes between nodal cities are built at different speeds due to geographic, political, or other reasons related to the nodal cities. Importantly, backbone expansion planning typically does not consider on-route towns due to their insignificant population size compared to nodal cities (see, e.g., Williams et al., 2011).

[24] To not confound our control group, we do not consider towns getting an access point in the post period as control towns in our main specification.

As a consequence, some remote towns exogenously benefit from their location on the route between nodal cities that are connected before SMC arrival. Note that the comparison group are other remote towns that often lie on route between nodal cities, too, but are connected later. Thus, the staggered nature of national backbone rollouts creates spatial variation in internet availability at the time of SMC arrival for remote towns in SSA. We discuss a typical country example in detail in Section A.1.1.

To factor out further confounding factors, we include two types of fixed effects as well as additional controls. Time-constant differences across towns are captured by town fixed effects $\alpha_{c(i)}$. Differences across calendar years common to all towns within a country are absorbed by country-year fixed effects $\alpha_{c(i)t}$. Note that this allows for country-specific time trends such as differential growth rates and also captures variation in satellite sensor quality over years. In addition, we account for mobile internet network expansion by using spatial coverage of each town with GSM signal, GSM_{it}. Lastly, we include a set of geography controls $\mathbf{X}_i$ interacted with the connection indicator $\text{connection}_{c(i)t}$ to allow for time-variation in the effect of geographic factors related to town-level growth. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. We use robust standard errors clustered at the level of access points to account for serial correlation in the error term $\varepsilon_{ic(i)t}$.

The key identifying assumption for β_1 is that treated towns would have evolved similarly to control towns in absence of treatment, i.e., if internet had not become available. The same underlying trends assumption cannot be tested. Its plausibility can, however, be examined by investigating pre-treatment differences in time trends between the treatment and the control group. It is a necessary, although not sufficient, but testable condition for same underlying trends that there are no trend differences between treatment and control group before the treatment. To this end, we conduct an event study and analyze the dynamic impact of internet availability on local economic activity by running the regression

$$\begin{aligned} y_{ic(i)t} = \mu_0 + \sum_{j(c(i))=\underline{T}}^{\overline{T}} \left[\mu_{1j} \left(t_{j(c(i))} \times \text{access}_i \right) \right] + \mu_2 \text{GSM}_{it} + {}_3(\mathbf{X}_i' \times \text{connection}_{c(i)t}) \\ + \delta_{ic(i)} + \delta_{c(i)t} + e_{ic(i)t}, \end{aligned} \tag{1.2}$$

where $t_{j(c(i))}$ indicates the year relative to treatment year, i.e., the year when internet became available in country $c(i)$, starting in relative year $j(c(i)) = \underline{T}$ and ending with relative year $j(c(i)) = \overline{T}$. The treatment year is normalized to $j(c(i)) = 0$. We omit $j(c(i)) = -1$ as the reference point. The interaction $t_{j(c(i))} \times \text{access}_i$ indicates if town i in country $c(i)$ is part of the

treatment group and restricts the coefficient to relative year $j(c(i))$. Thus, the coefficients μ_{1j} capture the dynamic effect – i.e., the effect for each relative year – of internet availability on local economic growth.

We further assume that there is no other time-varying within-country variation net of controls that correlates with the interaction of SMC arrival and backbone access and affects local economic growth independently of internet availability. There are three main threats to identification: measurement error, omitted variables, and model misspecification. We discuss all of these in Section 1.5.2.

1.5 Results

We use the difference-in-differences model in Equation 1.1 to estimate the average treatment effect on the treated (ATT) of Internet availability on local economic growth at the town level. The regression results are presented in Table 1.1. In line with our expectations, we find a positive relationship between Internet availability and local economic growth. Models (1) to (3) show a statistically highly significant effect of Internet availability on the standard light intensity composite measure – the logarithmic sum of light intensities of a towns' pixels. We translate these effects into GDP growth effects by using the elasticity between changes in night time light and GDP growth from Henderson et al. (2012) of $\epsilon_{\text{GDP, light}} = 0.283$. The resulting GDP growth effects are reported in the last row of Table 1.1 and are economically significant in size. The effect from our preferred specification in model (3) corresponds to a 3.26 percentage point higher GDP growth in connected towns in the five years after SMC connection relative to control towns connected later.

The time-varying control for GSM mobile coverage is only weakly statistically significant but still economically sizable yet smaller than the main Internet effect. Its inclusion leads to more precise estimation of the main effect, which increases slightly. As mobile Internet is the main alternative form of Internet infrastructure in rural Sub-Saharan Africa at the time, this suggests that the Internet access points and complementary last-mile infrastructure are in fact driving the main effect and not by simultaneous expansion of mobile coverage in treated towns. We discuss the role of mobile coverage in more detail in Section 1.5.2.

Increasing model flexibility by including geography controls interacted with an indicator for the post-connection period in model (3) improves model fit and reduces size and precision in the estimates of Internet access effects. This specification allows the effects of geographic factors such as distance to transport infrastructure or markets to vary over time. In fact, recent

literature suggests market access and travel times have become less important over the last decades in developing countries (see, e.g., Henderson and Kriticos, 2018; Brülhart et al., 2020). There is also evidence that ICT contributes to decreasing importance of geography as it improves communication with and thereby increases integration into larger markets (Steinwender, 2018). Model (3) shows that the main effect is not driven by changes in the economic benefits from geographical factors common to all towns.

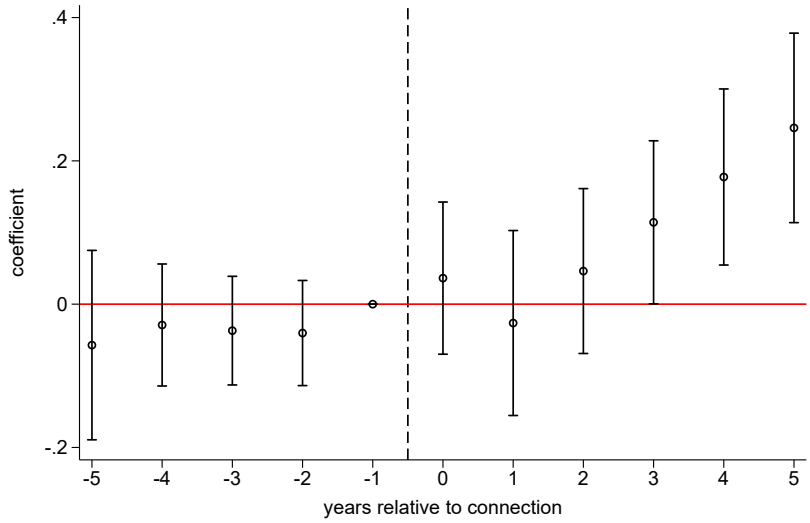
To assess the plausibility of the same underlying trend assumption as well as the dynamics of the effect, we plot the estimated event study coefficients μ_{1j} from the regression in Equation 1.2 in Figure 1.3. We omit the year before SMC arrival as reference point. There are no differences in pre-trends between connected and unconnected towns before SMC arrival, depicted by insignificant estimates close to zero for all pre-treatment years. About two years after SMC arrival the trends diverge and connected towns start to grow substantially faster compared to unconnected towns, conditional on controls. From the third post-treatment year onwards these dynamic estimates are significant. We exploit a shock in Internet availability and therefore it is expected that there is a lag until an economic effect materializes as adoption or behavioral adjustments take time. Our dynamic results suggest a sustained growth advantage due to internet availability in connected towns up to five years post treatment, the end of our observation period, but do not speak to the persistence of the growth advantage beyond this period.

Table 1.1: The effect of internet availability on local economic growth

	NTL growth			NTL growth margin	
	(1) composite	(2) composite	(3) composite	(4) intensive	(5) extensive
Connection × access	0.129***	0.134***	0.109***	0.0769***	0.0817**
	(0.0427)	(0.0433)	(0.0383)	(0.0237)	(0.0330)
Town FE	×	×	×	×	×
Country × year FE	×	×	×	×	×
GSM coverage		×	×	×	×
Geography controls × connection			×	×	×
Observations	2,310	2,310	2,310	2,310	2,310
Countries	10	10	10	10	10
Towns	210	210	210	210	210
Share treated	*.462*	*.462*	*.462*	*.462*	*.462*
Adjusted R^2	0.936	0.936	0.948	0.923	0.919
Economic growth effect	3.90	4.06	3.26	—	—

Notes: NTL intensity in models (1) to (3) is measured as the logarithmic sum of light intensities. The corresponding economic growth effect in percentage points is calculated as $[[\exp(\hat{\beta}_{\text{connection}\times\text{access}}) - 1] * \epsilon_{\text{light, growth}}] * 100$ using the elasticity $\epsilon_{\text{light, growth}} = 0.283$ from Henderson et al. (2012). The intensive margin in model (4) is measured by the logarithmic mean light intensity and for the extensive margin in model (5) as logarithmic sum of lit, i.e., non-zero, pixels, all on the same area. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Figure 1.3: Dynamic effect of internet availability on local economic growth



Note: The figure plots event study coefficients μ_{1j} based on Equation 1.2. The outcome is the logarithmic sum of light intensities. Bars represent 95% confidence intervals using robust standard errors clustered at the level of the closest access point. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

1.5.1 Mechanism

We investigate potential mechanisms behind the main effect in three ways. First, we decompose NTL into proxies for intensive- and extensive-margin growth. Second, we consider migration via changes in high-resolution population grids. And third, we explore effect heterogeneity with respect to market access, transport infrastructure, and sectoral employment.

Growth margin. Our composite NTL measure includes nightlight emissions as a result of both geographic expansion due to more lit pixels ('extensive' growth margin) and increased light intensity of previously lit pixels ('intensive' growth margin). Both channels are suggestive of different sources of growth. An increasing number of lit pixels points more towards potentially increased population, especially as rural towns in SSA typically do not accommodate population by increased inhabitants per square kilometer but through geographic expansion (Sakketa, 2023). In contrast, increased light intensity suggests growing economic activity. We distinguish these channels by estimating separate models for the number of lit pixels and average light intensity in models (4) and (5) in Table 1.1. Results show both channels play a role, but the intensive growth margin plays a more important role in terms of statistical

significance.

As the extensive margin measured via NTL data might be confounded by blurring intensive-margin pixels, we study the extensive margin more explicitly using high-resolution population grids. For each town, we compute population estimates from the *Gridded Population of the World* data available every five years and use its logarithmic values as outcome variable in our baseline specification. Table A.15 reports the results for different sample specifications. We find insignificant but mostly slightly positive point estimates, although the sign is not stable in all specifications. We interpret these results as pointing to a subordinate role of migration to connected towns, i.e. the extensive growth margin, consistent with the NTL-based finding of a more pronounced intensive growth effect.

Market access. Market access has been identified as key complement to ICT (Freund and Weinhold, 2004). We therefore assess heterogeneity with respect to multiple market access measures in Table 1.2. First, we estimate the impact of a standard deviation increase in distance to the next port on the treatment effect by a triple interaction on SMC connection, internet availability, and (standardized) distance to port. The estimate suggests a statistically weakly significant negative effect of 6.6 percentage points reduced economic growth when distance to port increases by a standard deviation (290 km). Second, we calculate a market access measure following Baragwanath et al. (2021) from weighted geographic distances to a country's population as

$$\mathrm{MA}_i = \sum_{i \neq j} \frac{\mathrm{pop}_i}{(\mathrm{dist}_{i,j})^2}, \tag{1.3}$$

for each town j and settlements in the country i using the 2015 *Africapolis* location and population data. We exclude town j when calculating this measure (Donaldson and Hornbeck, 2016). Relative to the other measures used, this metric gives more weight to local and regional markets and less to distant but larger metropolitan areas. A standard deviation increase in this market access measure yields a 3.7 percentage point higher growth effect that is statistically more precisely estimated. As third proxy for market access, we use landlocked status on the country level and find a large but statistically only marginally significant heterogeneity. The point estimate suggests that the effect on towns in landlocked countries on average is only one quarter the size compared to towns in coastal countries. Together, the results on market access lack statistical power but point towards market access as a key complement to improved connectivity, in line with existing literature (see, e.g., Steinwender, 2018). Our findings suggests that the growth effect is present particularly in towns with local and regional market access, although international market access seems important too, e.g. for landlocked

Table 1.2: Internet availability and market access

	(1)	(2)	(3)
Connection × access	0.110***	0.101***	0.205***
	(0.0378)	(0.0367)	(0.0721)
Connection × access × distance port	-0.0667*		
	(0.0400)		
Connection × access × market access		0.0369**	
		(0.0175)	
Connection × access × landlocked			-0.145*
			(0.0807)
Town FE	×	×	×
Country × year FE	×	×	×
GSM coverage	×	×	×
Geography controls × connection	×	×	×
Observations	2,310	2,310	2,310
Countries	10	10	10
Towns	210	210	210
Share treated	*.462*	*.462*	*.462*
Adjusted R^2	0.943	0.942	0.942

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

countries.

Transport infrastructure potentially affects economic growth (Boopen, 2006). As market access seems important for the growth effect of connectivity, other infrastructure is potentially complementary. We investigate heterogeneity with respect to road and railroad access in two ways. First, we estimate triple interaction models with distance to roads and railroads in Table A.6. Results show insignificant estimates implying no different effect for connected towns that are closer to infrastructure. Second, we vary our sample and include only towns

with access to roads or railroads, finding similarly-sized and statistically indistinguishable effects. These results do not support a high relevance of transport infrastructure for harnessing the growth effects of connectivity. However, it is important to acknowledge limited statistical power due to the vast majority of sample towns being located alongside roads.[25] The significantly higher point estimate for towns with railroad access, where there is more variation in our sample, is suggestive of some relevance of transportation infrastructure.

Structural change. ICT typically impacts sectors differently and is more complementary to services and manufacturing than agriculture (Acemoglu and Autor, 2011). We use individual-level employment from repeated cross-sectional surveys to investigate if internet availability is associated with different patterns of structural transformation. For five SSA countries, there is a survey for both before and after SMC arrival.[26] Data is geolocated at sub-national regional level. Therefore, we switch to the regional level for this analysis and define treated regions as regions with at least one access point during the observation period. Figure A.16 plots regional employment shares by sector and treatment status. Agricultural employment dominates with over two thirds of respondents, followed by services and manufacturing employment.

Regression results of our baseline model with industry shares as outcome show that regions with internet availability experience a shift in employment shares different to regions with no internet availability. Specifically, regions with internet availability at the time of SMC arrival feature an about 1.3 percentage point higher share of manufacturing workers in the survey after countrywide SMC connection. Given the spatial and temporal coarseness of the available data and the large informal sector, the marginal statistical significance of this finding is expected. While no economically and statistically meaningful effect is detected for service employment, there is an economically significant reduction in agricultural employment, although statistically insignificant. Overall, these results suggest a slightly faster structural transformation of regional economies towards manufacturing employment in connected regions. With manufacturing employment only at 11% on average, a 1.3 percentage-point increase reflects a sizable employment-based growth of 12% of the manufacturing sector.

1.5.2 Robustness

Measurement. Measurement is a key challenge in our setting (cf. Section 1.3). Therefore we conduct a battery of robustness checks with respect to the measurement choices implicit in our preferred specification. Importantly, we vary our choice regarding the buffer around built-

[25] Note that this is a direct result of our empirical strategy focusing on on-route towns and a reassuring property of the sample.
[26] Table A.8 reports survey years for available countries.

Table 1.3: Internet availability and sectoral employment

Sector:	(1) agriculture	(2) manufacturing	(3) services
Connection × access	-0.0194	0.0129*	0.00642
	(0.0163)	(0.0074)	(0.0107)
Region FE	×	×	×
Country × year FE	×	×	×
GSM coverage	×	×	×
Observations	956,454	956,454	956,454
Countries	5	5	5
Regions	99	99	99
Share treated	*.208*	*.208*	*.208*
Adjusted R^2	0.128	0.039	0.100

Notes: Employment shares are measured at the region level. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, IPUMS International, Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

up areas (Table A.10), the population threshold for nodal cities (Table A.16 and Table A.17), and the required distance to an access point (Table A.20 and Table A.19). Furthermore, we re-estimate our baseline model using different specifications to measure the intensive margin growth effect (Table A.7) and on a larger sample, relaxing our requirement for town-level NTL data every year (Table A.11 and Table A.12). All robustness checks are extensively discussed in the dedicated Section A.1.3 in the Appendix. Our checks demonstrate the robustness of the results with respect to measurement choices.

Omitted variables. Factors affecting the outcome and correlated with treatment are a potential threat to identification. In our context, parallel infrastructure rollout is a potential concern. Other infrastructure that boosts local economic activity and is built in treatment but not control towns at the time of SMC arrival confounds our estimates. Except for mobile internet infrastructure, time-varying local infrastructure data are unfortunately unavailable. Therefore, we resort to alternative ways to assess robustness for possibly growth-enhancing infrastructure other than mobile connectivity.

Mobile connectivity. Our main specification already accounts for changing connectivity due to improved mobile signal. Generally, fiber infrastructure improves mobile signal as well, but at the time most cell towers in rural SSA are too far from the fiber network and relied on satellite or microwave transmission technology (Ngari and Petrack, 2019). In Table A.22, we additionally account for the possibility of time lags before improved mobile connectivity affects economic activity. We achieve this by introducing lagged mobile GSM coverage to the model. Results show that the main effect remains robust in all lag specifications. The strongest effect of mobile coverage on economic growth is estimated with a lag of one year. Afterward, the point estimate shrinks and loses statistical significance.

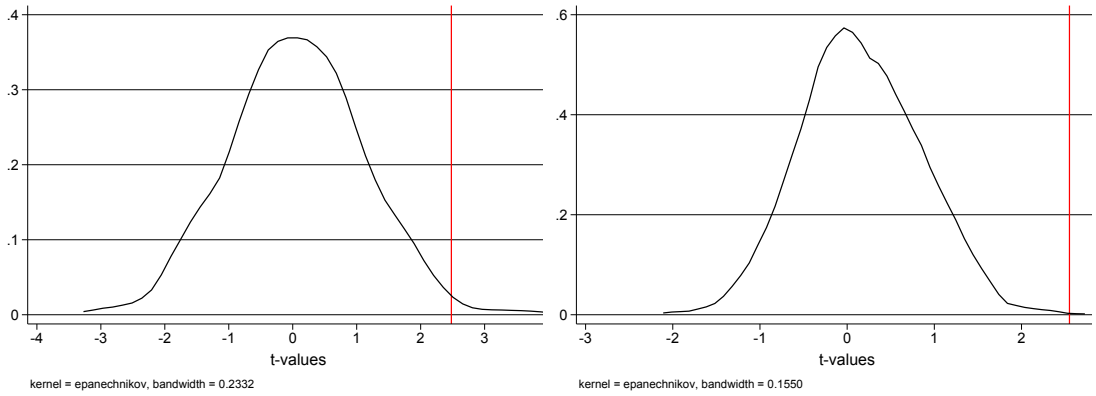
Electricity. Electricity is often found growth-enhancing in developing countries (see, e.g., Best and Burke, 2018; Rud, 2012). Consequently, simultaneous rollout of the electricity grid in treated but not control towns might be a thread to isolate the effect of internet availability. Their stable NTL emission of towns in our sample suggests electricity availability in the whole period (Falchetta et al., 2020; Dugoua et al., 2018). Nevertheless, to empirically test for spatial and temporal simultaneity, we draw on georeferenced survey data from *Afrobarometer* (BenYishay et al., 2017).[27] From the repeated cross-sections, we select data from the first four rounds of the survey between 1999 and 2009.[28] We aggregate household-level electricity availability to the town level and estimate our baseline model with town-level electricity availability either weighted and unweighted by sample size. The resulting samples are small both in terms of towns and countries. We therefore relax other sample restrictions. The specification and data are discussed in detail in Section A.1.3. Results provide no indication for an overlap in the expansion of electricity grid and internet backbone. Additionally, we estimate a triple interaction model with distance to the electricity grid in column (3) of Table A.6 to assess effect heterogeneity with respect to electricity access and find an insignificant on growth.

Placebo tests. Identification concerns regarding simultaneous expansion of other infrastructure are warranted only if they affect economic growth in treated but not control towns at the same time as a SMC arrives in a country. This means that simultaneous infrastructure rollouts nationally for internet and other infrastructure alone, for which we find no evidence, does not threaten our empirical design. The growth effect of simultaneously rolled-out infrastructure additionally would have to be systematically related to SMC arrival, which we consider highly unlikely. To assess empirically to what extent the captured effect is indeed specifically related

[27] *Afrobarometer:* `https://afrobarometer.org/`, accessed 07/12/2022.
[28] We restrict the data to country-years to the time before the major SMC upgrades.

to our empirical design we conduct two types of placebo exercises relating to the exogenous variation from national backbone rollout and SMC arrival.

Figure 1.4: Access and connection placebos



(a) access placebo **(b)** connection placebo

Note: The figure depicts the estimated kernel density function for the t-test statistics of the main effect for 1,000 permutations of our baseline specification with randomly assigned treatment years. *Sources:* Africa Bandwidth Maps, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

For the first placebo, we randomly assign treatment status to towns while maintaining each country's connection date. We then follow Chetty et al. (2009) and re-estimate our preferred specification on 1,000 permutations. Panel (a) of Figure 1.4 plots the density distribution of the resulting t-statistics. The vertical red line indicates the t-statistic of the estimate from our preferred (true) specification. The t-statistics of the randomly assigned hypothetical access samples are normally distributed and center around zero. Only in 18 of 1,000 permutations of internet access (1.8%) a higher t-statistic than in the true sample is observed. Similarly, we conduct a second placebo exercise randomly allocating the country connection years. Again, the distribution of t-statistics for 1,000 permutations plotted in panel (b) of Figure 1.4 is normally distributed, centering around zero. Only 1 out of 1,000 permutations (0.1%) yields a higher t-statistic than our true estimate. This implies the effect we find is statistically highly specific to both the exact timing of countrywide SMC arrival and town-level internet access at the time of SMC-arrival. Alternative growth-enhancing shocks are therefore unlikely to drive our effect if they do not exhibit a very similar structure both temporally and spatially.

Model specification. We assess robustness regarding model specification in various ways. Importantly, our empirical design considers a selected sample of treated and control group towns following a conservative approach focused on clean identification. The difference-in-differences setting generally allows for different outcome levels and relies on the same

underlying trends assumption. We already established that treated towns are somewhat larger; see Figure A.13. Nevertheless, common event study pre-trends point towards a robust design.

Sample balance. Still, a potential concern is that initially connected towns differ in terms of an economically more favorable location. The exogenous shock is at the country level and in Figure A.1 we point out that the timing of the countrywide internet connection is associated with countries' rollout progress. As the rollout of national internet networks is not random, we test whether observable time-invariant geographic controls correlate with treatment status in the cross section, given country fixed effects. If treatment status cannot be predicted from the controls, this adds additional credibility to our identification as it implies a like-for-like comparison. Figure A.5 show results of cross-sectional balance tests with respect to initial internet access. Internet access point rollout typically follows existing (transportation) infrastructure. Unsurprisingly, we therefore find a negative correlation between initial access and distance to capital cities, roads, and railroads. Our preferred specification includes all of these geographic factors interacted with the connection indicator to allow for changes in their importance for economic growth over time. There is no statistically significant correlation with other observables such as geographical characteristics and points of interest like educational or health infrastructure (Figure A.5 and Figure A.4).[29] We conduct a similar balance test at the country level using (weighted) averages of the same observables and their relation to connection year. We report the largely similar results in Figure A.6.

Control group. Designing a control group from towns getting an access point only after the post-treatment observation period ensures access points established near or in control towns during the post-treatment period do not contaminate the control group. However, this design also leads to a gap in the connection years between treated and control towns. In Table A.21 we re-run our baseline model not allowing late connected towns in the control group to have access points after certain calendar years (columns (1) to (4)) and with different post-SMC cutoff years (columns (5) to (7)). Although this significantly impacts sample size, the effect remains relatively stable and significant. Going into the other direction, relaxing this restriction further by allowing also late-connected or (to date) untreated towns in the control group increases sample size. Results are robust to these changes, too, and are reported in columns (2) and (3) of Table A.10.

[29] An exception are colleges, which show a marginally significant association with treatment status. At the same time, other educational infrastructure such as universities and schools are insignificantly related to treatment status.

Further specifications. National backbone expansion typically focus on major trade routes from the landing point in or near the capital to the second-largest economic center such that the treatment group often includes towns on these routes. Our baseline specification controls for town fixed effects as well as changes in importance of geography over time. At the same time, heterogeneity analyses points to market access as important amplifier of the growth effect of connectivity. In column (3) of Table A.9 we exclude towns on a countries' main trade corridor to address concerns the effect is purely driven by a selected group of favorably located towns. Despite significant reduction in sample size, the effect remains stable. We provide more detail and report further robustness tests with respect to model specification in Section A.1.3 in the Appendix. These include econometric model choice like standard error clustering, effect stability regarding countries, and additional industry heterogeneity results.

1.6 Conclusion

Digital infrastructure is a key precondition for locations to harvest digital dividends from internet connectivity. In rural areas of SSA, infrastructure provision is particularly costly due to remoteness and low population density. At the same time, due to differences in the structure of rural economies it is unclear if such locations are able to reap similarly high benefits from connectivity and therefore if closing the digital divide simultaneously narrows the economic gap between rural areas and economic centers. In this study, we exploit the unique setting when internet first became available in SSA with the arrival of sub-marine cables during the 2000s. We show that even low-speed internet predominantly accessed in community based internet centers, cybercafés, significantly improves economic development of remote towns in rural SSA.

In particular, we study the arrival of the first sub-marine internet cables in ten SSA countries in the 2000s, which first brought international bandwidth and therefore internet connectivity to SSA countries. We assess the causal effect of internet availability on local economic growth using a difference-in-difference setup that additionally makes use of the rollout of national backbone infrastructure to design appropriate treatment and control groups. Our quasi-experimental comparison relies on incidentally connected towns on-route between economic centers connected by national internet backbones at the time of country-wide internet arrival. Together with plausibly exogenous variation in the timing of SMC arrivals, this allows us to causally estimate the effect of internet availability by comparing initially connected on-route towns to a control group of similar towns not (yet) connected to the national digital infrastructure but that get an access point later. In this comparison, we track economic activity

of each town using nighttime lights as a proxy measure.

We find that the connection of remote towns in SSA to the *World Wide Web*, on average, leads to an increase in light emissions of about 11 percent, relative to similar towns not (yet) connected. This translates into about 3.3 percentage points higher growth in terms of GDP. Moreover, we decompose light emissions into growth in lit pixels (extensive margin), and in light intensity (intensive margin) and find higher significance for intensive-margin growth. Together with an assessment of changes in population showing no effect, this is more in line with growth in per capita productivity in connected towns rather than a spatial redistribution of economic activity. Further analyses suggest higher effects in towns with better market access and show local internet availability is associated with a shift in regional employment shares towards manufacturing. Overall, our results suggest significant effects even of low-speed internet in remote towns in rural SSA that are predominantly served via cybercafés.

Our findings have several implications for policy makers. Importantly, that internet infrastructure drives economic growth in remote towns beyond the large urban areas of developing countries. Internet infrastructure investments therefore are an important lever for regional development policy aiming to narrow the digital and economic divide within the developing world. When planning national backbone expansions, decision makers should take into account positive spillovers of connectivity on smaller, on-route towns and consider maximizing the number of access point along routes between nodal cities of the backbone. Evidence suggests there is a complementarity between internet infrastructure and market access. Moreover, our findings point to significant economic growth effects even with low-speed internet and through a low-cost local access mode that does not require high investments in ‘last mile’ infrastructure.

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

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.[1]

Keywords: geography; digitalization; networks; knowledge economy; colocation
JEL-No: L84; O18; O30; R32

[1] Versions of this chapter have been published as ifo Working Paper No. 386 and CRC Discussion Paper 422. 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, the 2nd CESifo Workshop on Big Data and the 12th European Meeting of the Urban Economics Association for valuable comments and suggestions. I am grateful to Lena Abou El-Komboz and Thomas Fackler for sharing data. Further, I thank Lara Mai, Raunak Mehrotra, Svenja Schwarz and Gustav Pirich for excellent research assistance and gratefully acknowledge public funding through DFG grant number 280092119.

2.1 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 of other products (Nagle, 2019; Andreessen, 2011). Yet, comprehensive empirical evidence on spatial collaboration of software developers is lacking.[2] 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 (see, e.g., Jones, 2021), 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

[2] 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).

virtual space (Emanuel et al., 2023).[3] 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).

In this paper, 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 market with relatively few cultural and language barriers and, thus, lower barriers to collaboration across space. The data is representative of the overall activity of software developers and offers unique and comprehensive insights into the industries' production process 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;

[3] 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).

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, which 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 increasing 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 than lower-quality links, 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 2.2 discusses related literature. In Section 2.3, I provide a brief background on digital collaboration in software development and present the data. The empirical analysis in Section 2.4 first explores the role of colocation and distance for collaboration in the highly digital setting of software developer networks (Section 2.4.1), compares the observed spatial collaboration pattern to less digital networks (Section 2.4.2), and explores the drivers colocated collaboration (Section 2.4.3). Section 2.4.4 presents robustness assessments and Section 2.5 concludes with a discussion.

2.2 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; Waltman et al., 2011), 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).[4]

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., 2023). 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 opportunities for

[4] 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).

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. This is in line with recent findings by Duede et al. (2024) for intellectual influence in science and the descriptive findings on the internal geography of firms by Bartelme and Ziv (2024). 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 communication 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.

(2018b) construct regional connectedness from *Facebook* data. Analyses of this data reveal a high degree of spatial clustering in social networks (Bailey et al., 2020a) and a strong association with travel (Bailey et al., 2020b) 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.

2.3 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.[5] 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.[6] 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;[7] 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 information, 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.[8] 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

[5] *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`, last accessed 02/15/2023.

[6] 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.

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

[8] 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.

public user profile. Thereof, 34% (778 thousand) can be georeferenced to a location within the United States.[9] 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.[10] 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.

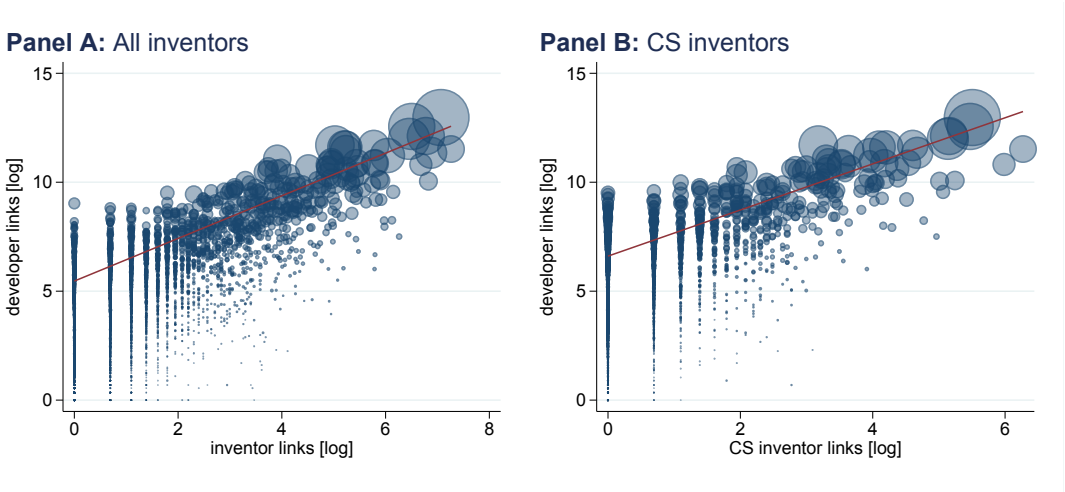
Summary statistics. In-sample users contribute to about 4.29 million *repositories*, i.e., open-source code projects on the platform. In total, they make roughly 97.3 million single code contributions to these projects, so-called *commits*. The most popular programming languages used on the platform are JavaScript, Python, as well as C and related languages (see Figure B.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.[11] See Figure B.3 for a visual impression.

Each user on average contributes to 28.5 projects (median: 14) in the observation period. 28% of projects are one-time uploads with one (inital) *commit*. To projects that are not one-time

[9] 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.
[10] New users in the last time interval are regarded as active if they contribute in this time interval.
[11] See, e.g., Luca (2015) for a review of user content generation on social media platforms.

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 2.1: 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 B.1. To distinguish different types of collaboration I use information provided in the data on the organizational affiliation, forks, stars, and followers (see Section B.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.[12] 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 B.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 B.2). Figure 2.1 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.[13] 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.

2.4 Empirical analysis

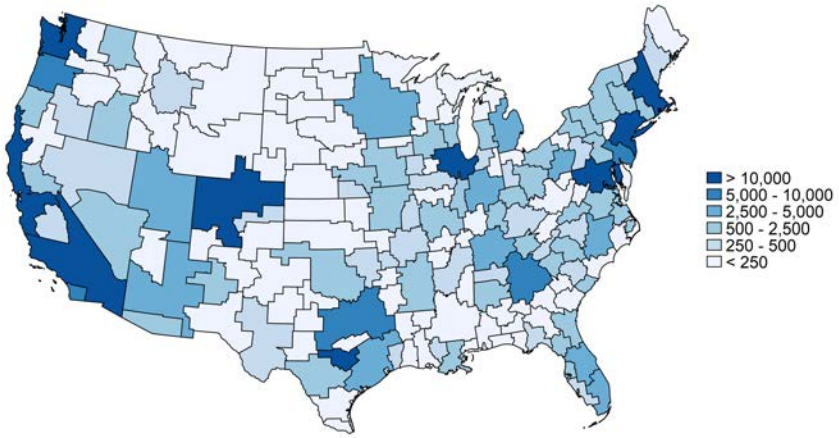
2.4.1 Main results

Concentration. Users are extremely concentrated in space. Figure 2.2 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

[12] For detailed information on supplementary data used here see Section B.1 in the Appendix.
[13] For detailed information on supplementary data used here see Section B.1 in the Appendix.

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.

Figure 2.2: Geographic distribution of users

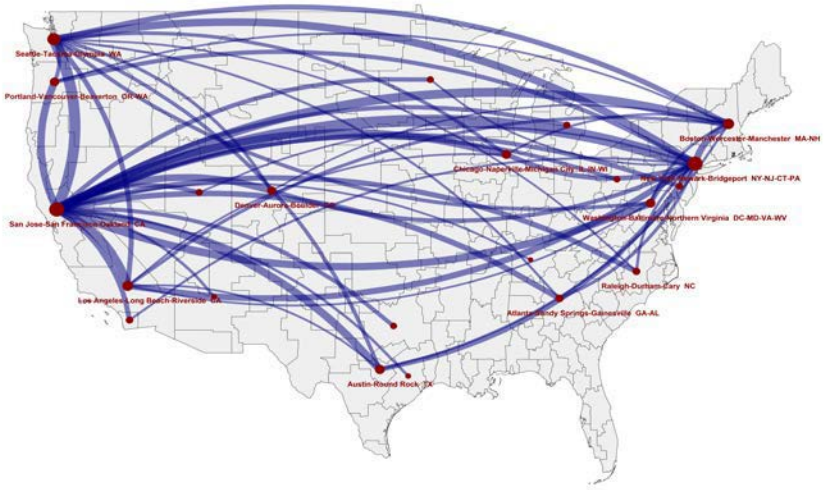


Notes: Map shows the number of (in-sample) users per economic area. The remote economic areas Anchorage, AK, and Honolulu, HI, are not shown. *Sources:* GHTorrent, own calculations.

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 B.5 for more complete information on the largest twenty economic areas according to these metrics.

Collaboration. Figure 2.3 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,

Figure 2.3: Inter-regional collaboration of users



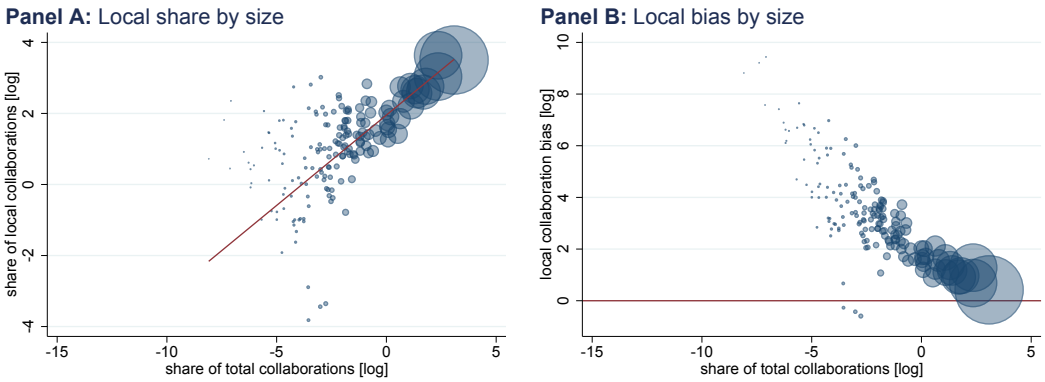
Notes: Map shows the structure of the U.S. software developer collaboration network. Important edges of the network, defined as links between economic areas above 20,000 connections, are shown in blue and scaled by the logarithm of the number of links. Economic areas shown in gray with their centroids 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.

central nodes correspond to the economic areas with the highest numbers of users (see Figure 2.2). 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 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.[14] 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.

[14] See Figure B.6 for a distribution plot.

Figure 2.4: (Local) collaboration and distance



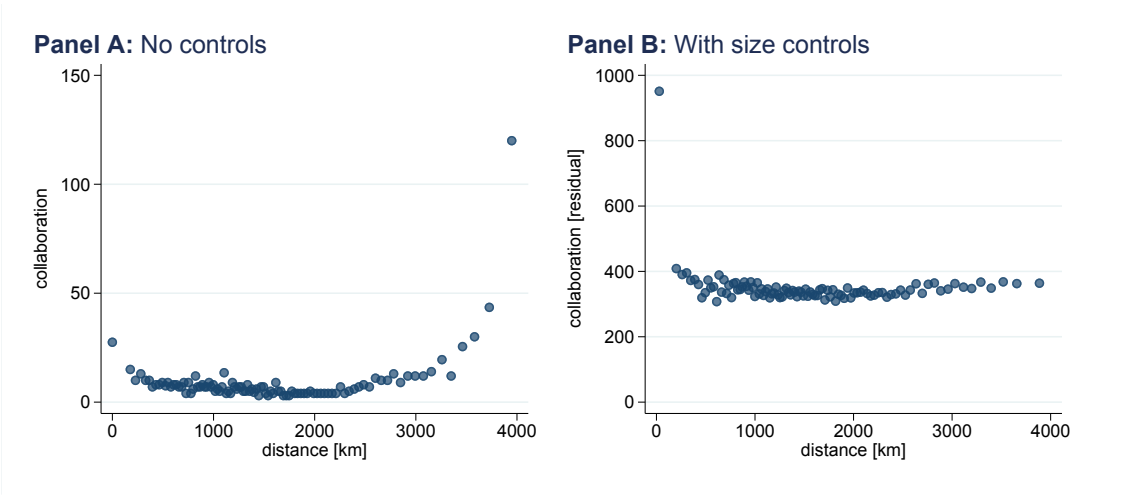
Notes: Plots depict localization patterns in the software developer network. Panel A shows the relationship between the share of collaborations of an economic area's users in all collaborations. The red line represents the best linear fit weighted by total collaboration share as economic-area size measure. 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.

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 2.4, 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 2.5 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-based geodesic distance between economic areas; Figure B.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.[15] 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.

Figure 2.5: 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

$$\text{links}_{i,j} = \beta_0 + \beta_1 \mathbb{1}\{\text{coloc}_{i,j}\} + \beta_2 \text{dist}_{i,j} + \mathbf{X_{i3}} + \mathbf{X_{j4}} + \mathbf{X_{ij5}} + \epsilon_{i,j} \tag{2.1}$$

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

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.[16] 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, $\mathbf{X_{ij}}$.

Table 2.1: 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 100 kilometers. 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.

The main results confirm collaboration is strongly positively associated with being colocated.

[16] 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.

Estimates in Table 2.1 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 characteristics, results suggest 0.1% to 0.6% fewer collaborations when distance increases by 100 kilometers. 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 2.1 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.

2.4.2 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.

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.[17] I arrive at a sample of around 17,000 U.S. inventors that filed a collaborative computer-science patent in this time period.

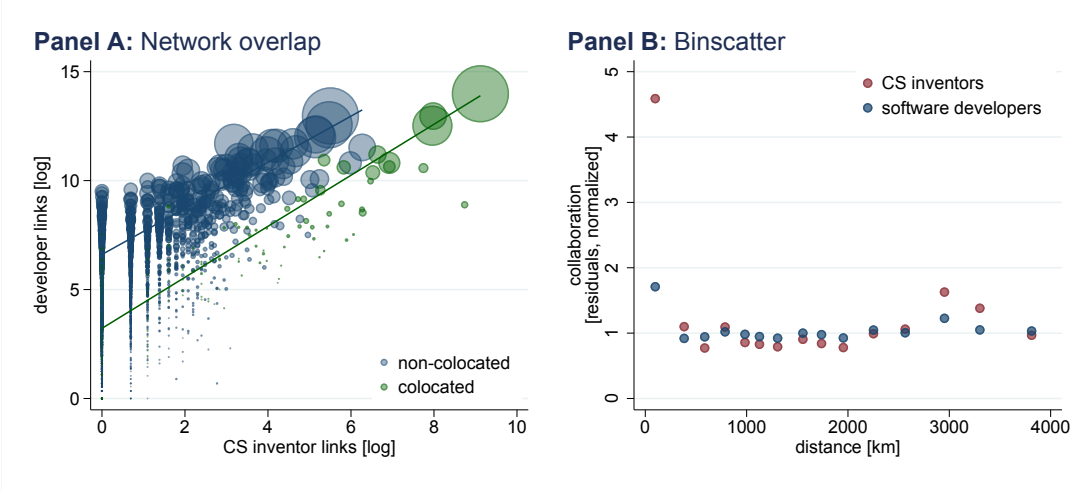
Network relatedness. Panel A of Figure 2.6 plots the relation between software developer and computer-science inventor networks and differentiates between (blue) and within (green) economic-area collaborations. Marker size represents a measure of economic-area size. There is a strong linear relationship between the two networks. This high inter-regional network overlap means that software developers and inventors exhibit similar inter-regional collaboration patterns.[18] 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 2.6, 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.

To quantify the difference in colocation effect size between the two networks, Panel B of Figure 2.6 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

[17] More information on data preparation is provided in the Appendix.
[18] Figure B.11 shows a similar plot for all inventors, a larger sample of around 76,000 individuals.

Figure 2.6: 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.

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 B.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.

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.[19] Bailey et al. (2018a) 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

$$\text{index}_{i,j} = \frac{\text{links}_{i,j}}{\text{users}_i * \text{users}_j}. \tag{2.2}$$

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).[20] Figure B.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 2.7; a data example for the economic area containing Los Angeles in Figure B.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

[19] See Bailey et al. (2018b) for a detailed discussion.
[20] For details on index construction and aggregation see Section B.1 in the Appendix.

to the SCI and, therefore, is independent of economic-area size. This shows that software developer and general friendship networks measured through size-independent indices such as GHCI and SCI feature no significant regional overlap.[21] Intuitively this is explained by general friendships typically being much more tied to one's geographic center of life.

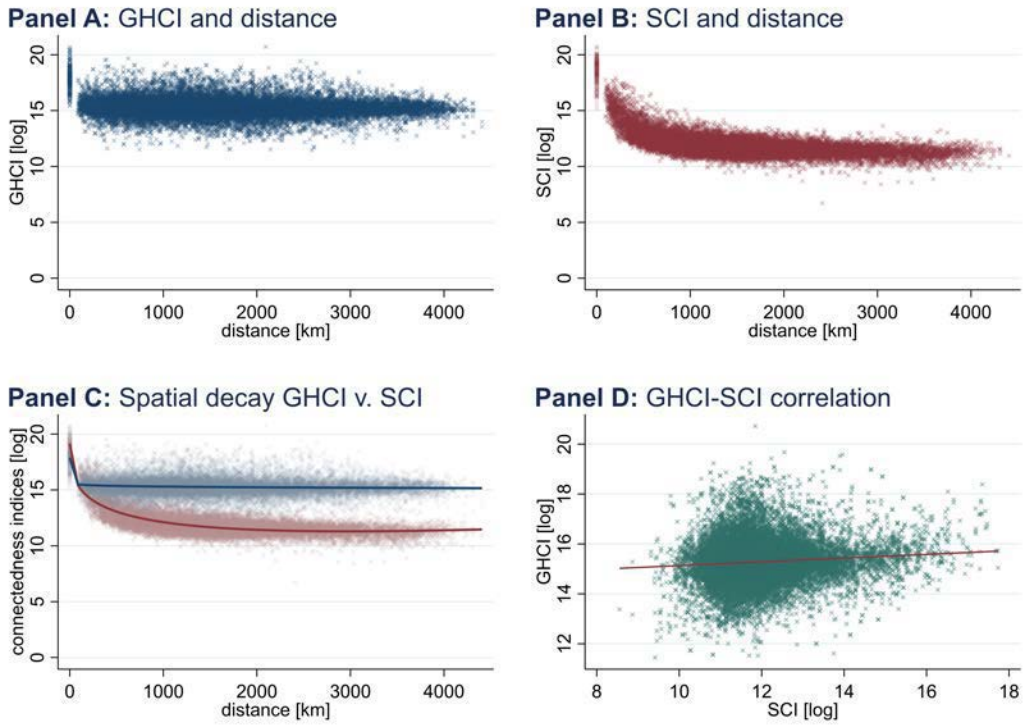
Comparing spatial decay. Data confirms the presence of a strong colocation effect in both networks. Figure 2.7 plots raw data from scaled GHCI (Panel A) and SCI (Panel B) after logarithmic transformation. A large colocation effect is already clearly visible in the raw data, represented by the sharp upward 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 500 kilometers.

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.[22] Panel C of Figure 2.7 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 B.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.

[21] 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 2.6 cannot be analyzed here.
[22] See Section B.1 in the Appendix for detailed information on the fractional polynomial model used.

Figure 2.7: 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. (2018a) 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. (2018a), U.S. Census Bureau, own calculations.

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.

2.4.3 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 2.4 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 B.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 (Duede et al., 2024; 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.

Table 2.2: 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 2.1. 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.

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 2.2 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.[23] 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 2.2. Similarly, looking at only users affiliated with one of the big tech firms (Amazon, Google, Apple, Microsoft, or Facebook) yields within-firm collaborations 35% less colocated compared to between-firm links with involvement of a big tech firm user. 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 B.10. Detailed regression results are presented in Table B.6 in the

[23] The organization size distribution is plotted in Figure B.4 in the Appendix.

Appendix.

Quality. Colocated and non-colocated collaboration potentially systematically differs in quality. Theoretically, there are two opposing forces at play (Lahiri, 2010). 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 (Lin et al., 2023). 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 2.2 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. This metric is especially insightful since most new knowledge today recombines existing works (see, e.g., Uzzi et al., 2013; Weitzman, 1998). 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 face-to-face requirement for high-quality

projects. Detailed regression results are presented in Table B.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 2.2 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 2.2.

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 B.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 B.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 2.2 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 B.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 (Wu et al., 2019). 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 2.2 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 B.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 B.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.

2.4.4 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 2.1 with alternative definitions of colocation. The results are shown in Table B.2. Including centroid-based distances of less than 100 kilometers captures only seven economic-area pairs but leads to a substantially smaller colocation effect of 7.73. Allowing distances up to 200 kilometers 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 100 kilometers.

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 B.8 plots the resulting coefficient estimates of these distance bin indicators. The coefficient for distances greater than 3200 kilometers 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 100 kilometers is estimated less precisely than others and is not significantly different from zero. Except for the last estimate, the

coefficient estimates tend to gradually become smaller for higher distances. This shows that the colocation effect is confined to small distances only and essentially vanishes thereafter, confirming findings from Panel B in Figure 2.5. 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 B.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 B.4 exhibit a similar pattern with respect to distance as the preferred specification (see Figure B.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 B.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 B.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 B.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.

2.5 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.

3 Virtually Borderless? Cultural Proximity and International Collaboration of Developers

Are national borders an impediment to online collaboration in the knowledge economy? Unlike in goods trade, knowledge workers can collaborate fully virtually, such that border effects might be eliminated. Here, we study collaboration patterns of some 144,000 European developers on the largest online code repository platform, *GitHub*. To assess the presence of border effects, we deploy a gravity model that explains developers' inter-regional collaboration networks. We find a sizable border effect of –16.4%, which is, however, five to six times smaller than in trade. The border effect is entirely explained by cultural factors such as common language, shared interests, and historical ties. The international border effect in Europe is much larger than the state border effect in the US, where cross-border cultural differences are much less pronounced, further strengthening our conjecture that culture is a main driver of the border effect in virtual collaboration.[1]

Keywords: digitization; software development; knowledge work; culture; language
JEL-No: F66; J61; O31; O33; O36

[1] This chapter is based on joint work with Lena Abou El-Komboz. We thank Raunak Mehrotra for excellent research assistance and gratefully acknowledge public funding through DFG grant number 280092119.

3.1 Introduction

Border effects, the reduction of economic exchange that flows across international borders, are one of the most robust and consistent empirical findings in international economics. Border effects (or home bias) are present, for example, in trade (e.g., Anderson and van Wincoop, 2003; McCallum, 1995), investment (e.g., Chan et al., 2005; Strong and Xu, 2003; French and Poterba, 1991) and innovative activity (e.g., Peri, 2005; Maurseth and Verspagen, 2002). Today however, digital exchange enabled by modern information and communication technologies (ICT) accounts for a sizable part of economic activity. In such settings of the digital economy, traditional explanations for the presence of border effects, such as trade or transportation costs, do not apply (Blum and Goldfarb, 2006).

In this paper, we therefore ask if a border effect is present in virtual collaboration, as well, and explore its relationship with cultural factors. Using unique data on the inter-regional collaboration of around 144,000 European developers on the largest online code repository platform, *GitHub*, we estimate the border effect in virtual collaboration in a parsimonious region-level gravity framework. We then assess potential drivers of the border effect via the inclusion of a large set of potential cultural determinants while controlling for confounding factors. As a reference, we estimate the border effect using the same model and data for US state borders, where cross-border cultural differences are much less pronounced compared to national borders in Europe.

The setting of developers collaborating online is particularly suitable here as it not only represents an important and representative sector of the knowledge economy (Korkmaz et al., 2024), but at the same time also precludes many of the traditional explanations driving border effects for multiple reasons. First, online code projects technically allow for fully virtual interaction and IT professionals' adoption of such technologies is high. Second, code development is not affected by transportation costs nor are open-source developers constrained by tariffs or bureaucratic barriers. Third, programming is codified to a higher degree compared to other knowledge work, which facilitates cross-border communication. And lastly, language barriers are likely less important as many developers speak English and use similar (universal) programming languages.

We find a significant digital border effect for developer collaboration in Europe of –16.4% after accounting for collaboration potential and geographic factors in the baseline gravity framework. Although this effect is sizable, it is five to six times smaller than in goods trade. The border effect is particularly high when at least one of the involved countries is small in

terms of hosted users. Our results further suggest cultural factors fully explain the digital border effect. Specifically, common interests, a common spoken language, and a shared history are significantly associated with the border effect while religious proximity and most political circumstances are unrelated to the border effect. Investigating several widely-used frameworks of cross-country cultural differences shows some relation of the border effect to preferences and interest. There is a particularly strong relation to shared interests in non-local business. In contrast, social ties do not explain much of the border effect but rather the distance gradient. Comparison with the state border effect in the US, a setting where cultural and language differences are largely absent, suggests that indeed culture is a main driver of the international border effect since the domestic border effect is much smaller.

This work entails several contributions that have important managerial and policy implications. It is one of few studies to investigate digital border effects, i.e., border effects in collaboration that technically can be shifted completely into the virtual space. To the best of our knowledge, we are the first to thoroughly examine border effects in software developer collaboration on a online platform. Estimated digital border effects are several magnitudes smaller compared to goods trade, where border effects are studied extensively. Generally, this points to fewer and less important barriers to international collaboration. While existing works on international collaboration are mainly concerned with travel costs or geographic factors, we relate the observed border effect to cultural factors. As geography increasingly becomes less relevant in the knowledge economy, the importance of cultural factors for international collaboration in the digital economy increases. We demonstrate which among the many dimensions of culture, broadly defined, are most strongly related to the digital border effect among software developers in Europe.

The remainder of this paper is organized as follows. We discuss the related literature in Section 3.2. Section 3.3 introduces the data. In Section 3.4, we discuss the empirical model. Results are presented in Section 3.5 and Section 3.6 concludes.

3.2 Related literature

ICT and remote collaboration. This study contributes to three related strands of literature. First, there is a growing literature in economics on the impact of ICT on remote collaboration. Existing work shows that ICT tends to foster inter-regional trade (Visser, 2019; Steinwender, 2018; Jensen, 2007), research and innovation (Forman and van Zeebroeck, 2019; Agrawal and Goldfarb, 2008), and entrepreneurship (Agrawal et al., 2015). However, geographically

close exchange tends to increase disproportionately (Akerman et al., 2022; Agrawal and Goldfarb, 2008), in line with theoretical considerations that ICT and geographic proximity are complements (Gaspar and Glaeser, 1998). In knowledge work, colocation is especially important (see, e.g., Goldbeck, 2023; Urry, 2002; Olson and Olson, 2000) and average collaborator distance in teams increases with ICT adoption (Adams et al., 2005). In non-collaborative office settings, remote work is feasible and may even increase productivity (Choudhury et al., 2021; Bloom et al., 2015). Yet, studies find that face-to-face interaction opportunity remains valuable in many settings (e.g., Gibbs et al., 2023; Atkin et al., 2022; Brucks and Levav, 2022; Yang et al., 2022; Pentland, 2012), partly due to improved learning (Emanuel et al., 2023; van der Wouden and Youn, 2023; Eckert et al., 2022; Akcigit et al., 2018; De La Roca and Puga, 2017; Glaeser and Mare, 2001). Still, Chen et al. (2022) find that the costs of distributed teams tend to fall over time as remote collaboration technology improves and learning effects materialize and Forman and Zeebroeck (2012) show internet adoption leads to more geographically dispersed inventor teams.

Geography, gravity, and border effects. There is a large literature examining the determinants of geographic distribution of economic activity. Large parts of this literature center around the gravity model (Tinbergen, 1962; Bergstrand, 1985) that considers geographic distance and size to empirically explain economic exchange, most prominently trade (Anderson, 1979; Eaton and Kortum, 2002; Disdier and Head, 2008; Head and Mayer, 2010), but also knowledge flows (Bahar et al., 2022; Montobbio and Sterzi, 2013; Picci, 2010), foreign aid (Alesina and Dollar, 2000), online behaviour (Steegmans and de Bruin, 2021), or migration (van der Kamp, 1977; Lewer and van den Berg, 2008). For trade, the impact of distance has fallen steadily over time (Yotov, 2012), especially between rich countries (Brun et al., 2005). Blum and Goldfarb (2006) were first to show that the gravity model holds even for digital goods, where there are no trade costs, but also find no distance effect for non-taste dependent products such as software. Hanson and Xiang (2011) confirm gravity for movie exports, another product with no trade or transport costs. In contrast, Lendle et al. (2016) find distance irrelevant in e-commerce. Virtual proximity is positively associated with services trade (Hellmanzik and Schmitz, 2016, 2015) and investment (Hellmanzik and Schmitz, 2017). Recent evidence from gravity applications for developer collaboration shows smaller effects of distance globally when compared to trade (Fackler and Laurentsyeva, 2020) and a negligible distance effect for the US but significant colocation effects (Goldbeck, 2023).

Within the gravity framework, McCallum (1995) was first to explicitly estimate border effects for trade and Anderson and van Wincoop (2003) refine the empirical model and provide theoretical

foundations. There is vast empirical evidence on border effects in trade (e.g., Head and Mayer, 2021; Havranek and Irsova, 2017; Anderson et al., 2014; Millimet and Osang, 2007; Chen, 2004; Helliwell and Verdier, 2001; Wolf, 2000) with recent work on European international borders (Santamaría et al., 2023a,b) pointing to still very large effects. In comparison, investigations of the border effect in collaboration and knowledge flows are relatively scant. Singh and Marx (2013) find significant but diminishing border effects in patent collaboration. However, Li (2014) shows that the decrease over time is driven by age effects. Griffith et al. (2011) point out that the speed of patent citations, as measure for knowledge spillovers, steadily increased with improved ICT and travel cost reductions.

Cultural proximity in the knowledge economy. A growing strand of literature studies the role of cultural factors as deep determinants of economic activity (Alesina and Giuliano, 2015; Guiso et al., 2006). Considering cultural factors in gravity applications is widely established. Deardorff (1998) distinguishes trade barriers related to transport costs and unfamiliarity. Since then, the gravity literature routinely found cross-country cultural differences important determinants of trade (e.g., Gokmen, 2017; Felbermayr and Toubal, 2010; Boisso and Ferrantino, 1997) and other economic outcomes including innovation (e.g., Gorodnichenko and Roland, 2017), collaboration (e.g., Bercovitz and Feldman, 2011; Cummings and Kiesler, 2007; Hinds and Bailey, 2003), and productivity (e.g., Stewart and Gosain, 2006). Since culture is a fuzzy concept, the literature investigates more tractable sub-dimensions of culture such as preferences (Kondo et al., 2021; Guiso et al., 2009; Huang, 2007), institutions (Hoekman et al., 2010; Acemoglu et al., 2005), shared history (Alesina and Dollar, 2000), social ties (Bailey et al., 2021; Agrawal et al., 2006), or language (Visser, 2019; Falck et al., 2012; Melitz, 2008; Baier and Bergstrand, 2007).

Cultural factors play an important role in knowledge-intensive and innovative sectors, as well. Several studies identify common language as important, e.g., for effective team communication (Koçak and Puranam, 2022), research performance (Cao et al., 2024), or knowledge transfer (Parrotta et al., 2014). Gomez-Herrera et al. (2014) study e-commerce and also find linguistic borders important but no difference in the border effect compared to offline trade. A large strand of literature examines the role of social ties on knowledge worker collaboration (e.g., Bercovitz and Feldman, 2011) and knowledge flows (e.g., Diemer and Regan, 2022; Reagans et al., 2005). As social ties are closely related to geographic distance (Bailey et al., 2018b; Breschi and Lissoni, 2009) they are an important channel to explain the robust distance effect in gravity applications (Diemer and Regan, 2022; Garmendia et al., 2012; Bercovitz and Feldman, 2011; Breschi and Lissoni, 2009) as well as for collaboration success

more generally (Hahn et al., 2008; Cowan et al., 2007; Grewal et al., 2006). Organizational links (Duede et al., 2024; Fadeev, 2023; Adams et al., 2005) as well as immigration (Tadesse and White, 2010) attenuate negative border effects associated with culture.

Specifically for (open-source) software development, existing works in the organizational economics literature study culture extensively. For example, von Engelhardt and Freytag (2013) shows that cultural and institutional factors explain software developers' open-source software (OSS) activity differences across countries. OSS activity differences are partly driven by social identity (Bagozzi and Dholakia, 2006) and intellectual property rights (O'Mahony, 2003), and Stewart and Gosain (2006) show shared values make OSS teams more effective. Furthermore, culturally diverse teams are associated with improved performance (Ren et al., 2016; Daniel et al., 2013; Page, 2010; van Knippenberg and Schippers, 2007) and creativity (Jang, 2017), at least up to a certain threshold (Ren et al., 2016; van Knippenberg and Schippers, 2007).

3.3 Data

Virtual collaboration. We compute regional collaboration networks of software developers on the by far largest online code repository platform, *GitHub*, with about 73 million users worldwide in 2021 (GitHub, 2021). To this end, we draw on the *GHTorrent* database by Gousios (2013), which mirrors the data publicly available via the *GitHub* API and generates a queryable relational database in irregular time intervals.[2] This paper relies on ten *GHTorrent* snapshots dated between 09/2015 and 03/2021, which contain data from public user profiles and repositories as well as a detailed activity stream capturing all contributions to and events in open-source repositories.[3] *GitHub* projects ("repositories") are maintained using the integrated version control software *git*. Importantly, the nature of the *git* version control system allows us to observe each users' activity and collaborators in public repositories. Additionally, users can indicate their location on their *GitHub* profile. We assign users to cities via exact matching to city names in the *World Cities Database*. Goldbeck (2023) validates the location information using various benchmarks, finding no systematic bias at the regional and region-pair level. Defining a collaboration as active contribution during the observation period to at least one joint project, we compute the regional collaboration network at the

[2] *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`.
[3] 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.

NUTS2 level.[4]

Figure 3.1: Regional collaboration network

Notes: Map shows the structure of the European software developer collaboration network. Important edges of the network, defined as links between economic areas above 25,000 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. Ireland not shown. *Sources:* GHTorrent, own calculations.

Overall, our data contains 290 NUTS2 regions in 34 European countries[5] and captures the activity in open-source repositories of 144,121 active, geolocated, and collaborating users. Users are highly concentrated in space with 39% of users in the ten largest regions.[6] The London metro area is by far the biggest region with more than 19,000 users, followed by Paris metro (Île-de-France) with 11,496 and Amsterdam metro (Noord-Holland) with 4,794. The left map in Figure C.2 shows the spatial distribution of users across European regions. Generally,

[4] We merge the NUTS2 regions for London, UKI3 through UKI7, to increase comparability, as this is the only capital city metro area that is split into multiple NUTS2 regions.
[5] Table C.1 reports user numbers by country.
[6] Note, however, that this concentration is much less pronounced than in the US where this number is about 80% (Goldbeck, 2023).

this pattern is also reflected in the regional collaboration patterns depicted in Figure 3.1, which shows the most important nodes and edges in the regional collaboration network. The red nodes are scaled by the total number of collaborations and edge width represents bi-regional collaboration intensity. London as the central hub for software development in Europe is clearly visible and we observe most collaborations between the large cities in terms of the number of software developers. We are interested in the border effect, i.e., the relation of international versus national collaborations after controlling for geographic factors in a gravity framework. Figure C.1 plots distance histograms for cross-border and within-country network edges and shows there is a large region of common support in the distributions to facilitate robust estimation.

Cultural proximity. We associate potential border effects to various measures of cultural proximity, drawing on multiple data sources. First, we use a composite measure of cultural proximity derived from detailed data on online behaviour (Obradovich et al., 2022). This large-scale data collection effort systematically queries the *Facebook* marketing API to dissect societies' interests along hundreds of thousands dimensions. The API offers insights derived from users' self-reported interests, clicking behaviours and likes on the platform, as well as software downloads and behaviour on other websites employing *Facebook* ads. Due to the large number of active users on *Facebook* and the representativeness of in-sample users to the general population (Bailey et al., 2018b), this source provides insight into cultural differences at unprecedented scale. Specifically, from the universe of *Wikipedia* articles on *DBpedia*, Obradovich et al. (2022) extract 60,000 interest dimensions with at least 500,000 users worldwide to create a composite as well as sub-indices for cultural proximity as cosine distance between the interest vectors of populations k and l

$$\cos\text{dist}(k,l) = 1 - \cos(\theta) = 1 - \frac{S_k * S_l}{\|S_k\|\|S_l\|} \tag{3.1}$$

where S_k denotes a n-dimensional vectors with components s_{ik} that measure the share of population k holding a particular interest $i = 1, \ldots, n$ and θ is the angle between S_k and S_l. Consequently, the resulting index is independent of n. Obradovich et al. (2022) validate this composite index using traditional composite measures of culture and find a high overlap. Still, their index improves in granularity and represents a bottom-up approach in contrast to top-down measurement along few select dimensions. We use the cross-country composite measure as well as the sub-indices for the 14 main interest dimensions.

Second, we relate border effects to genetic distance, a well-established proxy for cultural

factors associated with ethnicity (Spolaore and Wacziarg, 2009; Creanza et al., 2015). We use the cross-country genetic distance data from Creanza et al. (2015), which measures the degree of similarity in vertically transmitted characteristics as aggregated differences in allele frequencies for highly predictive parts of a chromosome. In particular, we follow the literature and use the co-ancestor coefficients (also: F_{ST} distance) that is based on heterozygosity, i.e., the probability of two specific areas of genes being different. By this measure, we proxy for co-ancestral distance between national populations, a measure found highly relevant for economic outcomes (see, e.g., Bove and Gokmen, 2018; Spolaore and Wacziarg, 2009).

Third, we account for important cultural factors traditionally used in the gravity literature and captured in the *CEPII Gravity Database* (Conte et al., 2022). As language is commonly found to be an important factor for collaboration, we use the indicator for common spoken language (Melitz and Toubal, 2014). Likewise, we control for religious proximity measured as the product of the shares of Catholics, Protestants, and Muslims in origin and destination countries (Disdier and Mayer, 2007; La Porta et al., 1999). As measures for a shared history we account for two factors: whether countries ever were part of the same nation, and whether they have a colonial history, both sourced form the *CEPII GeoDist Database* (Mayer and Zignago, 2011).

Fourth, we assess the relationship to traditional survey-based cultural dimensions as measured in the Hofstede model (Hofstede, 2011) and the *Global Preferences Survey* (Falk et al., 2018). The Hofstede model measures national cultural dimensions quantitatively along six dimensions: power distance, uncertainty avoidance, individualism/collectivism, achievement and success, long/short-term orientation, and indulgence/restraint. The *Global Preferences Survey* elicits cross-country differences in preferences along the six dimensions patience, risk taking, positive/negative reciprocity, altruism, and trust.

Supplementary data. We further use regional-level social connectedness measures derived from *Facebook* (Bailey et al., 2018b) to investigate potential mechanisms of collaboration. For better comparability, we compute the *GH Connectedness Index* (GHCI; see Goldbeck, 2023) similarly to the *Social Connectedness Index* (SCI) as the relative probability of connection between users in two regions

$$\text{index}_{i,j} = \frac{\text{links}_{i,j}}{\text{users}_i * \text{users}_j}, \tag{3.2}$$

and scale between 1 and 1,000,000,000. Note that these indices are independent of regions size by design. Furthermore, we use various additional variables traditionally used in gravity

applications from the *CEPII Gravity Database* (Conte et al., 2022). In addition, we use *Fraser Institute's Economic Freedom of the World Index* and the *Freedom House Index of Political Rights* from Graafland and de Jong (2022) and compute bilateral differences in these indices.

3.4 Empirical model

To estimate border effects in software developer collaboration, we deploy the gravity model, which is widely used to explain economic exchange between countries such as bilateral migration, trade, and FDI flows (see, e.g., van der Kamp, 1977; Anderson, 1979; Frankel and Rose, 2002). In the innovation literature, the gravity model is applied to describe knowledge flows and collaboration measured through patenting activity (e.g., Bahar et al., 2022; Montobbio and Sterzi, 2013; Picci, 2010). While traditionally applied in cross-country settings the model is equally suitable at the sub-national regional level, where it is routinely used to estimate border effects (e.g., Anderson and van Wincoop, 2003; Wolf, 2000; McCallum, 1995). Note that border effects gravity models are theory-consistent and, because they feature domestic flows by design, even more so than traditional cross-country gravity (Yotov, 2022). In our context, the gravity model, in its simplest form, states that regional collaboration is proportional to the product of the regions' masses (measured by the number of local users) and inversely proportional to the distance between the regions. We take the parsimonious gravity model from McCallum (1995), which includes an indicator for cross-border collaboration, as starting point for estimating the border effect:

$$\ln(y_{i,j}) = \beta_0 + \beta_1 \text{crossborder}_{i,j} + \beta_2 \text{coloc}_{i,j} + \beta_3 \ln(\text{dist}_{i,j}) + \delta_i + \delta_j + \epsilon_{i,j} \tag{3.3}$$

where $y_{i,j}$ represents the number of bilateral collaborations between regions i and j including domestic collaborations $i = j$. The dummy variable $\text{crossborder}_{i,j}$ indicates if region i is located in a different country than region j, and $\text{dist}_{i,j}$ denotes the geographic distance between the regions. We further add a colocation indicator, $\text{coloc}_{i,j}$, to account for strong colocation effects in collaboration (Goldbeck, 2023; Urry, 2002; Olson and Olson, 2000). Origin and destination fixed effects δ_i and δ_j account for unobserved regional determinants of collaboration common across all partner regions. The coefficient β_2 captures the elasticity of collaboration with respect to geographic distance, which we expect to be negative from theory. The border effect is given by our coefficient of interest β_1, expected to be negative or zero depending on the presence of a border effect in the population.

It is important for the interpretation of the effect to discuss how the border effect is

conceptualized in the model. The key identifying assumption for the border effect in the gravity model is that there are no third factors related to the border indicator that drive collaboration. The plausibility of this assumption depends on how we think of the border effect. If we think of the border effect narrowly, in the sense that the border itself causes collaboration to decrease, this assumption is clearly implausible. However, if we conceptualize the border effect as a proxy measure of all things that vary across borders and possibly determine collaboration, it is plausible yet tautological. Put differently, the border effect estimated from Equation 3.3 represents a quantification of how much inter-regional collaboration declines, on average, for cross-border links as compared to within-country links. Therefore, the border effect should be interpreted as descriptive proxy measure of many potential deeper determinants rather than a causal estimate of the effect of the border itself.

To assess the specific drivers of this broadly defined border effect we extend the baseline model to include variables at the country-pair level measuring different cultural dimensions that vary across borders:

$$\begin{aligned}\ln(y_{i,j}) = \beta_0 + \beta_1 \text{crossborder}_{i,j} + \beta_2 \text{coloc}_{i,j} + \beta_3 \ln(\text{dist}_{i,j}) \\ + \mathbf{X}'_{c(i),c(j)}\boldsymbol{\gamma_4} + \mathbf{X}'_{i,j}\boldsymbol{\gamma_5} + \delta_i + \delta_j + \epsilon_{i,j}\end{aligned} \tag{3.4}$$

where $\mathbf{X}_{c(i),c(j)}$ is a vector of variables that measure differences between the respective country of region i, $c(i)$, and the country of region j, $c(j)$. By definition, these differences are zero if region i and j belong to the same country, i.e., $c(i) = c(j)$. Thus, the coefficients $\boldsymbol{\gamma_4}$ capture the part of the border effect that is attributable to a particular cross-border difference while β_1 is the residual part of the average border effect not explained by the included variables in $\mathbf{X}_{c(i),c(j)}$. $\mathbf{X}_{i,j}$ is a vector of region-pair level determinants of collaboration and $\boldsymbol{\gamma_5}$ are the related coefficients.

As in the baseline model, the main assumption for causal interpretation of the coefficients $\boldsymbol{\gamma_4}$ is that there are no omitted factors related to $\mathbf{X}_{c(i),c(j)}$ that determine inter-regional collaboration. Note that the cross-border indicator isolates the remaining part of the border effect and therefore provides indication for the presence of omitted variables when significant. Nonetheless, country-pair explanatory variables that are related to unobserved determinants of collaboration are a threat to identification. Together with potential measurement error, especially in related explanatory variables, this cautions us of a narrow interpretation of the separate coefficients in $\boldsymbol{\gamma_4}$.

Especially since cultural factors are often interrelated and can have common deep determinants, a narrow causal interpretation is likely inappropriate. Rather, the model provides some indication of possible determinants as it points to dimensions that are statistically associated with the border effect. Plausible, theory-guided selection of explanatory variables is therefore paramount to avoid spurious correlation issues. We return to this discussion in Section 3.5.3. Note that Equations 3.3 and 3.4 are partial equilibrium models and, as such, estimated border effects should not be misconstrued as counterfactual for border removal, as wiedely acknowledged in the literature (see, e.g., Santamaría et al., 2023a; Havranek and Irsova, 2017).

3.5 Results

3.5.1 Digital border effect

Table 3.1 reports estimation results of the border effect for online collaboration among software developers in Europe. We start with a model that does not consider gravity and subsequently control for size and geographic distance. The raw correlation in model (1) suggests a large border effect of 60% less collaborations. Controlling for size in terms of logarithms of multiplied user bases in origin and destination regions halves the effect. The large positive coefficient on multiplied user bases demonstrates the importance of collaboration potential. Model (3) drops the functional form assumption for the size effect and instead includes unobserved regional characteristics using origin and destination region FE. This more flexible model slightly increases the estimate of the border effect. Finally, our preferred specification in model (4) resembles a typical parsimonious gravity model that additionally controls for geographic distance. We include logarithmic distance between origin and destination region centroids as specified in Equation 3.3. Since our data features within-region collaborations and Goldbeck (2023) finds colocation hugely important for collaboration, we also add a colocation indicator. As expected, results show a highly significant negative relation of collaboration and distance and a substantial collaboration premium for colocation.

There still is a border effect in our preferred baseline specification, with 16.4% fewer collaborations for region-pairs that are located in different countries compared to within-country pairs. While the border effect is economically significant, it is much smaller than for trade. The meta-analysis by Havranek and Irsova (2017) aggregates 263 estimates for the EU from similar gravity model specifications and finds a border effect of –91.5%[7], a slightly smaller effect size than the original estimates of McCallum (1995) and nearly identical to

[7] Cf. the unweighted mean coefficient for the EU in Table 1 in Havranek and Irsova (2017), expressed as home

Table 3.1: Border effect in collaboration

Collaboration	(1)	(2)	(3)	(4)
Cross-border	-0.906***	-0.371***	-0.446***	-0.180***
	(0.041)	(0.016)	(0.012)	(0.014)
Users, multiplied [log]		0.755***		
		(0.002)		
Colocation				0.862***
				(0.068)
Distance [log]				-0.129***
				(0.007)
Origin FE			×	×
Destination FE			×	×
Observations	84,100	84,100	84,100	84,100
Adj. R^2	0.011	0.837	0.919	0.922
Border effect	-59.6%	-31.0%	-36.0%	-16.4%

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. Users, multiplied, is the natural logarithm of the multiplication of the number of users in origin and destination. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, own calculations.

the border effect for Europe in Santamaría et al. (2023b) of –90.4%[8] estimated from recent granular freight data. Thus, a comparison to their results suggests a five to six times larger border effect in (goods) trade compared to (online) software developer collaboration. This is generally in line with our conjecture that national borders should play a minor or no role for virtual collaboration of software developers. Still, there is significant heterogeneity in the border effect. Table C.2 demonstrates that the border effect is systematically related to the number of country-wide users. Model (2) shows the border effect roughly doubles when a small country is involved, defined as hosting an above-median number of users. Model (3)

bias of $\exp(2.55) - 1 \approx 11.8$, translated into a percentage border effect as defined here via $\left(\frac{1}{\exp(2.55)-1} - 1\right) * 100$.

[8] Cf. the border effect coefficient in Table 1 column (2) of Santamaría et al. (2023b), translated into a percentage border effect as defined here via $(\exp(-2.34) - 1) * 100$.

shows the effect does not differ depending on whether both countries are small or just one, meaning there is a smaller border effect among large countries.

3.5.2 The role of culture

As there still is a significant border effect present in virtual collaboration, we investigate potential channels through which cross-border collaboration of software developers might be affected. We elicit association of various cultural factors with the border effect and collaboration by including appropriate cross-country level variables as specified in Equation 3.4.

Table 3.2 reports the results of variations of our baseline model that consider cross-country cultural differences. Note that the metrics for culture are available only for a subset of countries. For consistency, we estimate all models on the same, reduced sample that features a slightly higher baseline border effect in model (1). In model (2), we add two distinct composite measures of culture. First, we take the cultural distance metric from Obradovich et al. (2022) derived from common interests on *Facebook* and validated using traditional, mostly survey-based, metrics of culture. Second, we control for genetic distance from Spolaore and Wacziarg (2009) as a well-established proxy for cultural factors associated with ethnicity. The coefficient estimates of both distance measures have the expected negative sign. Cultural distance is strongly negatively associated with collaobration while genetic distance is much less relevant and also features weaker significance. Importantly, the border effect is entirely explained by these cultural distance composite measures, as shown by the insignificant point estimate close to zero of the border effect coefficient.

In model (3), we further add specific cultural factors that have been identified as relevant in the previous literature, namely common language, religious distance, and a common history reflected by same country or colonial history. Religious distance is statistically and economically insignificantly related to collaboration.[9] In contrast, there appears to be a sizable benefit from common spoken language of around 8.4% more collaborations, although imprecisely estimated. On the one hand, this makes sense as it eases communication. On the other hand, most knowledge work professionals speak English and code projects in software development are written in computer code. Reassuringly, the magnitude of the language effect is almost 14 times smaller compared to trade, where the corresponding semi-elasticity is 0.775 (Melitz and Toubal, 2014).[10] A shared colonial history is often highly

[9] Note that this might reflect that religious differences in Europe are generally small.
[10] Cf. column (2) in Table 3 of Melitz and Toubal (2014). Note that estimate magnitudes for common (spoken)

Table 3.2: Collaboration and cultural proximity

Collaboration	(1)	(2)	(3)	(4)
Cross-border	-0.233***	-0.009	-0.014	0.013
	(0.012)	(0.035)	(0.037)	(0.038)
Colocation	1.341***	1.485***	1.476***	1.472***
	(0.066)	(0.069)	(0.070)	(0.070)
Distance [log]	-0.046***	-0.016**	-0.018**	-0.009
	(0.007)	(0.008)	(0.008)	(0.008)
Cultural distance		-0.097***	-0.081***	-0.080***
		(0.016)	(0.017)	(0.017)
Genetic distance		-0.001**	-0.001*	-0.001*
		(0.000)	(0.000)	(0.000)
Common language			0.082**	0.062*
			(0.034)	(0.034)
Religious distance			-0.005	-0.007
			(0.020)	(0.020)
Same country history			-0.071**	-0.078***
			(0.028)	(0.028)
Colonial history			0.011	0.001
			(0.016)	(0.016)
Social connectedness				0.013***
				(0.004)
Origin FE	×	×	×	×
Destination FE	×	×	×	×
Observations	55,169	55,169	55,169	55,169
Adj. R^2	0.947	0.947	0.947	0.947

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. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, Obradovich et al. (2022), Creanza et al. (2015), Bailey et al. (2018b), CEPII, own calculations.

predictive in gravity models but does not explain collaboration today. This is likely due to the few colonial relationships within Europe. History as a same country is associated negatively with collaboration, which is surprising only at first and likely relates to the fact that this indicator captures mostly historical occupations in the former Yugoslavia and Austria-Hungary that lead to disrupted relationships until today (e.g., Kešeljević and Spruk, 2023).

Model (4) additionally adds social connectedness between regions as explanatory variable for collaboration. Social connectedness is highly statistically and economically significantly and positively related to collaboration. Controlling for social connectedness leads to irrelevance of geographic distance and a smaller language effect, but otherwise does not significantly alter the results. This points to the distance effect being driven by social connections and is reassuring towards the other effects. Note, however, that social connectedness might constitute a bad control in our setting as it likely is determined by cultural factors, as well. Therefore our preferred specification is model (3). While the relevance of colocation remains highly important throughout all specifications, geographic distance is statistically significant at a lower level and the coefficient size shrinks considerably. This is in line with empirical evidence on knowledge worker collaboration suggesting a high relevance of face-to-face meeting possibility (e.g., Emanuel et al., 2023; Atkin et al., 2022) but irrelevance of geographic distance otherwise (cf. Goldbeck, 2023) and feeds into the discussion that geography, in most models, is to a large extent merely a proxy for deeper determinants of outcomes (see, e.g., Waldinger, 2012; Azoulay et al., 2010).

We further investigate the relation between culture and international collaboration using established frameworks for particular cultural dimensions. First, we exploit the decomposition of the cultural interest composite measure by Obradovich et al. (2022) into 14 subcategories of interest. The results reported in Table C.3 reveal that especially different interests in the category *non-local business* explain the border effect. This means that international software developer collaboration is associated with overlapping professional interests with respect to industries and companies. It is, however, unclear if common professional interests are responsible for increased collaboration or if the presence and relation to local industries are a common driver of both collaboration and interests. Existing literature points towards an important role of organizations in shaping software developer collaboration (e.g., Duede et al., 2024; Goldbeck, 2023). Other subcategories are relatively unimportant, but mostly show positive associations. This points to cultural differences not being unidimensionally

language in log-specifications are generally quite robust in the trade literature (Melitz, 2008). Yet, most semi-elasticities refer to a worldwide sample. Still, estimates for European samples are comparable in size (see, e.g., Fidrmuc and Fidrmuc, 2014).

negatively related to collaboration but rather paints a more nuanced picture that some cultural differences, e.g. with respect to food or lifestyle, might in fact spur collaboration.

Second, we explore how cross-country differences in preferences relate to international collaboration. To this end, we use the six preference dimensions from the *Global Preferences Survey*: patience, risk taking, trust, altruism as well as positive and negative reciprocity. Table C.4 reports the results and shows that especially patience and positive reciprocity are negatively related to collaboration. Negative reciprocity explains collaboration to a lesser extent and is only weakly significant and the other dimensions are statistically insignificant, although point estimates are negative throughout. Generally, cross-country differences in preferences partly explain the border effect but only to a small extent.

Third, we use the established traditional cross-country measures of culture by Hofstede (2011) to study possible associations with collaboration. Of the six standard dimensions (power distance, individualism, achievement and success, uncertainty avoidance, long-term orientation, and indulgence), only power distance is significantly and negatively related to collaboration as shown in Table C.5. Individualism is also negatively related to collaboration but only weakly significant. Overall, the Hofstede cultural dimensions do not prove useful to explain the border effect as the point estimate is only slightly reduced when including differences in the six cultural dimensions.

3.5.3 Robustness

We demonstrate the robustness of the digital border effect estimated in Table 3.1 in multiple ways. First, we follow the methodology in Santamaría et al. (2023b) and compute an independence benchmark that disregards everything but the size component of gravity. This essentially corresponds to a theory in which all user-pairs feature equal probability of collaboration independent of their locations. We then relate observed collaborations to the benchmark in Panel (a) of Figure C.9 and distinguish cross-border, within-country, and colocated links. This shows the strong predictive power of the logarithmic multiplication of region size in terms of users. It is reassuring that the relationship between collaboration potential measured by multiplied user size is not significantly different between cross-border and within-country collaborations. Importantly, the analysis confirms that collaboration probability is significantly increased for within-country compared to cross-border collaborations, depicted by a shift to the right of the distribution in Panel (b) of Figure C.9.

Second, we plot residuals of fixed-effects models disregarding the cross-border indicator in Figure C.10. Panels (a) and (b) plot the averages and distributions of residuals for cross-border and within-country collaborations for the baseline fixed-effects model without and with geography controls, respectively. We generally observe well-behaved residual distributions, which is reassuring of our model specification. The significant right-shift of the residual distribution for within-country collaborations points to omitted variables bias in models that disregard border effects and, therefore, the presence of border effects in virtual collaboration. The narrowing of this gap between the distributions in Panel (b) compared to Panel (a) while still retaining statistical significance shows that geographic factors are important but do not fully explain the raw border effect. This is corroborated by models featuring a non-parametric distance specification. Figure C.5 compares non-parametric models with and without the cross-border indicator. Results show that considering the cross-border indicator significantly flattens the distance gradient and decreases the colocation effect.

Third, we calculate the size-independent *GH Connectedness Index* (GHCI; cf. Goldbeck, 2023), which is similar to the *Social Connectedness Index* (SCI) by Bailey et al. (2018b), and directly plot the relation to distance for within-country and cross-border links, respectively, in Figure C.6. As depicted in Figure C.7, GHCI and SCI feature similar distributional shapes, but are unrelated at the region-pair level (Figure C.8). Generally, the relationships of the within-country and cross-border GHCI to distance are largely overlapping, i.e., have significant common support, and a border effect for software developers is not clearly visible. This is due to the relatively small size of the border effect that, in fact, is statistically highly significant. In contrast, for the SCI there is a magnitudes larger and visually easily identifiable upwards shift for within-country collaborations. In line with expectations, this comparison suggests that the border effect in virtual collaboration of knowledge workers is much smaller compared to the border effect present in social networks, which is reassuring of our analysis.

Although cultural factors explain the border effect in Europe well, our parsimonious gravity model does not allow causal interpretation. Still, model fit and explanatory power point to cultural proximity as important driver of virtual collaboration. To strengthen the conjecture that culture plays an important role as deep determinant of (online) collaboration, we compare border effects in software developer collaboration for European nations and US states (Figure C.4). The idea is that there are far fewer and less pronounced cultural differences across populations in different US states than in culturally much more diverse European countries. Thus, we use the same data on regional collaboration in the US at the economic-area level from Goldbeck (2023) and estimate the state border effect using the same approach as in

Table 3.1. Table C.6 reports the results. The raw border effect, disregarding geographic factors, in the US in model (1) is 0.69 of the European estimate. Similarly, the preferred specification that takes into account size and distance in the US is 0.58 the size of the border effect in Europe, as shown by model (4). This is in line with expectations of cultural factors such as language barriers as a key determinant of the digital border effect.

Further, we assess the robustness of the coefficient estimates for the culture variables in Tables C.7 and C.8. We demonstrate that all estimates remain stable when we include various other potential control variables, e.g., regarding historical and political circumstances. Table C.7 shows robustness with respect to inclusion of contiguity, an indicator for a common border, a common control variable in gravity models that theoretically should be irrelevant in our setting. Models (2) through (7) demonstrate that all estimates remain stable when controlling for a common legal origin and shared communist history. Coefficients are similarly stable when including further control variables for political circumstances in Table C.8. For example, we account for a diplomatic disagreement score, EU membership, regional trade agreements, hegemonic relationship, relationships between monarchies as well as differences in economic and press freedom scores. Again, our coefficient estimates remain robust throughout all specifications.

In Table C.9, we examine different alternative measures for language and religion. Similarly to the trade literature (e.g., Melitz and Toubal, 2014), where continuous language proximity variables show weaker relation to trade, we find only common spoken language relevant to collaboration. Various other metrics such as other binary indicators like common native language but also continuous metrics of linguistic proximity are insignificant. This is in line with expectations that only speaking the exact same language benefits collaboration and closely related but still different languages have no impact. Model (7) in Table C.9 switches to an alternative continuous metric for religion that uses a different methodology but is also insignificantly related to collaboration. Importantly, the other coefficients remain robust and largely unchanged throughout all specifications.

3.6 Discussion and conclusion

We provide evidence of border effects in virtual collaboration that are, however, five to six times smaller compared to trade. This is consistent with trade and transportation costs being largely absent in the digital economy. The digital border effect is particularly high whenever a small country, in terms of hosted users, is involved. Generally, the remaining

border effect in software developer collaboration in Europe is entirely explained by cultural factors, especially shared interest, a common language, and history. Most other political and historical circumstances are unrelated to the digital border effect. Compared to the digital border effect at the domestic borders between US states, where cultural differences are comparably negligible, the European digital border effect is about twice as large.

This study has limitations that open up avenues for further research. Notably, our settings lacks a quasi-experimental approach where stronger identification could be achieved. Yet, already few settings exist where border effects can be estimated at all, as estimation requires domestic flow data. Opportunities to causally estimate border effects are extremely rare (e.g., Santamaría et al., 2023a). Additionally, culture evolves endogenously, which makes it hard to causally explore the intricate patterns of mediation and co-determination among the countless cultural factors. Further, our data contains information on public repositories only. While the geographical collaboration pattern is representative of the entire population of software developers (Goldbeck, 2023), it is less clear if the relationship between cultural factors and collaboration differs between open- and closed-source developers. Ideally, the measurement of culture is conducted on a more granular scale both population-wise and geographically as, e.g., software developers might be different to the general population.

Our work has several practical implications relevant to management and policy makers. Importantly, we show that there is a significant border effect for international collaboration of developers on online code repository platforms. Still, the digital border effect is much smaller compared to other outcomes, which generally points to improved feasibility of international collaboration in digital knowledge work. Since the digital border effect is entirely explained by cultural factors, they merit more attention. Together with decreasing role of geography in ICT-intensive settings of the knowledge economy this suggests that management and policy makers should shift their attention to cultural barriers to collaboration as they are relatively more important in the digital economy when fully virtual collaboration is technically possible.

4 Career Concerns as Public Good: The Role of Signaling for Open-Source Software Development

Much of today's software relies on programming code shared openly online. Yet, it is unclear why volunteer developers contribute to open-source software (OSS), a public good. We study OSS contributions of some 22,900 developers worldwide on the largest online code repository platform, *GitHub*, and find evidence in favor of career concerns as a motivating factor to contribute. Our difference-in-differences model leverages time differences in incentives for labor market signaling across users to causally identify OSS activity driven by career concerns. We observe OSS activity of users who move for a job to be elevated by about 16% in the job search period compared to users who relocate for other reasons. This increase is mainly driven by contributions to projects that increase external visibility of existing works, are written in programming languages that are highly valued in the labor market, but have a lower direct use-value for the community. A sizable extensive margin shows signaling incentives motivate first-time OSS contributions. Our findings suggest that signaling incentives on private labor markets have sizable positive externalities through public good creation in open-source communities, but these contributions are targeted less to community needs and more to their signal value.[1]

Keywords: software; knowledge work; digital platforms; signaling; open source; job search

JEL-No: L17; L86; H40; J24; J30

[1] This chapter is based on joint work with Lena Abou El-Komboz. Versions of this chapter have been published as ifo Working Paper No. 405 and CRC Discussion Paper 453. We thank Jean-Victor Alipour, Florian Englmaier, Thomas Fackler, Oliver Falck, Manuel Hoffmann, and Muhammed Yildirim as well as seminar participants at ifo Institute for valuable comments and suggestions. Further, we thank Raunak Mehrotra for excellent research assistance and gratefully acknowledge public funding through DFG grant number 280092119.

4.1 Introduction

Today's digital economy relies heavily on open-source software (OSS) (Hoffmann et al., 2024; Lifshitz-Assaf and Nagle, 2021). While the role of patents in IT decreases (see, e.g., Acikalin et al., 2022), OSS has long become an important mode of software production (Osterloh and Rota, 2007) with a 2019 investment equivalent of about 37 billion USD in the US alone (Korkmaz et al., 2024). Numerous modern products and services are built using OSS, including electronic devices, web applications, and AI algorithms. Estimates for 2022 suggest 96% of software codebases contain OSS (Synopsys, 2023). Yet, OSS is often created by a decentralized community of volunteer developers (Nagle, 2022). Because OSS is both non-rival in consumption and non-excludable due to open-source licensing (Lerner and Tirole, 2005b), OSS is a public good. This model of open community-based software development has always been "startling" to economists (Lerner and Tirole, 2002) as the motivation of individual contributors to exert private effort in order to create an openly available public good is hard to rationalize.

One potential rationale behind private contributions to OSS is it allows developers to signal valuable information and communication technology (ICT) skills to potential employers (Lerner and Tirole, 2002) since individual contributions are directly and transparently observable on online OSS platforms. Generally, ICT abilities are highly valued skills in the labor market (Draca et al., 2007; Bresnahan et al., 2002) that yield significant returns (Falck et al., 2021). At the same time, high skill obsolescence (Deming and Noray, 2020) and the inability of formal education to certify job-relevant technical skills (Fuller et al., 2022; Marlow and Dabbish, 2013) lead to information asymmetries that make it difficult for employers to assess individuals' ability. Publicly visible OSS contributions could represent a valuable signal to potential employers (Marlow and Dabbish, 2013; Long, 2009) with respect to the most job-relevant skill in software development: practical programming ability (see, e.g., Wagner and Ruhe, 2018; Surakka, 2007). This implies that, besides private benefits from learning and improved labor market outcomes, signaling activity driven by developers' career concerns might directly generate considerable positive externalities (Leppämäki and Mustonen, 2009) in the form of a public good, open source software.

In this paper, we investigate whether career concerns are indeed a driver of OSS development. To this end, we exploit variation in individual incentives to signal over time. Specifically, because signaling is costly and its value quickly depreciates, individuals economize on the signal and dynamically allocate OSS activity to times of immediate job search in order to signal skill to employers. This allows us to test for the presence of the signaling motive

empirically by studying OSS contributions of software developers who move for a job on the largest online code repository platform, *GitHub*. We focus on movers as job changes are often associated with moving (Balgova, 2020; Amior, 2019), especially for the high-skilled (Haussen and Uebelmesser, 2018; von Proff et al., 2017), which might confound our results when not explicitly considered in the empirical model. We, therefore, compare developers relocating for a new job to developers moving to a new location for other reasons in a difference-in-differences design. We argue that while job movers face elevated signaling incentives driven by immediate career concerns in the period prior to moving, the 'job search period,' these incentives are absent for developers who relocate for other reasons. Consequently, OSS activity attributable to signaling is captured by the difference in OSS contributions between job movers and other movers during relative to outside of the job search period.

Our data comprises all *GitHub* users with changing location information from ten snapshots of the *GHTorrent* database dated between 2015 and 2021. Due to this sample selection approach, we are able to capture typical volunteer developers who occasionally contribute to OSS (Vidoni, 2022). In total, our sample contains some 22,900 movers worldwide, of which around a third simultaneously change their job. Besides location and organizational affiliation, we observe in detail each user's public activity on the platform such as the monthly number of commits in open-source projects, their collaborators, or quality metrics such as stars, followers, and forks. This allows us to investigate not only whether career concerns drive OSS activity, but also if there are systematic shifts in OSS activity when motivated by signaling incentives with respect to the types of projects, usefulness to the community and quality, or user groups.

We find significantly elevated OSS activity by about 16% of job movers in the job search period compared to developers moving for other reasons. Assuming an average job tenure of three years applies to OSS developers and constant (base) activity levels over time, this translates to 6.8% of overall OSS activity being caused by signaling incentives during job transitions. Within the job search period effect size steadily decreases, consistent with stronger incentives during the application preparation phase. Notably, our analysis points to the importance of the extensive margin, inducing first-time contribution to OSS. In general, the effect derives from a broad base of job movers rather than a specific group. But we observe a larger effect for users relocating internationally and for users moving to academia. The signaling effect tends to be smaller for users with new jobs at large firms and especially at big tech companies, where we do not see a signaling effect. Multiple classifications of projects based on programming languages indicate that the effect is mainly driven by contributions to web development and data engineering projects, and to projects using top-paying programming languages.

However, signaling projects are starred and forked less by other users, pointing to a lower direct use-value to the OSS community. In general, our results are in line with career concerns motivating significantly increased OSS contributions during the job search period as we observe activity shifts to projects that increase the visibility of existing works or necessitate skills highly valued in the labor market. Additional analyses with respect to model choice and other empirical decisions emphasize the robustness and conservativeness of our preferred specification.

This work makes several contributions. In contrast to most existing studies that follow a stated preferences approach, we deploy a quasi-experimental framework and are therefore able to achieve high internal validity of our results and causally link career concerns to OSS activity under reasonable assumptions. In addition, we improve on external validity by selecting our sample from the near-universe of OSS activity on *GitHub*, the by far largest online code repository platform. Therefore our data includes not only the most active OSS developers but also volunteer developers who only occasionally contribute to OSS, but together make up the vast majority of OSS contributors. We also add to the labor market literature by showing that employees indeed signal ability through OSS activity, which groups are especially likely to signal, and how this motivation impacts the type of projects users engage in. Importantly, we contribute to the literature on private public good provision by pointing out that there are significant positive externalities from private career concerns while, at the same time, the direction of public good creation changes when labor market considerations are prominent.

Our findings have multiple managerial and policy implications. Notably, they highlight an important but neglected channel of public good creation: the positive externalities from individual labor market signaling incentives. We show that these externalities are significant with respect to overall OSS activity and signaling incentives systematically induce first-time contributions of users previously inactive in the OSS community. To increase public good creation and platform growth, both management and policymakers should take these positive externalities of career concerns into account in platform design and public policy. For example, platform design that considers the signaling needs of their users explicitly could further boost growth at the extensive (user) and intensive (activity) margin. At the same time, decision-makers should be aware of the shift in focus towards labor market requirements and away from direct use-value for the OSS community in signaling projects. For labor market and education policy as well as HR professionals, our findings point out the continued shift away from formal (public) skill certification and emphasize greater importance of more fluid and practical skill signals that directly showcase work product. Lastly, innovation policy aiming

to foster public good creation in the knowledge economy may consider maximizing positive externalities from signaling incentives, e.g. via adopting open science policies that create synergies between funded and signaling activities.

The remainder of this paper is organized as follows. First, we discuss related literature in Section 4.2. Section 4.3 introduces the data. In Section 4.4, we present the empirical identification strategy. Results are provided in Section 4.5 and Section 4.6 concludes with a discussion.

4.2 Related literature

Economics of open source. This project is related to the economics of open source. Literature in this area examines the distinct innovation model of OSS, which is based on volunteer contributions of often decentralized teams and is governed by open licenses (Osterloh and Rota, 2007; Lerner and Tirole, 2005b). As such, open innovation contrasts sharply with traditional ('closed') innovation featuring exclusive intellectual property rights (Lerner and Tirole, 2005a, 2002). These unique properties, combined with the lasting success of OSS and the growing importance of software in general, spurred dedicated research (see, e.g., von Krogh et al., 2003; Lifshitz-Assaf and Nagle, 2021). Compared to volunteer developers, firms are of less significance as in traditional innovation models, but increasingly incorporate OSS in their business models (Butler et al., 2019; Lee and Cole, 2003), for example to increase visibility (Conti et al., 2021) or learn from community feedback (Nagle, 2018). OSS research addresses a wide variety of topics such as productivity effects (Nagle, 2019), team organization (Raveendran et al., 2022; Puranam et al., 2014), geography (Wachs et al., 2022), or innovation and entrepreneurship (Wright et al., 2023; Wen et al., 2016; Colombo et al., 2014; Bitzer and Schröder, 2007).

Naturally, a large literature revolves around the reasons volunteer developers contribute to OSS and broadly distinguishes between internal factors and external rewards (Krishnamurthy, 2006; Hars and Ou, 2002). von Krogh et al. (2012) cluster motivations into intrinsic (ideology, altruism, kinship, fun), internalized extrinsic (reputation, reciprocity, learning, own use), and extrinsic (career, pay). Empirically, researchers elicit the prevalence of different motivations to contribute predominantly through surveys. These works generally find evidence for mixed motivation, but internal factors tend to be most important (von Krogh et al., 2012). For example, a survey of *Linux* contributors by Hertel et al. (2003) emphasizes the role of group belonging, identification, and a feeling of indispensability while acknowledging own use-value

as another motivator. Likewise, Stewart and Gosain (2006) show that *SourceForge* contributors are more involved because of shared values. Hars and Ou (2002) conduct an e-mail survey among OSS developers, who state that self-determination, learning, and reputation are the main reasons to contribute. Community surveys by Lakhani and Wolf (2003) and Nagle et al. (2020) explicitly stress that external and monetary factors are far less important than intrinsic motivation from creativity and intellectual stimulus. In a survey by Hann et al. (2004), *Apache* developers state own use-value, recreational value, and career impact most often as motivating factors. Gerosa et al. (2021) elicit from survey responses that reputation-building as a motive became more important in recent years, and that learning and career incentives are especially relevant for novice contributors. Shah (2006) finds motivational dynamics, where initial participation is typically driven by own use-value whereas maintainers of OSS are often intrinsically motivated. Roberts et al. (2006) note that motivations interact with each other in complex ways as, e.g., being paid increases status but at the same time is associated with a lower use-value. Indeed, Krishnamurthy et al. (2014) shows that monetary reward can crowd out other motivations. Investigating behavioral changes of developer contribution after being sponsored, both Conti et al. (2023) and Wang et al. (2022) find evidence in favor of a net-positive effect of monetary incentives on activity. Projects with fast feedback and a non-commercial nature are associated with a higher probability of receiving contributions (Smirnova et al., 2022).

Our study adds to this literature in that it broadens the scope in terms of contributors being studied. While existing work mainly focuses on the most active OSS developers, often partly paid for their work, we investigate typical users on the platform, i.e., volunteer developers who sporadically contribute to open-source projects (Vidoni, 2022). The importance of economic benefits and motives for this group of OSS contributors is neglected in the literature, and this study is among the first to study the role of career concerns in a causal identification framework. As such, it sharply contrasts with the prevailing methodological approaches used in existing research on this topic. These works are largely based on surveys, which feature the important caveat of only eliciting stated preferences as opposed to the revealed-preference approach embodied in our causal framework. As a result, we are able to make quantifiable causal claims on the importance of career concerns motive for typical volunteer OSS developers under reasonable assumptions. Our findings suggest a sizable portion of OSS activity is driven by career concerns, and that motivations dynamically change over time, which in turn alters the content of contributions.

Labor market signaling. This article focuses on one specific motivating factor to contribute

to OSS, career concerns, and therefore adds to the vast literature on signaling originating from Spence (1973). Subsequent theoretical models explicitly relate career concerns to signaling via observable effort (Holmström, 1999; Chevalier and Ellison, 1999), even when beliefs on ability are precise (Miklós-Thal and Ullrich, 2015). While basic signaling models yield separation of skill types even if signaling has no real effects, Leppämäki and Mustonen (2009) provide a model where signaling activity generates (positive) product market externalities. Empirically, Miklós-Thal and Ullrich (2016) test the career concerns hypothesis in soccer and find confirmatory results for marginal individuals. Pallais (2014) shows detailed public performance records on the online marketplace *oDesk* improved workers' subsequent employment outcomes, especially for the inexperienced. Also on an online platform for contract labor, Agrawal et al. (2016) find standardized and verifiable information important for developing-country candidates' employment probability. For software developers, Xu et al. (2020) find career concerns increase reputation-generating activity in an online community forum. Experimental evidence by Piopiunik et al. (2020) reveals basic IT skills signals in CVs on the broader white-collar labor market significantly increase the probability of receiving a job interview invitation. Apart from this causal evidence, surveys show reputation-building, signaling, and career concerns are important motivations for developers to contribute to OSS (e.g., Gerosa et al., 2021; Marlow and Dabbish, 2013; Hann et al., 2004; Hars and Ou, 2002). Similarly, employers state they regard OSS contribution as a credible and valuable signal. For example, in a survey, Long (2009) finds tech companies value OSS experience of applicants. More specifically, Marlow and Dabbish (2013) surveys recruiting managers who state *GitHub* activity is used in hiring as a signal for technical abilities and motivation, and is regarded as a stronger signal than the applicants' resume with respect to these areas. A survey among developers by Hakim Orman (2008) shows OSS activity and traditional education are seen as complements and not substitutes. However, Bitzer and Geishecker (2010) finds formally educated individuals are underrepresented in the OSS community. For developing-country candidates, Hann et al. (2013) claim that valuable OSS activity is an effective and credible signal as it is associated with significant wage premiums for *Apache* project participants. Huang and Zhang (2016) associate improved outside options from OSS signaling with job-hopping, but also acknowledge retaining effects from learning.

The contribution of this research to this strand of literature is twofold. First, in contrast to most work in this area, we follow a quasi-experimental approach using observational data from the near-universe of OSS developers. This allows us to make causal claims under reasonable assumptions leading to a comparably high degree of internal validity. Furthermore, because we are able to study a large and diverse group of OSS contributors and do not limit our scope to

the most active users, the results also feature a higher level of external validity in comparison to the fairly specific and small groups typically studied in existing works thus far. Our second contribution, which received limited attention, is asking to what degree signaling activity is wasteful or productive from a content perspective. Our empirical evidence suggests lower but still positive direct use-value for the community of signaling activity, and therefore adds an empirical perspective to the notion of positive externalities of signaling, which has only been examined theoretically to date (Leppämäki and Mustonen, 2009).

Public good provision. The paper is also connected to the broader literature on private public good provision. In contrast to traditional innovation models that rely on private property, open innovation models like OSS largely depend on voluntary contributions by individual developers and thus can be framed as private public good provision (Lerner and Tirole, 2002). Traditional theory emphasizes group size as the main factor influencing the provision of the good (e.g., Chamberlin, 1974; Bliss and Nalebuff, 1984; Palfrey and Rosenthal, 1984; Bergstrom et al., 1986; Hendricks et al., 1988; Bilodeau and Slivinski, 1996). Explicitly modeling intrinsic motivation, Bitzer et al. (2007) show provision is more likely maintained when OSS programmers value gift benefits and the intellectual challenge, have a long time horizon (i.e., are younger), are patient, face low development cost, and derive a high own use-value. In a model of OSS development, Johnson (2002) shows how own use-value considerations drive the direction of software production. Incorporating own use-value considerations and provision costs, Myatt and Wallace (2002) model a public good provision game and show multiple equilibria can arise. Ignoring intrinsic motives, Bitzer and Schröder (2005) derive joining and exiting dynamics from signaling in a model of repeated contribution. Regarding the licensing regime, Fershtman and Gandal (2004) show that contributions are higher when OSS licensing is less restrictive. Athey and Ellison (2014) model a world where OSS projects can be successful when developers are motivated by reciprocal altruism if customer support is not needed. Zeitlyn (2003) emphasizes the gift economies motivation. Empirically, O'Neil et al. (2022) define contribution territories for firms and individuals in the space of possible innovation to rationalize why certain areas are neglected. Recently, del Rio-Chanona et al. (2023) find public good generation on *StackOverflow* is impacted negatively by large language models, a substitute to online forums.

Our empirical results are important to inform on the applicability of theoretical models depending on their presumptions. Our findings emphasize that external motives are relevant and that considering the dynamic evolution of motivation is important. At the same time, external motives such as career concerns likely do not explain OSS activity entirely. Hence,

theoretical models that aim to capture OSS contribution comprehensively should consider modeling multi-dimensional motivations to contribute that include both internal and external motivations and incorporate their dynamic evolution. In general, our study emphasizes the importance of labor market incentives of high-skilled professionals for the private provision of an important public good in the knowledge economy, which likely features considerable positive spillovers both on the private market and in the form of public follow-on innovation in the OSS community.

4.3 Data

We study software developers on *GitHub*, the by far largest online code repository platform. *GitHub* was founded in 2008, reached 10 million users by 2015, and in 2021 reported 73 million users worldwide (GitHub, 2021; Startlin, 2016). Around a fifth of all code contributions on the platform are made to public repositories, i.e., open-source projects (GitHub, 2021). Repositories are maintained using the integrated version control software *git*. Importantly, the nature of the *git* version control system allows us to track each user's contribution to open-source projects over time as it records and timestamps all activity in public repositories. *GitHub* provides access to public user profiles and repositories via API. 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.[2] 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 open-source repositories. This paper relies on ten *GHTorrent* snapshots dated between 09/2015 and 03/2021.[3]

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*. Goldbeck (2023) finds no systematic bias in the location information provided on the platform, even though only a fraction of users indicates their location. We assign users to cities via exact matching to city names in the *World Cities Database*.[4] Users can also provide an indication of

[2] *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 authors. Data from the *GHTorrent* project is publicly available at `ghtorrent.org`, last accessed 02/16/2023.

[3] 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.

[4] A fraction of 0.25% of users (total: 58) are not matched to a city in the database but rather a state or a country. We do not geocode cities or states with a name that exists multiple times.

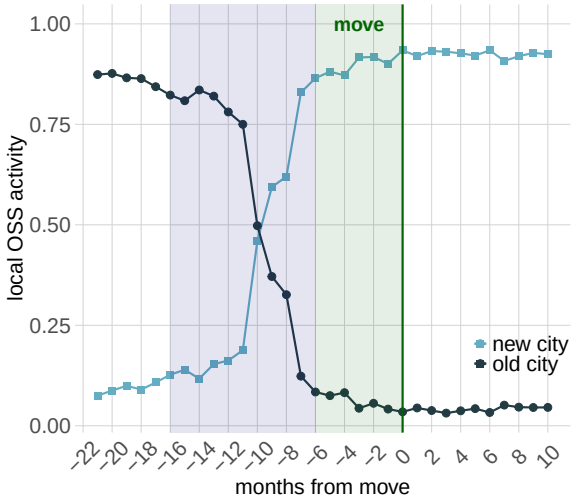
their organizational affiliation, which we use to elicit job changes. Location and organization information is observed only on snapshot dates – i.e., roughly every six months – while user activity is timestamped. We aggregate users' timestamped activity to monthly data to obtain a panel structure. Since the data is highly skewed and most users are inactive (see, e.g., Vidoni, 2022; Luca, 2015), we restrict our sample to users with an observed minimum activity of three months with non-zero commits.

Movers. From the data, we select movers, i.e., users who change their city-level location once in the observation period. Our empirical strategy elicits signaling activity from time-varying incentives around a job change. When people change jobs, they often simultaneously move (Balgova, 2020; Amior, 2019), which is especially the case among high-skilled professionals (Amior, 2015; Machin et al., 2012; Greenwood, 1975, 1973). To attain a meaningful comparison and get rid of any confounding factors associated with moving we, therefore, compare users who move for a job to users who move for other reasons. We infer the reason for moving from changes in the organizational affiliation of users. Whenever there is no affiliation change around the move date we regard a user as moving for other reasons. Conversely, if a new affiliation appears around the move date we consider a user as job mover. To implement this, we extract users' move (and job change) dates from the data.

We infer the move date from user-level location information as the month of the first snapshot with a new city indication. There is some uncertainty regarding the actual move date for two main reasons. First, users manually enter (new) location information data on the platform themselves and do this not necessarily exactly at the time of moving. On the one hand, users might be busy during the time period of moving and enter their move late. On the other hand, it might be beneficial to communicate the future location early, maybe even before actually moving, to let peers know about their relevant location as soon as possible. We empirically investigate the plausibility of the move dates attained through the snapshots by looking at team member locations in the projects a user actively contributes to each month. To this end, we assign locations to projects depending on other members' locations. Specifically, we define a user's project as localized in a particular city if the current location of more than 60% of the team members is in that city. This is only possible for a subset of projects as few members share their location and team members can be distributed. Nevertheless, it allows us to get an impression of changes in the spatial collaboration pattern of users in our sample.

Figure 4.1 plots the share of users' activity in localized collaborative projects by origin and destination city. The dark blue line represents a users' activity share in projects where team

Figure 4.1: User collaboration around relocation date



Notes: Graph shows in-sample users' commits to new- and old-city repositories as a share in users' total commits to repositories with an assigned location. Location is assigned to repositories for which at least 60% of the team members indicate a common city as current location. *Sources:* GHTorrent, own calculations.

members are predominantly located in her origin city while the light blue line represents activity in projects with team members predominantly located in the destination city. The graph shows a clear pattern. Most localized activity is in old city projects up to ten months prior to the estimated move date. This starts to reverse afterward and most localized activity is measured in destination city projects from six months prior to moving until the end of the observation period. It is plausible that users start collaborating with teams in their destination city prior to moving and activity in old-city projects fades out. Importantly, this graph shows user-provided locations systematically and meaningfully relate to collaboration patterns, which validates our measurement of moving. Similarly to the move date, we elicit job changes from users' affiliation indication as the first month the new city location is observed in the data.

Summary statistics. The resulting sample of users comprises 22,896 movers, of which 7,211 (32%) simultaneously change their job.[5] Naturally, since most registered users are inactive, this sample is very different compared to the universe of users in the data and comprises more active users, which is confirmed by the summary statistics in Table D.1. More interestingly, Table 4.1 provides an overview of our sample and compares job movers and other movers.

[5] Figure D.2 reports the moves by data snapshot and shows a similar distribution for job movers and other movers.

In general, job movers and movers are comparable in terms of activity, collaboration, and quality metrics. At the same time, there are also some differences between the groups. The median mover has five followers, contributes around 170 commits to open-source projects in the observation period, and has 15 projects with on average 2 to 3 team members. Job movers contribute a bit less to team projects and the average team size is smaller compared to other movers, and their team projects also receive fewer stars and forks. Projects in our sample are very diverse both in terms of programming languages (cf. Table D.7) and topics covered and range from web development to data engineering (cf. Figure D.5).

Table 4.1: Summary statistics

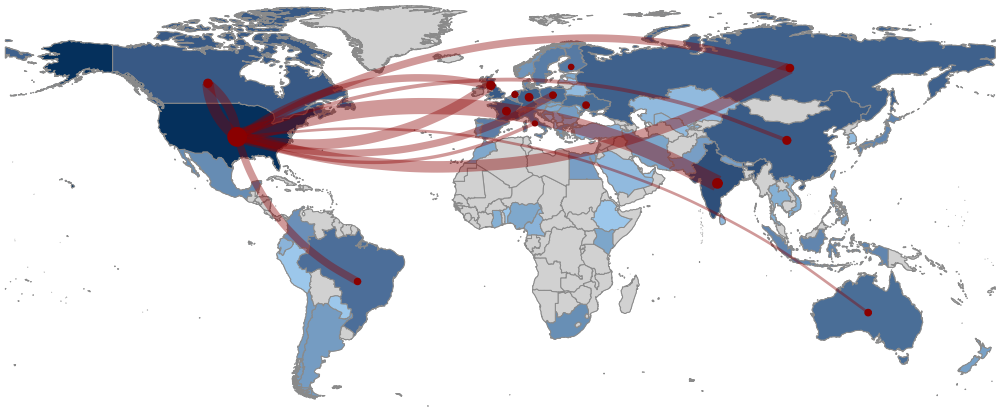
Median	**Movers**			
	job	other	Δ	%Δ
Activity				
Commits	163	188	−25	13.3%
commits single projects	*72*	*76*	*−4*	*5.3%*
commits team projects	*59*	*80*	*−21*	*26.3%*
Experience	37	42	−5	11.9%
Collaboration				
Projects	14	16	−2	12.5%
single projects	*9*	*9*	*0*	*0.0%*
team projects	*5*	*6*	*−1*	*16.7%*
Project members	2.21	2.82	−0.61	21.6%
Quality				
Followers	5	5	0	0.0%
Stars	1.10	1.88	−0.78	41.5%
stars single projects	*0.09*	*0.12*	*−0.03*	*25.0%*
Forks	0.62	1.11	−0.49	44.1%
forks single projects	*0*	*0*	*0*	*0.0%*

Notes: Experience is measured as tenure on the platform in months since the first commit at the move date. Column Δ reports the absolute difference in median between job movers and other movers. Column %Δ sets this difference in relation to other movers' median. *Sources:* GHTorrent, own calculations.

The differences between job movers and other movers regarding team project behavior is one reason why we look at single projects, i.e., projects in which only the focal user is active. But there is a more important reason derived from theoretical considerations and a practitioner's

perspective with respect to labor market signaling through OSS activity. Not all contributions to OSS communities constitute equally valuable signals of ability and thus generate reputation (Xu et al., 2020; Marlow and Dabbish, 2013). In particular, for potential employers, it is difficult and time-consuming to assess individual contributions to collaborative projects even if transparently available (Tubino et al., 2020). In contrast, single-authored projects can be assigned entirely to individual users. At the same time, quality metrics such as stars and forks make assessment effortless and enable non-software developers like HR professionals to perform such assessments. Consequently, using OSS activity in single projects as the main outcome metric ensures a close practical and theoretical relation to actual signaling potential.

Figure 4.2: Domestic and international user relocations



Notes: Blue country coloring shows the number of domestic movers after logarithmic transformation. There are 73 countries with domestic movers; grey indicates no domestic movers. The size of the red country centroids indicates the number of international moves a country is involved in. 14 countries are associated with international relocations. Red arcs represent edges in the directed country mover network, i.e., the number of international relocations from one country to another, and are scaled logarithmically. For clarity, only edges above 75 are shown. *Sources:* GHTorrent, own calculations.

Although we look at users moving worldwide, 71% are relocations to another city within the country. About 29% of relocations are international, and 19% of movers or two-thirds of international movers even move inter-continentally. This mirrors the fact that software developers are disproportionally mobile internationally (see, e.g., Adrian et al., 2017; D'Mello and Sahay, 2007; Solimano, 2006). The average relocation distance is 5,324km and there are no significant differences in these statistics between job movers and movers relocating for other reasons (cf. Figure D.1). Figure 4.2 maps the observed migration flows in our data in more detail. Countries are colored in darker blue the higher the number of domestic relocations and the width of the network edges represents the number of international relocations. The dominance of the USA as the central hub both in terms of domestic moves and as a receiving country is clearly visible even on the logarithmic scale. Domestic moves are observed most

frequently in the USA (63.5%), India (7.5%), and the United Kingdom (3.9 %). Table D.4 shows the ten countries with the most domestic moves, which account for over 90% of domestic moves and 65% of all relocations. The most important origin countries are shown in Table D.5. Table D.3 reports the ten largest origin and destination cities, which are predominantly the world's big software industry hubs, e.g., San Francisco and New York. Notably, for international relocations, we observe that users tend to move to richer countries as indicated by per capita GDP increasing on average by USD 9,780 (Figure D.3), with no systematic differences between job movers and other movers.

Users are affiliated with a diverse range of organizations. Most firms in the data are small, but the distribution is highly skewed to the right (Figure D.4). On average, each organization has four affiliated users and 23 users are affiliated with the median organization.[6] Table D.2 reports organizational affiliations and job transitions by organization type. As a consequence of the skewness, about 29% of users are affiliated with the 100 largest firms and 7.2% with the big technology firms (i.e., *Google, Apple, Meta, Amazon, Microsoft*; GAMAMs). Job transitions point out net movements towards larger, and especially big tech, firms and away from academic and small-firm affiliations. This is confirmed by Table D.6 depicting top origin and destination affiliations. While top origin affiliations include mostly students, universities, and freelancers the biggest destination shares almost exclusively are held by large software companies such as the GAMAMs or *Red Hat*, *IBM*, and *LinkedIn*.

4.4 Empirical strategy

The key idea behind our empirical model setup is to exploit temporary differences in signaling incentives across users. Specifically, we compare the activity of users who move for a job and movers who move for other reasons. The reasoning behind this is that users who move for a job experience increased incentives to signal their ability on the platform to potential employers prior to their move during the job search period, whereas movers who relocate for other reasons do not experience this temporary increase. As already discussed above, we focus on movers since job changers typically simultaneously relocate, which is widely acknowledged in the literature (Balgova, 2020; Amior, 2015) and especially the case for high-skilled professionals (see, e.g., Abreu et al., 2015; Haapanen and Tervo, 2012; Venhorst et al., 2011; von Proff et al., 2017; Kodrzycki, 2001; Ciriaci, 2014; Haussen and Uebelmesser, 2018). Thus, comparing movers leads to improved comparability as it accounts for confounding

[6] Note that these numbers are not to be confused with the number of employees since not all employees are active OSS contributors on *GitHub* and provide their affiliation.

factors associated with moving.

Figure 4.3: Adapted difference-in-differences model

From a theoretical perspective, we structure signaling incentive dynamics into three phases, where each phase is governed by a distinct incentive regime. This is illustrated in Figure 4.3. In the first phase, which we call the pre-period, an eventual mover is still working in her previous arrangement and does not actively prepare to change jobs. In this phase signaling incentives are not entirely absent and are at a normal level as there is no immediate pressure to signal skill in the labor market. In the decisive second phase, the 'job search period,' the job mover then actively searches for a new employer and prepares to relocate while movers who relocate for other reasons only prepare to relocate. In this phase, job movers face elevated incentives to signal skill to potential employers. Finally, there is a third phase after the move, which we call post-period, in which movers have relocated and the job mover has started to work for her new employer. Movers who relocated for other reasons are still with their old affiliation. In this phase, as job movers just started a new job, signaling incentives vanish and are likely even lower than in the pre-period and compared to other movers because job movers have to settle in to their new job environment, and the especially low signaling incentives.

As a result of these theoretical considerations, we expect elevated OSS activity of users who move for a job compared to users who move for other reasons in the job search period if career concerns are an important factor for OSS contribution. Additionally, we expect to see lower OSS activity of job movers compared to other movers in the post-period. We empirically investigate the dynamics of OSS activity by estimating the following baseline event study model:

$$y_{it} = \beta_1 + \sum_{j=\underline{T}}^{\overline{T}} \left[\beta_j \left(t_j \times \text{JobChanger}_i \right) \right] + \delta_i + \delta_{s(t)} + \delta_{a(i)t} + e_{it}, \tag{4.1}$$

where y_{it} is the number of commits of user i in relative-to-move month t to single-authored repositories ('signaling projects'). Note that the event study panel is balanced in the job search and pre-period but unbalanced in the post-period as some moves happen during the end of our observation period. The variable JobChanger_i indicates if user i moves for the job, i.e., simultaneously changes her affiliation and location. The core element is the interaction term

of JobChanger$_i$ with relative months to the moving month t_j. Coefficients of interest are β_j and reveal the difference in the temporal pattern of signaling activity around the move date between users who simultaneously change their job and users who do not. To control for time-constant unobserved user characteristics relevant to their level of OSS activity, we add user fixed effects δ_i. Calendar month fixed effects $\delta_{s(t)}$ account for unobserved factors affecting all users' activity in a given month. We include experience fixed effects $\delta_{a(i)t}$ to account for differences in platform tenure across users that impact OSS activity. Standard errors are clustered at the user level.

Starting from this flexible dynamic model, we adapt the standard difference-in-differences model to estimate the average treatment effect on the treated such that three phases around the move date are considered: a pre-period, a job search period, and a post-period. The reference period is the pre-period, and the temporary treatment of increased incentives to signal using OSS activity is present only during the job search period. In the post-period, signaling incentives for job changers are lower relative to the pre-period because of diminished career concerns and the new job crowding out OSS activity. The resulting model specification is

$$\begin{aligned} y_{is} = \beta_1 &+ \beta_2(\text{SearchPeriod}_{s(i)} \times \text{JobChanger}_i) \\ &+ \beta_3(\text{PostMove}_{s(i)} \times \text{JobChanger}_i) + \delta_i + \delta_s + \delta_{a(i)s} + e_{is}, \end{aligned} \tag{4.2}$$

where SearchPeriod$_{s(i)}$ is one if calendar month s falls in user i's job search period prior to the move. To account for generally reduced incentives of job switchers to make OSS contributions after the move relative to users who move for other reasons, we interact an indicator for the post-move period, PostMove$_{s(i)}$, with job changer status. The coefficient of interest β_2 captures the ATT of increased signaling incentives during the job search period, i.e., differences in OSS activity between job movers and other movers in the job search period relative to the period before. Similarly, β_3 represents the average difference in OSS activity between job movers and other movers in the post-move period relative to the pre-period.

Although the inclusion of the post-period is not formally needed for identification, we consider it explicitly in our model for two reasons. First, it adds credibility to the signaling effect estimated from the difference between the pre-period and the job search period if signaling activity declines when taking up a new job, which we assume reduces immediate signaling incentives. Second, validation of parallel trends between job movers and other movers in both the pre- and post-period helps to further assess the validity of our design. And third, although not the main goal of this analysis, estimating the effect of taking a new job on OSS

activity is interesting in itself. The three-period specification with the pre-period as reference is superior to alternatives. Taking the post-period as reference neglects the crowding-out of OSS via time constraints of formal work. Combining pre- and post-period as reference attenuates this issue, but leads to potential overestimation due to the same mechanism.

Empirical results from the event study specification guide the selection of appropriate time frames for the three phases in the ATT model. In addition, a priori theoretical and empirical considerations set our expectations. In his classical framework, Blau (1994) divides the job search period into three steps. The first step is the preparation phase, where applicants prepare their application package. Then there is the actual application step in which applicants undergo the formal application process. Finally, the third step is the decision step, in which employers and applicants decide on whether to enter an employment relationship or not. Signaling activity is expected to occur predominantly in the first step, i.e. preparation (Chamberlain, 2015). Recent statistics for the US show hiring time for complex jobs such as software development averages around four months prior to applying (Firaz, 2022), and people start thinking about and preparing for job search likely much earlier. Additionally, there is some fuzziness in our measurement of the move date due to only observing locations about every six months. Therefore, we expect to see most OSS signaling activity in the preparation phase of the job search period somewhere between six and 15 months prior to our estimated move date.

Note that our model specification provides a conservative and incomplete estimation of the role of career concerns for individual OSS activity for multiple reasons. First, signaling incentives are not entirely absent in the pre-period. Career concerns are not binary and we exploit time variation in their strength rather than presence or absence. Second, our estimates are downward biased due to measurement error when some control group movers in fact move for the job, as well, but do not change their affiliation. Third, our focus on movers implies we study a group of users who face significant additional time constraints relative to users who are not relocating and therefore trade-off their time allocation between more activities, potentially leading to less time spent on signaling activity in this group. Finally, the dynamics within the job search period as well as the fact that towards the end of our signaling period, the share of users who already found a job increases biases the ATT downward. Consequently, our estimates should be interpreted as a lower bound to the importance of career concerns for OSS activity.

Our key identifying assumption is that in the absence of signaling incentives for job changers,

their activity would have evolved similarly to movers not changing jobs simultaneously, conditional on controls. Although we cannot test this assumption directly we assess it by showing parallel trends in periods when signaling incentives are absent, i.e., both the pre- and post-period. The main remaining threats to our identification strategy are factors unrelated to signaling incentives that affect the user activity of job changers in the job search period prior to the move but not the user activity of movers that do not change jobs or vice versa. One such concern could be due to potentially reduced work ethic of job movers in their old job as it comes to an end and, as a consequence, more time for side projects. However, one could also expect the old job claims more time towards the end as, e.g., projects have to be handed over. Another potential concern is an increased prevalence of learning motives during periods of unemployment between two jobs. This is, however, not only unlikely due to generally short unemployment spells for IT professionals; the median duration of unemployment in the US, for example, is only eight weeks.[7] It is especially unlikely given that our design focuses on movers, and relocating to another city or even country is generally time-consuming and stressful. Nevertheless, in Sections 4.5.2 and 4.5.3 we address these concerns and assess related channels by investigating the kind of OSS activity of job movers and how it differs from other movers to validate if the observed activity can likely be attributed to signaling or not.

4.5 Results

4.5.1 Main effect

Figure 4.4 plots the event study coefficients for user activity around the relocation date resulting from the model in Equation 4.1. The dynamics are consistent with signaling as a driver of OSS activity and the hypotheses derived from our theoretical considerations. In the pre-period, there are no statistically significant differences in OSS activity between users who move for a job and users who move for other reasons. Similarly, after moving we observe a lower activity level for job movers compared to other movers but the dynamic development is, again, parallel to each other. This absence of differential trends between treatment and control group users is reassuring of the validity of our empirical design as it provides confidence that our key identifying assumption holds. Importantly, during the period prior to moving, OSS activity of job movers is significantly elevated relative to other movers conditional on controlling for time, user, and experience fixed effects. We claim this increase is driven by immediate career concerns in the period of job search which incentivizes signaling activity.

[7] Statistic retrieved from *Bureau of Labor Statistics* based on the *Current Population Survey* 2018: `https://www.bls.gov/web/empsit/cpseea37.htm`, last accessed on 11/10/2023.

Figure 4.4: Event study estimates

Notes: Estimates for $t_j \times \text{JobChanger}_i$ based on Equation 4.1 with user, experience and calendar month fixed effects. The outcome is IHS-transformed commits to single-authored projects. The reference month is $t = -16$. Bars show 95% confidence intervals. Standard errors are clustered at the user level. *Sources:* GHTorrent, own calculations.

The dynamic activity pattern during the job search period is consistent with signaling behavior, too. Signaling activity is strongest at the beginning of the job search period 10 to 14 months before the move month with activity in signaling projects being elevated by up to 24.5% for job movers. The effect then declines steadily to substantially lower levels before the move date around 6-10% before returning to a permanently lower, stable, activity level from the move month onward, with estimates centering around -7 to -10%. Model (3) in Table D.13 provides estimates for each period. This pattern is in line with our theoretical considerations predicting more intense signaling in the preparation step of the job search period as users generally have an incentive to have their signal ready by the time of application which is likely earlier in the job search period. In addition, more and more users finding a job during the job search period or moving earlier than the observed move month, both leading to reduced incentives to signal.

Because of sparsity, we transform the dependent variable using the inverse hyperbolic sine transformation in order to retain zero-valued observations (Bellemare and Wichman, 2020). At the same time, this transformation approximates the natural logarithm and is commonly interpreted in a similar way (Burbidge et al., 1988; MacKinnon and Magee, 1990). As our data typically features right-skewed but low numbers of commits, we do not rescale the dependent variable prior to transformation. Estimates are generally sensitive to scaling and

as there is no overarching guideline, scaling choice is described as a data fitting problem in the econometric literature (Aihounton and Henningsen, 2021). As rescaling typically leads to larger estimates our choice with respect to dependent variable scaling is conservative (Chen and Roth, 2023).[8] The effect size of the resulting coefficient estimates thus is not only statistically highly significant but also economically sizable as we estimate between 5 and 25% higher OSS activity of job movers compared to other movers in the job search period, depending on the month relative to move date.

Table 4.2: Difference-in-differences model

IHS(single commits)	(1)	(2)	(3)
Job mover × job search	0.3621***	0.2962***	0.1646***
	(0.0137)	(0.0144)	(0.0141)
Job mover × post move	-0.2608***	-0.2208***	-0.1036***
	(0.0189)	(0.0203)	(0.0190)
User FE	×	×	×
Month FE		×	×
Experience FE			×
Adjusted R^2	0.289	0.308	0.359
Observations	1,946,413	1,946,413	1,946,413
Users	22,896	22,896	22,896

Notes: Results from estimation of Equation 4.2. experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$. *Sources:* GHTorrent, own calculations.

The dynamic event study specification validated by theoretical and empirical evidence from the literature informs our definition of the job search period. We identify the period of distinctly elevated OSS activity in the 15 months prior to the month of moving as job search period. Using this definition of the job search period allows us to estimate the average treatment effect on the treated (ATT) per Equation 4.2. Table 4.2 provides the ATT estimates of our adapted three-period difference-in-differences specification. As expected, job movers OSS activity is elevated during the job search period relative to other movers and is lower in the post period. The inclusion of calendar month and experience fixed effects considerably improves model fit as described by adjusted R^2. The coefficient(s) of interest are attenuated as a result. Our preferred specification in model (3) estimates that job movers contribute about 16.5% more

[8] We discuss model specification in more depth in Section 4.5.3.

on average in the job search period compared to other movers.

While the ATT effect size as such is suitable in assessing the importance of signaling incentives for individuals' OSS contributions during a job transition, we are further interested in the broader relevance of this motivation for the OSS community. Because our definition of the job search period is broad and includes periods with only moderately elevated signaling incentives, the ATT is best interpreted relative to the length of the job search period by performing a back-of-the-envelope calculation. Recent statistics state average job tenure in the US is around four years and only two years for software developers (Firaz, 2022). Assuming an average job tenure of three years applies to OSS developers, constant (base) activity levels across users and over time, and using our estimates ATT coefficient implies 6.8% of overall OSS activity is caused by signaling incentives during job transitions.[9] This suggests career concerns are a significant motivation for software developers and causes a sizable portion of contributions to OSS.

4.5.2 Heterogeneity

A natural question that arises from our main finding is whether there are systematic shifts in job movers' OSS activity during the job search period. This not only improves our understanding of how the signaling motive impacts users and activities differently but provides further validation of the signaling as the motive behind increased OSS activity. In particular, we explore two main dimensions of heterogeneity. First, we ask if job movers systematically focus their OSS activity during the job search period on certain types of projects, e.g., projects that are especially valuable as signal in the labor market. Second, we investigate if particular groups of job movers exhibit significant differences in effect size or if the effect size derives from all job movers equally.

We investigate effect heterogeneity with respect to the type of projects users contribute to during the job search period in Table 4.3. For this purpose, we use information on the main programming languages of projects and classify them into categories to distinguish broad project types. Our classification is documented in Table D.7 in the Appendix. This project-level approach requires using the number of contributions to each project type as outcome variable in user-level regressions. Thus, we run separate regressions of the model in Equation 4.2 for each project type. Results show significant differences in the ATT effects.[10] Notably, we

[9] Calculated as: $\hat{\beta}_2 * \frac{\#months_{JobSearch}}{\#months_{JobTenure}} = 16.46\% * \frac{15}{36}$.

[10] Note that increased sparsity leads to a loss of quantitative comparability to the main results in favor of comparability between project-type regression estimates.

obtain the largest effects for web development and data engineering projects. Low-level programming, program routine, and app development projects experience much smaller increases in the job search period. These results are consistent with, first, job movers focusing on web development because such projects are a way to showcase their work product and thus skill in existing works. Secondly, job movers might signal more through data engineering projects as skills related to such projects are especially valuable in the labor market.

To investigate the second channel in more detail, we classify programming languages directly by their valuation in the labor market as stated in the *StackOverflow* list of top-paying technologies.[11] Using the same method as above, we compare the ATT for programming languages listed as top-paying technologies compared to non-listed programming languages. Among top-paying programming languages, we further separate the top 30 best-paying from other listed programming languages. Which languages are in each category is shown by Table D.8 in the Appendix. According to survey evidence by *StackOverflow*, programming languages in the best-paying category are associated with about USD 16,500 higher total annual compensation compared to other listed languages, a 24% premium. Table 4.4 displays the estimation results. While job movers significantly increase OSS activity during the job search period in all groups, the increase is by far the largest for the best-paying programming languages. Compared to the increase in languages lower on the list, the increase in OSS activity in projects using best-paying programming languages is about twice as large. The effects in the other two categories are not statistically distinguishable. This provides further indication that job movers focus their signaling activity on projects requiring skills especially valuable in the labor market.

[11] Available at: `https://survey.stackoverflow.co/2023/#technology-top-paying-technologies`, last accessed on 11/03/2023.

Table 4.3: Heterogeneity by project type

IHS(single commits)	(1) low-level	(2) data eng.	(3) app dev.	(4) web dev.	(5) routine	(6) other
Job mover × job search	0.0136**	0.0426***	0.0256***	0.0607***	0.0277***	0.0353***
	(0.0061)	(0.0082)	(0.0051)	(0.0109)	(0.0073)	(0.0072)
Job mover × post move	-0.0047	-0.0177*	-0.0068	-0.0852***	-0.0145	0.0015
	(0.0077)	(0.0107)	(0.0077)	(0.0144)	(0.0098)	(0.0089)
User FE	×	×	×	×	×	×
Month FE	×	×	×	×	×	×
Experience FE	×	×	×	×	×	×
Adjusted R^2	0.26051	0.26955	0.29500	0.28444	0.28765	0.33629
Observations	1,946,413	1,946,413	1,946,413	1,946,413	1,946,413	1,946,413
Users	22,896	22,896	22,896	22,896	22,896	22,896

Notes: Results from estimation of Equation 4.2 with IHS-transformed number of commits to single-authored projects featuring main programming language of the respective class. Classification of programming languages according to Table D.7. Experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$. *Sources:* GHTorrent, own calculations.

Table 4.4: Heterogeneity by labor market value

IHS(single commits)	listed		
	(1) top 30	(2) other	(3) not listed
Job mover × job search	0.0842***	0.0456***	0.0396***
	(0.0095)	(0.0104)	(0.0076)
Job mover × post move	-0.0181	-0.0703***	-0.0165*
	(0.0126)	(0.0132)	(0.0094)
User FE	×	×	×
Month FE	×	×	×
Experience FE	×	×	×
Adjusted R^2	0.23914	0.24635	0.27395
Observations	1,946,413	1,946,413	1,946,413
Users	22,896	22,896	22,896

Notes: Results from estimation of Equation 4.2 with IHS-transformed number of commits to single-authored projects featuring main programming language of the respective class. Classification of programming languages according to Table D.8. Experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$. *Sources:* GHTorrent, own calculations.

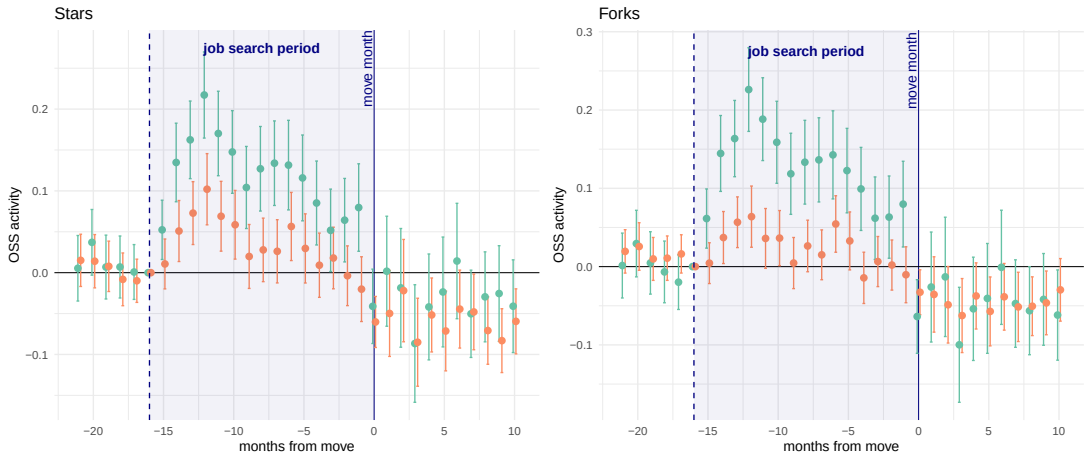
As an alternative method to classify projects, we tap project descriptions and deploy a keyword-based NLP approach (Gentzkow et al., 2019). Only about one fourth of projects in our sample have descriptions and descriptions are typically brief. Therefore, we use a bag-of-words representation of all project descriptions and create a list of keywords associated with five project categories (education, data(base), website, code, and files) from analyzing the most frequently appearing words.[12] We then assign projects to a cluster when their description contains at least one associated keyword.[13] This approach naturally results in a smaller sample due to few project with description and strict requirements from the keyword list. Yet, using appropriate keywords is a targeted approach and increases the confidence in our classification. Estimating our baseline model for commits to the project types generated with this method yields similar results, reported in Table D.12. We obtain the largest effect for coding projects, followed by files and websites. These findings are generally in line with the programming language-based approach. Notably, we find no effect for educational projects,

[12] The keywords are reported in Table D.9.
[13] As a result, projects may be assigned to multiple clusters.

consistent with signaling rather than learning motives.

To distinguish whether career concerns induce job movers to start contributing to OSS, we formulate the model as a linear probability model (LPM) with an indicator for contribution rather than the number of contributions as recommended in Chen and Roth (2023). Estimation results are shown in Table D.10 and suggest a 7% higher probability of job movers contributing during the job search period relative to other movers. To investigate the extensive margin further, we run our baseline event study model using contributions to new projects, defined as projects initiated (i.e., first commit date) during the month under consideration and compare new single projects to new team projects. Results in Figure D.7 show that job movers especially start working on new single projects during the job search period. Together, these findings suggest the extensive margin plays a significant role, and job movers specifically engage in OSS activity that is unambiguously attributable to themselves, which is advantageous in order to signal personal ability.

Figure 4.5: Heterogeneity by community use-value



Notes: Estimates for $t_j \times \text{JobChanger}_i$ based on Equation 4.1 with user and calendar month fixed effects. The outcome is IHS-transformed commits to single-authored projects with (orange) or without (green) stars (left) or forks (right), respectively. The reference month is $t = -16$. Bars show 95% confidence intervals. Standard errors are clustered at the user level. *Sources:* GHTorrent, own calculations.

When thinking about the relevance of OSS contributions spurred by career concerns as a public good, quality is an important factor. On *GitHub*, projects may receive stars and can be forked by other users on the platform. Stars are a way for other users to indicate they find the project useful and to bookmark them for future reference. Forking refers to a process that copies a project into a new repository of the forking user so that she can use and alter the code in her own projects. Forking thus indicates other users' interest. We use both quality indicators and estimate the event study model, differentiating between OSS activity in projects with

and without stars or forks, respectively. Figure 4.5 depicts the results and shows most OSS contributions of job movers during the job search period are in low-quality projects. This implies other users do not find signaling projects immediately useful. However, we found before that many signaling projects are websites that likely do not contain new code but rather showcase existing work more clearly. Such repositories are rarely starred or forked since usage is mostly off-platform. This might explain why the selected quality indicators suggest low quality and does not necessarily mean that projects are perceived as not valuable. Rather, the value could lie in making existing works more visible and accessible to the community. Nevertheless, these findings do suggest a lower direct use-value of signaling projects for the OSS community regarding the usefulness of code in other projects on the platform.

Table 4.5: International relocations

IHS(single commits)	international		upward moves	
	(1) international	(2) inter-continental	(3) income group	(4) GDP p. c.
Job mover × job search	0.1461***	0.1472***	0.1620***	0.1625***
	(0.0158)	(0.0150)	(0.0146)	(0.0144)
Job mover × job search × indicator	0.0619**	0.0923***	0.0295	0.0450
	(0.0260)	(0.0313)	(0.0393)	(0.0452)
Job mover × post move	-0.1040***	-0.1038***	-0.1038***	-0.1038***
	(0.0190)	(0.0190)	(0.0190)	(0.0190)
User FE	×	×	×	×
Month FE	×	×	×	×
Experience FE	×	×	×	×
Adjusted R^2	0.35948	0.35949	0.35945	0.35945
Observations	1,946,413	1,946,413	1,946,413	1,946,413
Users	22,896	22,896	22,896	22,896

Notes: Results from estimation of Equation 4.2 adding a triple interaction which features an indicator variable to separate heterogeneous effects of interest. Upward income group moves are defined as moves from developing to developed countries. Upward moves in GDP per capita are based on current 2021 PPP USD. Experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$. *Sources:* GHTorrent, World Development Indicators, own calculations.

Labor market signaling via OSS activity might be valuable to a different extent for job movers. We, therefore, investigate whether the effect is broad-based among all users or driven by a

group of users with a particularly large increase in OSS activity during the job search period. For this purpose, we first explore heterogeneity with respect to followers comparing quartiles and find no significant differences (cf. Figure D.6). Second, we investigate whether signaling activity differs for users moving internationally by interacting dummy variables for types of moves to our baseline model. The results are reported in Table 4.5. Model (1) indicates that users moving internationally engage in 42% more labor market signaling via OSS compared to domestic movers. Likewise, inter-continental job movers signal even more and feature a 63% higher effect compared to non-intercontinental movers as shown by model (2). Models (3) and (4) suggest that the effect differences are especially driven by international movers relocating to higher-income countries, though the coefficients lack statistical significance. These results are in line with existing evidence (e.g., Agrawal et al., 2016; Hann et al., 2013) suggesting that OSS signals could substitute formal certification, which is less transferrable and accepted internationally, particularly for developing countries.

Table 4.6 shows that there is some heterogeneity in signaling activity depending on users' origin (old) and destination (new) affiliation. Importantly, users who obtain new jobs at big tech firms do not engage in labor market signaling through OSS activity to a significant extent. In contrast, users changing jobs to academic affiliations signal significantly more. There is no statistically significant difference in signaling activity depending on the old affiliation, but an economically significant point estimate for above-median firm size points towards more signaling activity by users coming from larger firms. These results, though weak, are consistent with an arguably generally greater role of open source in academia while large corporations like the big tech firms emphasize proprietary software more, and users qualified for a job at the big tech firms typically do not need (additional) ability signals from OSS activity as they tend to have the highest credentials anyways.

4.5.3 Robustness

We choose a model that uses the inverse hyperbolic sine (IHS) transformation of the outcome variable as the preferred specification, which has the mentioned advantages of retaining zeros while approximating the logarithmic transformation (see, e.g., Bellemare and Wichman, 2020; MacKinnon and Magee, 1990; Burbidge et al., 1988). A related and widely-used transformation is the logarithmic transformation and specifically $\log(y+1)$ (Bellégo et al., 2022). The challenge with these transformations is that they are scale-dependent, but this problem is more severe for high-valued and sometimes-zero outcomes (Mullahy and Norton, 2022; Chen and Roth, 2023). Aihounton and Henningsen (2021) frame scaling as a data fitting exercise. Since our data is low-valued and sparse, we opt for a conservative quantitative interpretation arising

Table 4.6: Heterogeneity by affiliation

IHS(single commits)	destination			origin	
	(1) median	(2) big tech	(3) academia	(4) median	(5) academia
Job mover × job search	0.1784***	0.1753***	0.1578***	0.1631***	0.1601***
	(0.0198)	(0.0144)	(0.0145)	(0.0142)	(0.0502)
Job mover × job search × indicator	-0.0219	-0.1460***	0.0930**	0.0843	-0.0114
	(0.0234)	(0.0480)	(0.0457)	(0.0999)	(0.0652)
Job mover × post move	-0.1038***	-0.1042***	-0.1032***	-0.1040***	-0.1693***
	(0.0190)	(0.0190)	(0.0190)	(0.0190)	(0.0528)
User FE	×	×	×	×	×
Month FE	×	×	×	×	×
Experience FE	×	×	×	×	×
Adjusted R^2	0.35946	0.35950	0.35947	0.35946	0.36126
Observations	1,946,413	1,946,413	1,946,413	1,946,413	1,406,169
Users	22,896	22,896	22,896	22,896	22,896

Notes: Results from estimation of Equation 4.2 adding a triple interaction which features an indicator variable to separate heterogeneous effects of interest. Median split refers to median size of affiliation in terms of users in the full *GHTorrent* sample. Big tech refers to Google, Amazon, Meta, Apple and Microsoft. Academia refers to students and university affiliations. Specifically, users stating *university, college, institute, universiteit, universidad, universität* or *student* in their affiliation are assigned to academia. Destination (origin) refers to users' affiliation before (after) the affiliation change. Experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$. *Sources:* GHTorrent, own calculations.

from IHS transformation of the unscaled dependent variable. Another class of alternative models are Poisson models such as the PPML estimator. These models are the established go-to choice in trade (Larch et al., 2019) and other applications with high-valued count data featuring zeros such as investment, profit, or revenue data (Cohn et al., 2022). However, these models perform poorly in practice on low-valued sparse panel data such as ours and there is no standard econometric approach yet. Additionally, our data features sparsity not only across units but also within. For such applications, IHS or logarithmic transformations are the preferred choice in practice, e.g. in Xu et al. (2020) or Bahar et al. (2022).

Apart from being conservative in our preferred model specification, we assess the robustness of our results by estimating several alternative models. Results are reported in Table D.10

in the Appendix. First, we show that the most widely-used alternative way to transform the dependent variable in similar applications (e.g., Xu et al., 2020), a logarithmic transformation, yields similar coefficient estimates. Second, we run two types of frequently used count data models: a negative binomial and a Poisson fixed effects model. Both models are known to frequently exhibit performance issues with fixed effects and convergence issues (Bellégo et al., 2022; Correia et al., 2019). The PPML model results in similar coefficient estimates for the job search period and an increased estimate for the post-period. The negative binomial model estimates are significantly inflated by a factor of three to four compared to our preferred specification. These findings indicate the robustness of our results with respect to model specification and confirm that our estimated effect size is conservative. Furthermore, we follow state-of-the-art best practices (Chen et al., 2022) in that we explicitly consider intensive and extensive margin effects. The formulation of our model as LPM suggests reasonably high importance of the extensive margin (see model (3) in Table D.10). Note that through our sample selection of active OSS contributors only, extensive margin effects are likely downward biased. At the same time, this implicit conditioning decreases potential bias of the intensive margin in our main specification (Hersche and Moor, 2020).

Measurement error in the move date possibly introduces bias in our estimates due to observing location data only every six months and users entering their new location after relocation. The event study results in Figure 4.4 partly alleviate this concern as there is a discontinuous drop in OSS activity of job movers at the proxied move date. Nevertheless, it is unclear whether the downward trend during the job search period is due to already-moved job movers still in the treatment group or, e.g., due to decreased signaling incentives of users who already found a job. We address this by varying the job search period definition and separately estimating a coefficient for the period for which we are unsure if the user actually already moved. This adjustment generally increases the estimated effect by up to three percentage points to about 19.5%. Note that although this introduces upward bias in our estimates it simultaneously alters the length of the job search period and, as a result, leads to a mechanic downward adjustment in the interpretation when thinking about overall OSS activity attributable to career concerns.

Our approach exploits the specific timing of elevated career concerns during the job search period. Still, coinciding increases in other motives are a potential concern. Specifically, if people disproportionately learn new skills in between jobs and this activity is conducted in public repositories on *GitHub*, our model would wrongly attribute such activity to career concerns. One of our project types in the keyword-based classification are educational

projects. This category captures repositories associated with coursework, assignments, or online education (e.g., *Coursera*). Table D.12 shows no effect on the activity in educational projects, suggesting that activity driven by learning motives does not drive our effect. In addition, we investigate projects not owned by the mover, such as company projects, or projects consisting of initial forks (a copy of existing repositories). We find no evidence for a significant relevance of these channels (see Table D.11).[14]

For completeness, we report estimation results for the event study specification in Table D.13 and, similarly as in Table 4.2 for the ATT, show the results for the models without experience and calendar month fixed effects, as well. Figure D.8 plots event study coefficients for variations of the baseline model. Further, we establish the robustness of our results to alternative sample definitions with respect to geocoding and job changes in models (3) and (4) of Table D.10. For user-level heterogeneity analyses using interaction terms, alternative model specifications based on separate regressions with redefined outcome variables similar to the project-derived heterogeneity analyses (Tables D.15, D.16, and D.17) show qualitatively similar results.

4.6 Conclusion

We show private career concerns of software developers induce significant contributions to open-source software, a public good. By exploiting temporal variation in signaling incentives in a quasi-experimental design, we establish a causal increase of OSS activity of job movers compared to users relocating for other reasons in the job search period by about 16%. These positive externalities of labor market signaling are sizable from both the individual and the community perspective but often neglected in existing works that predominantly emphasize other motives to contribute to OSS development. A broad base of users on the largest online code repository platform, *GitHub*, engages in labor market signaling during the job search period and signaling opportunity even attracts first-time contributors. OSS activity driven by signaling motives is disproportionately directed to projects that increase external visibility of existing works or are written in programming languages highly valued in the labor market. At the same time, signaling projects are starred and forked less by other users on the platform. This suggests OSS activity induced by career concerns is targeted less to the direct use-value of the OSS community and more to their value as a labor market signal.

[14] Note that project ownership is prone to measurement error, as it might wrongly capture the same individual as distinct persons, e.g., when committing to projects using two different e-mail addresses as identification or using multiple devices. Thus, it is not surprising that there is a small significant effect for non-own projects in Table D.11.

Our study has limitations. Data does not contain information on users besides activity on the platform, location, and affiliation and cannot be linked to other data on the individual level, which constrains the number of possible heterogeneity analyses. Furthermore, location and affiliation changes are only observed at snapshot frequency, i.e., roughly every six months. This leads to blurriness in the proxied move (and affiliation) change months and likely biases our estimates downwards. In general, we opt for a conservative model specification as a quantitative interpretation of our effect size depends on econometric choices regarding model class and outcome scaling and transformation. It should also be noted that although our empirical strategy identifies the causal effect of temporarily elevated signaling incentives under reasonable assumptions, it by no means captures all OSS activity attributable to labor market signaling and therefore should be interpreted as a lower bound estimate. Similarly beyond the scope of this work is to assess the extent to which OSS signals improve individual-level labor market outcomes.

Despite these limitations, our findings have several managerial implications. Importantly, decision-makers aiming to increase OSS activity should take into account career concerns as a significant motivating factor for developers. Platform design addressing the signaling needs of users explicitly might grow the platform at both the intensive (activity) and the extensive (users) margin. Measures that foster public visibility, transparency as well as accessibility for non-experts might contribute to this goal, e.g., through easily understandable activity metrics, skill badges, or lists of spoken programming languages on user profiles. At the same time, platform managers should be aware that signaling motives might steer OSS activity towards projects with lower direct use-value for the community whenever there is a gap between signaling value and community value of projects. For hiring managers, our results emphasize that OSS is a commonplace and potentially valuable signal of skill for developer talent. Consequently, it should receive attention in employee search and assessment.

Finally, our study provides several insights for public policy. In general, the positive externalities of career concerns on public good creation merit attention due to likely significant positive spillovers of OSS on the private sector and innovation. Innovation policy that enables and encourages publicly funded software development to be hosted and shared on online open-source platforms may increase the motivation of the funded developer teams while at the same time generating OSS, a public good that potentially spurs further innovative activity. With respect to labor market and educational policy, our results point to the continued shift away from (public) skill certification in occupations related to software development and emphasize a greater role of more fluid and practical skill signals directly showcasing work

product. Educational institutions should acknowledge both the labor market value of OSS activity for their students and the positive societal externalities from such activity and consider encouraging students to engage in OSS development or even explicitly integrate OSS projects into curricula.

Appendices

A Supplementary Materials to Chapter 1

A.1 Supplementary Information

A.1.1 Country example: Benin

To understand national backbone rollouts in SSA countries in more detail, we describe the case of Benin as a typical example. Benin is one of the countries connected via the SAT-3 SMC, which brought an international connection of 45 Mbps (Chabossou, 2007). The rollout of the national backbone was planned by Benin Telecoms SA, the fixed-line monopolist who manages the gateway to the national Internet, operates as the national carrier, and administers the national domain (*.bj). Benin Telecoms SA is state-owned and offered permanent ADSL connections with up to 2 Mbps at the time (Agyeman, 2007).

Infrastructure rollout. According to Chabossou (2007), the SAT-3 SMC landed in Cotonou, Benin's largest city, the seat of government, and located 40 kilometers away from Benin's official capital, the much smaller city of Porto-Novo. Close by, in Abomey-Calavi, Benin's largest digital hub is located as well. Together with Godomey, these cities form the largest agglomeration and metropolitan area in Benin with nearly 2.5 million inhabitants, which represents about a third of Benin's population. From there, first, a connection to Parakou with a 425 kilometers optical fiber cable was constructed in 2001. Parakou is Benin's next largest economic center with more than 150,000 inhabitants in the 2002 census and the capital of the Borgou department. This connection was constructed along Benin's railway line and roads network (Figure A.14). On its way, the backbone cable connected smaller, more remote towns such as Savalou with 30,000 inhabitants. The next national backbone connection was established between Parakou and the borders to Niger, in the north-east, and Burkina Faso, in the north-west. These connections were constructed along the road network and transformed Benin to a sub-regional digital hub interconnecting Togo, Nigeria, Burkina Faso, and Niger. Until 2001, only the first kilometers of the fiber-optic backbone and access points were under construction. 2001 was the year of most active national backbone development in Benin. Benin Telecoms SA's infrastructure investment peaked in 2001, with more than USD 80 billion. The connection to Burkina Faso and Togo was constructed through Natitingou, the capital of the Atakora department. Again, on-route remote towns like Kandi or Djougou were connected incidentally. Only later, during the construction of cross-links in the national backbone, further rural towns were connected. Cross-links are often added to hub-

and-spoke networks to increase network resilience and reliability through redundancies. In Benin, remote towns like Nikki, Ségbana, and Banikoara benefited from incidental connection through cross-links.

Internet use. In Benin, Benin Telecoms SA owns the transmission monopoly. Benin Telecoms SA, at the time, offered data transmission packages mostly to commercial clients (banks, hotels, ministries, etc.). The broader population mostly accessed the internet through cybercafés in the 2000s (cf. Section A.1.2). The number of cybercafés grew exponentially with internet infrastructure rollout in Benin and reached several thousands. In contrast to international institutions, universities, or major corporations, private individuals typically do not have home access (Chabossou, 2007). Still, in 2006 only 25 percent of Benin's population had used the internet at least once. Access is mainly at cybercafés (21 percent) or at the workplace (2.2 percent), while internet at home remains expensive (Ahoyo, 2006).

A.1.2 Cybercafés and 'last mile' technologies

As in most developing countries, internet in SSA countries before the era of smartphones was largely accessed through cybercafés (see, e.g. Osho and Adepoju, 2016; LeBlanc and Shrum, 2017; Southwood, 2022), especially in the rural areas (Williams et al., 2012). Cybercafés (also: internet cafés; or just: cyber) in rural SSA are community-based internet centers typically in the form of small shops or rooms with one or two computers with internet access (see, e.g. LeBlanc and Shrum, 2017; Mbarika et al., 2004), though cybercafés were sometimes (much) larger in cities (LeBlanc and Shrum, 2017). The photograph in Figure A.3 shows an example of a rural cybercafé. Cybercafés represented the first experience of going online for most people in SSA who used the internet during the 2000s and early 2010s (Lubwama, 2023) and became hubs for communication, research, and online entertainment (Kitimbo, 2023). Alternative (public) access points like libraries or telecenters were relatively rare (Gomez, 2014). In cybercafés, internet access is sold at pre-paid hourly rates.[1]

Other 'last-mile' technologies at the time offered only unstable connection and were limited and prohibitively expensive. Dial-up in via 56k modems is only possible in locations connected to the telephony network and therefore mostly restricted to selected neighborhoods or places in larger cities (Gitta and Ikoja-Odongo, 2003). In 2004, average costs of a dial-up internet account for 20 hours a month in Africa were prohibitively expensive for most households with around USD 68 per month (Mbarika et al., 2004). Internet connection via satellite (e.g., Very

[1] Southwood (2022) estimates hourly rates of 1-2 USD in cities around 2000, much cheaper than alternatives. Prices came down quickly with higher international bandwidth, increasing competition, and improved infrastructure (World Bank, 2016).

Small Aparture Terminals; VSAT) was even more costly while providing less stable connectivity, although available independent from telephony networks (McKague et al., 2009; Nyezi, 2012; Byanyuma et al., 2013). In contrast, cybercafés have wired connections to the national backbone providing reliable signal at relatively high speed (LeBlanc and Shrum, 2017). Even when mobile internet became available around 2010, at first internet access on personal devices remained much more expensive compared to cybercafés (LeBlanc and Shrum, 2017).

In the 2000s cybercafés quickly became places to interact and exchange information with the outside world (Mbarika et al., 2004) as they provide affordable, immediate and convenient access to the internet (Osho and Adepoju, 2016). Advantages of community-based internet access via cybercafés was that people could learn how to use the internet from other users at the café or share the hourly rates (Southwood, 2022). Users of cybercafés generally constitute a diverse group, although with a bias towards younger populations, especially educated males and local elites (Mwesige, 2004; Gitta and Ikoja-Odongo, 2003). Low-speed internet at 0.5-2 Mbps available in the 2000s allowed basic functionality such as web browsing, e-mail, and chat messaging but not video streaming or other data-intensive activities. In a 2003 survey in Uganda, users indicated the purposes of their internet use in cybercafés is communication via e-mail (89%), research (32%), entertainment (30%), education (27%), or sports and news (24%); a quarter of respondents indicated using the internet for trade and commerce (Gitta and Ikoja-Odongo, 2003). According to Williams et al. (2012), cybercafés are particularly important for rural internet access in Africa as they benefit small-scale knowledge-based businesses such as call centers, engineering companies, farmers, and other local firms relying on outside information. Similarly, Mbarika et al. (2004) acknowledges the role of cybercafés in Sub-Saharan Africa in maintaining business contacts. This is confirmed by ample anecdotal evidence. For example, in a blog post Ndiomewese (2015) writes:

> "Those days [early 2000s], you could almost certainly stroll into a cybercafé and meet the MD [managing director] of a bank in one corner working on his private laptop."

Around 2010, the era of internet access via cybercafés in SSA countries came to an end due to mobile internet (see, e.g. Olofinlua, 2015). With telecom companies starting to offer mobile-browsing packages and increasing adoption of internet-enabled mobile phones, an alternative to the "long queues, overstuffed rooms, [and] lack of privacy" in cybercafés established (see, e.g., Quadri, 2023). According to a survey in several African countries, by 2011/12 mobile internet was the most commonly used form to access the internet (Stork et al., 2014). Still today, for many people in SSA data can be prohibitively expensive and cybercafés remain a

prominent way to access the internet for low-income families (Quadri, 2023).

A.1.3 Additional robustness analyses

NTL precision, blurring, and buffer. In our main specification, we consider a buffer of 2 kilometers around built-up areas due to blurring of the NTL data (cf. Figure 1.2). In column (4) of Table A.21, we remove the 2 kilometers buffer and estimate on the original *Africapolis* built-up areas. This implies we lose pixels at the town borders, typically with lower light intensities. As a result, our sample shrinks as some towns feature lit pixels only outside the built-up area but within the 2 kilometer buffer zone. This also leads to losing the country Angola.[2] An advantage of this approach is that blurring spilling over from nearby agglomerations is less prominent. Note that this is a marginal problem as we consider remote towns. The main effect, in comparison to the relevant baseline sample specification in column (2) of Table A.4 is robust with a slightly higher point estimate. With this robustness check, we show that our results do not depend on the adjustment of the built-up area. It also suggests that local light emissions originate predominantly at the town center rather than its outskirts.

We elicit economic growth of towns from changes in NTL emissions. In the main specification, we require stable NTL emission of towns over time and restrict our sample to towns with light emission in all years after 1994 (the earliest year in the sample). This ensures we capture meaningful changes in local light emissions. As this comes at the expense of sample size, we relax this restriction and conduct two types of robustness analyses. First, in Table A.11, we allow the sample to have missing light emission in up to three years at any point in time. In columns (1) through (4) there is no other restriction, while the specifications in columns (5) through (8) further require stable light emission in early years. Sample size and the number of countries increases when allowing for more missing NTL years. Results remain remarkably robust, yet some feature slightly smaller point estimates and are less precisely estimated. We therefore estimate alternative specifications with imputed values in Table A.12, which improves statistical power on the estimates compared to Table A.11. While these techniques allow to include more and even smaller towns, it comes at the expense of precision and pushes the NTL data to its limits. We therefore prefer our baseline model featuring a sample of towns with stable NTL emission over time.

Nodal cities. Generally, classifying agglomerations into subgroups is a debated topic and depends on many factors such as the country context and development (see, e.g., Frey and

[2] Estimating on the sample of the main specification without Angola is shown in column (2) of Table A.4. The sample shrinks, but the main effect estimate remains stable.

Zimmer, 2001). For our classification of nodal cities, we follow Dijkstra et al. (2020), who classify cities as agglomerations with more than 50,000 inhabitants. Note that we do not consider population density as a second criterion. Our sample of towns also coincides well with the definition of Dijkstra et al. (2020) (between 5,000 and 50,000 inhabitants). Still, the threshold for nodal cities is somewhat arbitrary. Therefore, we present robustness analyses in Table A.16 and Table A.17. In Table A.16, we vary the absolute cutoff value around our preferred definition and present alternatives ranging from 30,000 to 100,000 inhabitants. Results are very stable and tend to become slightly larger when more large towns are excluded, providing reassurance that we do not include unreasonably large towns. Yielding similarly robust results, Table A.17 presents specifications using percentile thresholds.

Internet access. Our interviews with experts at *Africa Bandwidth Maps* suggest an average distance of 10 kilometers to access points is an appropriate proxy for internet availability, given the transmission technology used predominantly at the time.[3] Consequently, in our main specification we define towns with an access point to the national backbone within 10 kilometers as within-reach, i.e., having access to internet infrastructure. Note that, in general, internet infrastructure availability is best interpreted as intention-to-treat effect. Some sources (e.g., Ngari and Petrack, 2019) suggest access points have a wider average range up to 50 kilometers, depending on geographical characteristics. In Table A.19, we estimate heterogeneous effects for towns within 10, 10-30, and 30-50 kilometer distance of an access point, respectively. Results show the effect is present for towns within 10 kilometers and decreases but remains statistically significant, though on a lower level, for towns within 10-30 kilometers. There is no measurable effect for towns within 30-50 kilometers.

In Table A.20, we re-estimate our baseline model using alternative distance thresholds of 5, 7.5, 12.5, and 15 kilometers. Note that the distance threshold affects the sample. Specifically, the control group shrinks when allowing for higher distances. For identification, it is important that the treatment group contains only towns with internet infrastructure access while the control group has no access. Too low distance thresholds potentially violate the first condition; too high distance thresholds might lead to wrong attribution of treatment status to suitable control towns. Results show a stable effect throughout all specifications. The slight reductions in point estimates and statistical power suggest our preferred specification is appropriate.

Clustering. A potential concern is that model errors are spatially correlated within regions.

[3] In their own analyses of population catchment areas from 2009 onward, *Africa Bandwidth Maps* use 10, 25, and 50 kilometer distances, respectively, for different scenarios. During the 2000s, the early years of national backbones in SSA, we opt for 10 kilometers.

Whenever more than one town is located within 10 kilometers to the access point, an access point serves more than one town. Therefore, we cluster at the access point level in our preferred specification. Yet, most treatment and control group towns do not share an access point and are also not located close to one another. Moreover, access points might generate spillover effects in surrounding areas. To take this into account, we apply a higher level of clustered standard errors in column (2) of Table A.14 using the administrative units of states (Admin-1). In addition, we re-estimate our baseline model with grid cell level clustering at one- (column (3)) and three-degree (column (4)) grid cells, a frequently applied alternative clustering method (see, e.g., Määttä et al., 2022; Hjort and Poulsen, 2019). Reassuringly, all specifications yield close to unchanged results with barely moving confidence intervals.

Fixed effects. In our baseline specification, we apply country-year fixed effects to account for country-specific growth paths. For robustness, we relax fixed effect granularity and re-estimate our preferred specification with the classical two-way fixed effects (TWFE) only: towns and calendar years. This specification is less demanding in the set of fixed effects. A potential concern with a TWFE specification might be that countries on a higher growth path might construct more access points faster. Therefore, this specification serves as a robustness check and not as the main specification provided in column (1) of Table A.10. At the same time, it significantly increases the sample. With TWFE, the estimate significantly increases. As we consider country-specific growth trends likely, we opt for the more conservative set of fixed effects in our main specification.

Control group. Our baseline specification relies on a fairly conservative design of control (and treatment) group focused on identification. As a result, a potential concern might be that this imposes unnecessarily strict restrictions on our sample. In columns (2) and (3) of Table A.10, we therefore extend our sample by easing some restraints. In column (2), we allow towns that did not receive an access point until the end of our data period in 2020 in the control group. This increases our sample significantly both in terms of towns and countries. Although we show that the type of towns we study incidentally get access due to their on-route location, one might have concerns with this specification regarding potential selection issues. Results corroborate the validity of our empirical design and external validity as the estimates remain unaffected while sample size increases. Nevertheless, for our baseline specification we stick with the more restricted sample for cleaner identification.

In column (3) of Table A.10, we extend the sample by adding towns to the control group that were connected during the five-year period after connection. In our main specification, these

towns are excluded as they neither belong clearly to the treatment nor the control group and would thus confound our analysis. However, given our finding that the effect of internet on growth materializes with a lag of two to three years, these towns are unlikely to exert a strong confounding effect on our results. At the same time, they significantly increase our sample size as well as the number of countries. With this specification, results remain robust and show a highly significant and only slightly smaller effect. As this could be due to some confounding, we stick with our baseline specification excluding towns receiving access in the post-period.

Although this reduces concerns regarding the suitability of our control group, a related concern might be that towns being connected through an access point which was constructed many years after the first internet connection are not comparable to the treated towns which were connected through an access point constructed before the first internet connection. We address this concern in Table A.21 by re-estimating our baseline specification restricting the control group to towns receiving an access point just after the five-year post period. We apply different levels of stringency to trade-off the resulting reduction in sample size and improved identification. Columns (1) through (4) use calendar year cutoffs while columns (5) through (7) apply cutoffs in years relative to countrywide connection. In line with the notion of incidentally connected on-route towns, we find no strong impact on our estimate.

Our identification builds on the notion that the plausibly exogenous timing of SMC arrivals affects countries in different stages of their national backbone expansion. This implies some countries receive international bandwidth and therefore internet connection with little rural internet infrastructure while in other countries national backbone expansion already progressed to more regions. This is shown by Figure A.1, which plots progress in national backbone expansion against connection year for SSA countries. Although not strong, which is expected given the unpredictability of SMC arrival, we observe a positive relation, i.e., countries connected later progressed further in the expansion of national backbones when provided with international bandwidth. This supports our empirical strategy as it exemplifies the variation in national backbone access around treatment date.

Countries. In our baseline specification, we rely on caparison within countries. Still, given the large variation in country sizes and country sample sizes, a potential concern might be to what extent our results are driven by selected countries. There is a considerable heterogeneity between landlocked and coastal countries (cf. Section 2.4.4). Therefore, in Table A.4 we re-run iterations of our baseline regression and exclude each country in our sample. Similarly, in

Table A.5 we re-estimate the effect for coastal countries. Results are remarkably robust across all specifications and remain statistically significant at the 1%-level. This is not only reassuring with respect to the presence of the effect in all countries contained in our sample, but also points to low effect heterogeneity across countries.

Employment. Our heterogeneity analysis with respect to regional employment shifts using *IPUMS International* survey data features the same geography times connection controls as our baseline specification to allow for changes in the importance of geography over time. However, given the time resolution of the survey data is much less granular than years this specification might be too demanding. Therefore, in Table A.18 we omit the geography controls and instead rely on region and country-year fixed effects. The results remain unchanged for all sectors in significance, although point estimates consistently show slightly larger effects as measured in levels. This generally suggests robust effects. If anything, we slightly underestimate the effect strength in our more demanding main specification.

Ethnic favoritism. A concern regarding our empirical model might be that certain ethnic groups were favored during rollout. Though the exogenous shock comes from countrywide connections and parallel trends in the event study do not underpin this concern, this would still be problematic if certain ethnic groups are also favored along other dimensions with the same timing, causing the observed growth differences over time. Using the map of ethnic boundaries by (Murdock, 1959) digitized by Weidmann et al. (2010), we extract the ethnic group majority in the area of each access point. Figure A.17 descriptively shows that many countries construct access points for more than one ethnic group before the treatment period. For the countries in our analysis, all countries except Angola provide at least two different ethnic groups with access points.[4] This already provides some indicates counter ethnic favoritism. Second, we construct country-ethnic group entities instead of countries. By re-estimating our baseline specification including town and country-ethnicity-year fixed effects, treatment and control group towns are compared only within a particular ethnic group. If ethnic favoritism drives our effects, the estimate in this specification is expected to vanish. The results are shown in column (5) of Table A.10. Naturally, sample size reduces in this more demanding specification. The result remains robust with a slightly lower point estimate, showing that even when comparing treatment and control group towns in areas with the same ethnic group majority, internet availability has a positive effect on local economic activity.

[4] Angola generally established few access points prior to connection.

A.2 Tables

Table A.1: Connection years

Country	year	connection	landing point	upgrade
Namibia	1999	Neighboring country		2012
Djibouti	1999	Sub-marine cable	Djibouti City	2009
Senegal	2000	Sub-marine cable	Dakar	2010
Angola	2001	Sub-marine cable	Sangano	2012
Benin	2001	Sub-marine cable	Cotonou	2012
Ghana	2001	Sub-marine cable	Accra	2010
Cameroon	2001	Sub-marine cable	Douala	2012
Gabon	2001	Sub-marine cable	Libreville	2012
Nigeria	2001	Sub-marine cable	Lagos	2010
Ivory Coast	2001	Sub-marine cable	Abidjan	2010
Sudan	2003	Sub-marine cable	Port Sudan	2010
Mali	2004	Neighboring country		2010
Botswana	2004	Neighboring country		2009
Zimbabwe	2004	Neighboring country		2011
Burkina Faso	2005	Neighboring country		2010
Togo	2005	Sub-marine cable	Lomé	2012
Gambia	2005	Sub-marine cable	Banjul	2012
Chad	2005	Neighboring country		2012
Central African Republic (CAR)	2005	Neighboring country		2012
Guinea-Bissau	2005	Sub-marine cable	Suro	2012
Mozambique	2006	Sub-marine cable	Maputo	2009
Lesotho	2006	Neighboring country		2010
Niger	2006	Neighboring country		2012
Malawi	2007	Neighboring country		2010
Ethiopia	2007	Neighboring country		2012
Zambia	2007	Neighboring country		2011
Swaziland	2008	Neighboring country		2009

Notes: Table reports the connection years of all SSA countries being connected before 2009.
Sources: Africa Bandwidth Maps, Submarine Cable Map.

Table A.2: National backbone expansions

Country	ISO	connection via	connection year	national backbone	notes
Angola	AGO	SAT-3	2001	concentrated on the big cities along the coast; some routes to larger cities within the country; landing point for submarine cable in capital city in north-west of country	after initial expansion prior to the arrival of the SAT-3 cable in 2001, network expansion in AGO was non-existent until the African Cup (football) in 2010
Benin	BEN	SAT-3	2001	network expansion mainly to larger cities and towards border connection points with neigboring countries; landing point for submarine cables in south	access point at the border with BFA were present since 2009, but the actual connection was established as late as 2017 due to conflicts about land titles in the border area
Botswana	BWA	ZAF	2004	network expansion mainly to larger cities and state capitals as well as border points; denser network in the east, where larger cities and the capital are located; connection via south-easter border with ZAF	
Burkina Faso	BFA	SEN-MLI	2005	network is expanded focused on routes necessary for international conection and border points to further neigboring countries	access via SEN and MLI instead of the geographically more convenient CIV or GHA; civil unrest in CIV at the time
Cameroon	CMR	SAT-3	2001	network present in largest cities; landing point in capital city	network extends along an oil pipeline between CMR and TCD, with a stop in CAF; this route encompasses most of the CMRs backbone and connects TCD and CAF

Table continues on the next page.

Country	ISO	via	year	national backbone	notes
Chad	TCD	CMR-CAF	2005	network limited to south-west, the location of the capital; border connection close to capital	
Côte d'Ivoire	CIV	SAT-3	2001	extensive network expansion in the south but limited in the north; overall expansion mainly to larger cities	civil war during the early 2000s hindered network expansion to the north and made international connection through CIV unfeasible
Djibouti	DJI	SEA-ME-WE-3	1999	network expansion to larger cities as well as the border with ETH	no connection of neighboring countries until 2007 despite early connection
Eritrea	ERI	EASSy	2009	network expansion to limited number of larger cities	connected only in 2009 via the EASSy cable, long after all neighbor countries established somewhat extensive networks; there were border conflicts with ETH
Ethiopia	ETH	SDN	2007	network centered around capital and limited in Eastern regions	
Gabon	GAB	SAT-3	2001	small network; landing point in capital located in north-west	
Gambia	GMB	SEN	2005	network expansion along river, where larger cities are located	
Ghana	GHA	SAT-3	2001	extensive network expansion in the south; connections at northern border points only very late; landing point in capital at southern coast	

Table continues on the next page.

Country	ISO	via	year	national backbone	notes
Guinea-Bissau	GNB	SEN	2005	no network expansion; connection from Senegal	
Kenya	KEN	TEAMS	2009	network expansion focussed on south, except for larger cities in the north; landing point in capital	initiated a bilateral cable project with the UAE; although plans started as early as 2003, cable established in 2009, few years before the major multinational cable projects; therefore a unusually large part of the network established prior to sub-marine cable connection
Lesotho	LSO	ZAF	2006	network covers largest cities	
Madagascar	MDG	LION	2009	network covers the larger cities at the coasts	
Malawi	MWI	ZAF-MOZ	2007	network focused on the south	
Mali	MLI	SEN	2004	extensive network expansion with focus on populated south; few connections to the north	important transit country as connections from SEN run through MLI to the countries that could not connect via CIV or GHA
Mozambique	MOZ	ZAF	2006	extensive network expansion all over the country, but less dense in south	network expansion between major cities in the south prior to international connection via ZAF was established; connections between capital and larger cities are made through domestic submarine cables
Namibia	NAM	ZAF	1999	extensive and early network expansion all over the country, with connections to all borders	extensive network expansion before the international connection was established

Table continues on the next page.

Country	ISO	via	year	national backbone	notes
Niger	NER	BEN	2006	small network focussed on south, the location of the capital	
Nigeria	NGA	SAT-3	2001	extensive network expansion all over the country with connections to all borders; especially dense in coastal areas and around capital; landing point in south close to largest city	connection to NER in the North-west constructed on usually direct, straight route, leaving out some bigger cities
Rwanda	RWA	KEN-UGA	2009	network expansion to all regions	
Senegal	SEN	Atlantis-2	2000	network expansion to largest cities; landing point in capital	network partially present prior to international connection
South Africa	ZAF	SAT-2	1993	very dense network all over the country; two landing points for submarine cables	
Sudan	SDN	SAS-2	2003	network expansoin to all regional capitals; more dense in the east and along the Nile river; landing point at largest port	
Swaziland	SWZ	ZAF	2008	network covers largest cities	
Tanzania	TZA	EASSy	2009	network expansion with focus on the coast, but covers all major cities and regional capitals; landing point in capital	network expansion mainly prior to international connection
Togo	TGO	SEN-MLI-BFA	2005	network expansion from inland border with BFA to capital city at the coast	obtained connection via BFA instead of an own landing point or via NGA or GHA

Table continues on the next page.

Country	ISO	via	year	national backbone	notes
Uganda	UGA	KEN	2009	network expansion centered around capital	network expansion mostly prior to international connection
Zambia	ZMB	EASSy	2007	extensive network expansion all over the country	state-owned electricity grid operator used pre-existing powerlines to establish an unusually dense network
Zimbabwe	ZWE	ZAF	2004	network expansion covers larger cities and connections to border points	

Sources: Table A.23, Africa Bandwidth Maps, own research.

Table A.3: Summary statistics

	(1) Mean	(2) SD	(3) Min	(4) P25	(5) P50	(6) P75	(7) Max
Population							
in 2000	15,956.67	13,154.77	0.00	5,398.00	12,772.00	24,239.00	49,217.00
in 2015	36,504.72	27,033.93	10,209.00	17,156.00	28,011.00	46,439.00	205,943.00
Density (2015)	4,860.99	4,118.79	710.00	2,639.00	3,982.00	6,029.00	38,637.00
Agglomeration							
Built-up area (2015)	10.99	12.29	0.35	4.40	7.40	13.47	122.21
Light intensity (in t-1)	505.41	601.41	12.00	174.00	308.00	585.00	4,842.00
Light intensity (in t-1, avg.)	7.05	5.99	0.29	3.10	4.82	8.91	32.72
Geography							
Altitude	874.76	719.08	0.02	60.20	1,016.20	1,372.18	2,816.32
Distance to							
Capital	2.52	2.48	0.02	0.75	1.75	3.59	12.54
Coastline	3.70	2.89	0.00	0.94	3.84	5.47	11.57
River	0.57	0.52	0.00	0.15	0.47	0.91	3.36
Landing point	5.64	3.85	0.01	1.74	6.02	9.02	14.510
Road	0.03	0.12	0.00	0.00	0.00	0.00	1.13
Rrailroad	0.58	0.93	0.00	0.00	0.07	0.85	4.40
Border	1.20	1.21	0.00	0.20	0.85	1.77	5.02
Port	4.02	2.90	0.00	1.35	4.29	5.87	11.96
Electricity grid	0.12	0.32	0.00	0.00	0.00	0.05	2.25
Terrain ruggedness	10.60	1.63	0.00	9.80	10.84	11.63	13.36
Market access	14804589.90	107781232.19	119.00	1,256.00	4,987.00	12,922.00	988349824.00
Connectivity							
Distance to access point (2020)	1.28	2.58	0.00	0.00	0.00	1.13	9.43
Mobile coverage (in t-1, GSM)	0.59	0.48	0.00	0.00	1.00	1.00	1.00

Notes: Table reports summary statistics for the estimation sample. *Sources:* Africa Bandwidth Maps, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.4: Robustness: country exclusion

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Excluded country:	AO	BJ	BW	ET	MW	MZ	SD	SN	ZM	ZW
Connection × access	0.0908***	0.104**	0.113***	0.163***	0.111***	0.111***	0.103***	0.103**	0.113***	0.0817**
	(0.0313)	(0.0441)	(0.0390)	(0.0473)	(0.0396)	(0.0395)	(0.0391)	(0.0424)	(0.0401)	(0.0399)
Town FE	×	×	×	×	×	×	×	×	×	×
Country × year FE	×	×	×	×	×	×	×	×	×	×
GSM coverage	×	×	×	×	×	×	×	×	×	×
Geography controls × connection	×	×	×	×	×	×	×	×	×	×
Observations	2,200	2,057	2,200	1,859	2,222	2,200	2,211	2,002	2,101	1,738
Countries	9	9	9	9	9	9	9	9	9	9
Towns	200	187	200	169	202	200	201	182	191	158
Share treated	.48	.46	.44	.467	.47	.455	.478	.407	.455	.513
Adjusted R^2	0.945	0.945	0.938	0.947	0.942	0.941	0.944	0.942	0.941	0.930

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Sources: Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.5: Robustness: coastal country exclusion

	(1)	(2)	(3)	(4)	(5)	(6)
Excluded country:	AO	BJ	MZ	SD	SN	TG
Connection × access	0.220***	0.278**	0.338***	0.284***	0.299***	0.287***
	(0.0506)	(0.112)	(0.0742)	(0.0734)	(0.0881)	(0.0724)
Town FE	×	×	×	×	×	×
Country × year FE	×	×	×	×	×	×
GSM coverage	×	×	×	×	×	×
Geography controls × connection	×	×	×	×	×	×
Observations	836	748	902	935	715	979
Countries	5	5	5	5	5	5
Towns	76	68	82	85	65	89
Share treated	.605	.5	.5	.541	.369	.494
Adjusted R^2	0.908	0.919	0.888	0.901	0.901	0.896

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Sources: Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.6: Heterogeneity: infrastructure distance

	(1)	(2)	(3)	(4)	(5)
Connection × access	0.115**	0.107**	0.115***	0.119***	0.0949**
	(0.0490)	(0.0440)	(0.0437)	(0.0451)	(0.0466)
Connection × access ×					
distance roads	0.0306				
	(0.120)				
distance railroads		-0.0224			
		(0.0302)			
distance electricity grid			0.0765		
			(0.0492)		
distance border				-0.0421	
				(0.0508)	
distance capital					-0.0246
					(0.0541)
Town FE	×	×	×	×	×
Country × year FE	×	×	×	×	×
GSM coverage	×	×	×	×	×
Geography controls × connection	×	×	×	×	×
Observations	2,310	2,310	2,310	2,310	2,310
Countries	10	10	10	10	10
Towns	210	210	210	210	210
Share treated	*.462*	*.462*	*.462*	*.462*	*.462*
Adjusted R^2	0.943	0.942	0.942	0.942	0.942

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.7: Measurement: intensive margin

		pixel intensity				
	(1) lit 1995	(2) top 10%	(3) top 20%	(4) top 30%	(5) top 40%	(6) top 50%
Connection × access	0.0821**	0.0533*	0.0600**	0.0674**	0.0693**	0.0705**
	(0.0316)	(0.0280)	(0.0297)	(0.0311)	(0.0329)	(0.0337)
Town FE	×	×	×	×	×	×
Country × year FE	×	×	×	×	×	×
GSM coverage	×	×	×	×	×	×
Geography controls × connection	×	×	×	×	×	×
Observations	2,310	2,310	2,310	2,310	2,310	2,310
Countries	10	10	10	10	10	10
Towns	210	210	210	210	210	210
Share treated	*.462*	*.462*	*.462*	*.462*	*.462*	*.462*
Adjusted R^2	0.923	0.963	0.959	0.955	0.951	0.949

Notes: Table reports variations of intensive NTL measures. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.8: Census years

Country	connection	2010s	2000s	1990s
Benin	2001	2013	2002	1992
Ethiopia	2007	n.a.	2007	1994
Malawi	2007	n.a.	2008	1998
Mozambique	2006	n.a.	2007	1997
Zambia	2007	2010	2000	1990

Notes: Table reports available survey waves by country used in our analysis as well as their year of connection via SMC or neighboring country. *Sources:* IPUMS International, Submarine Cable Map.

Table A.9: Heterogeneity: transport infrastructure

	(1)	(2)	(3)
Sample:	road access	railroad access	non-main
Connection × access	0.107**	0.155**	0.0843**
	(0.0438)	(0.0672)	(0.0332)
Town FE	×	×	×
Country × year FE	×	×	×
GSM coverage	×	×	×
Geography controls × connection	×	×	×
Observations	1,892	957	2,024
Countries	10	10	10
Towns	172	87	184
Share treated	.465	.529	.418
Adjusted R^2	0.941	0.963	0.920

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.10: Robustness: control group

Sample:	(1) relax FE	(2) untreated	(3) all late	(4) no buffer	(5) ethnic
Connection × access	0.227***	0.105***	0.0976***	0.0835**	0.0933**
	(0.0424)	(0.0368)	(0.0357)	(0.0373)	(0.0364)
Town FE	×	×	×	×	×
Year FE	×				
Country × year FE		×	×	×	
Country × ethnicity × year FE					×
GSM coverage	×	×	×	×	×
Geography controls × connection	×	×	×	×	×
Observations	3,883	4,345	3,707	2,178	1,793
Countries	20	13	13	11	10
Towns	353	395	337	198	163
Share treated	.309	.268	.315	.455	.454
Adjusted R^2	0.916	0.937	0.944	0.981	0.946

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.11: Measurement: missing NTL years

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Missing years allowed:	0	1	2	3	0	1	2	3
Connection × access	0.101**	0.109***	0.0924**	0.0708*	0.109***	0.0897**	0.0833*	0.0853**
	(0.0448)	(0.0394)	(0.0388)	(0.0395)	(0.0383)	(0.0399)	(0.0426)	(0.0431)
Town FE	×	×	×	×	×	×	×	×
Country × year FE	×	×	×	×	×	×	×	×
GSM coverage	×	×	×	×	×	×	×	×
Geography controls × connection	×	×	×	×	×	×	×	×
NTL in early years					×	×	×	×
Observations	1,958	2,295	2,421	2,664	2,310	2,657	2,771	2,843
Countries	10	10	11	12	10	12	12	12
Towns	178	209	220	241	210	240	248	254
Share treated	.478	.45	.445	.452	.462	.446	.44	.433
Adjusted R^2	0.946	0.942	0.941	0.937	0.942	0.936	0.933	0.930

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.12: Measurement: missing NTL year imputation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Imputed years:	0	1	2	3	4	0	1	2	3	4
Connection × access	0.101**	0.0971**	0.0940**	0.0940**	0.0917**	0.109***	0.0853**	0.0861*	0.0822*	0.0822*
	(0.0448)	(0.0449)	(0.0443)	(0.0443)	(0.0439)	(0.0383)	(0.0410)	(0.0440)	(0.0440)	(0.0440)
Town FE	×	×	×	×	×	×	×	×	×	×
Country × year FE	×	×	×	×	×	×	×	×	×	×
GSM coverage	×	×	×	×	×	×	×	×	×	×
Geography controls × connection	×	×	×	×	×	×	×	×	×	×
NTL in early years						×	×	×	×	×
Observations	1,958	2,013	2,024	2,024	2,035	2,310	2,640	2,706	2,717	2,717
Countries	10	10	10	10	10	10	12	12	12	12
Towns	178	183	184	184	185	210	240	246	247	247
Share treated	.478	.464	.462	.462	.459	.462	.45	.451	.449	.449
Adjusted R^2	0.946	0.947	0.947	0.947	0.947	0.942	0.937	0.935	0.935	0.935

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.13: Robustness: electricity

Sample:	extended		capital and landing		all nodal	
Dep. var.: electricity access	(1)	(2)	(3)	(4)	(5)	(6)
Connection × access	0.000387	-0.0359	0.0411	0.0579	-0.0731	-0.0914
	(0.103)	(0.0688)	(0.114)	(0.0766)	(0.211)	(0.173)
Town FE	×	×	×	×	×	×
Country × year FE	×	×	×	×	×	×
GSM coverage	×	×	×	×	×	×
Geography controls × connection	×	×	×	×	×	×
Weights		×		×		×
Observations	270	270	250	250	102	102
Countries	6	6	6	6	4	4
Towns	94	94	88	88	37	37
Share treated	.351	.351	.307	.307	.351	.351
Adjusted R^2	0.680	0.806	0.675	0.784	0.720	0.814

Notes: Access to the electricity grid is aggregated at the town level. Weighting by the number of households. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Afrobarometer (rounds 1-4), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.14: Robustness: alternative clustering

			grid cell	
	(1)	(2)	(3)	(4)
SE cluster:	AP	state	1°	3°
Connection × access	0.109***	0.109***	0.109***	0.109***
	(0.0383)	(0.0384)	(0.0376)	(0.0388)
Town FE	×	×	×	×
Country × year FE	×	×	×	×
GSM coverage	×	×	×	×
Geography controls × connection	×	×	×	×
Clusters	159	69	106	52
Observations	2,310	2,310	2,310	2,310
Countries	10	10	10	10
Towns	210	210	210	210
Share treated	.462	.462	.462	.462
Adjusted R^2	0.942	0.942	0.942	0.942

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.15: Population growth

Dep. var.: population	(1)	(2)	(3)	(4)	(5)
Time window:	baseline	2000 - (SMC + 3)	incl. 1995	excl. 1995	pre/post
Connection × access	0.0116	-0.00283	0.0218	0.0124	0.0102
	(0.0183)	(0.00805)	(0.0374)	(0.0277)	(0.0191)
Town FE	×	×	×	×	×
Country × year FE	×	×	×	×	×
GSM coverage	×	×	×	×	×
Geography controls × connection	×	×	×	×	×
Observations	2,310	1,765	830	610	440
Countries	10	10	10	10	10
Towns	210	210	210	210	210
Share treated	*.462*	*.462*	*.462*	*.462*	*.462*
Adjusted R^2	0.999	1.000	0.997	0.999	1.000

Notes: Population is measured as the logarithmic sum of pixel-level population counts. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Gridded Population of the World, Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.16: Robustness: absolute population thresholds

	(1)	(2)	(3)	(4)	(5)
Threshold:	30,000	40,000	50,000	75,000	100,000
Connection × access	0.129***	0.119***	0.109***	0.102***	0.0940***
	(0.0418)	(0.0391)	(0.0383)	(0.0346)	(0.0347)
Town FE	×	×	×	×	×
Country × year FE	×	×	×	×	×
GSM coverage	×	×	×	×	×
Geography controls × connection	×	×	×	×	×
Observations	1,903	2,167	2,310	2,453	2,486
Countries	10	10	10	10	10
Towns	173	197	210	223	226
Share treated	.462	.452	.462	.471	.478
Adjusted R^2	0.929	0.938	0.942	0.947	0.950

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.17: Robustness: percentile population thresholds

	(1)	(2)	(3)	(4)	(5)	(6)
Threshold:	100%	90%	80%	70%	60%	50%
Connection × access	0.0908***	0.115***	0.154***	0.168***	0.158***	0.136**
	(0.0335)	(0.0379)	(0.0480)	(0.0547)	(0.0577)	(0.0646)
Town FE	×	×	×	×	×	×
Country × year FE	×	×	×	×	×	×
GSM coverage	×	×	×	×	×	×
Geography controls × connection	×	×	×	×	×	×
Observations	2,640	2,145	1,659	1,298	1,074	854
Countries	10	10	9	9	9	9
Towns	240	195	151	118	98	77
Share treated	.5	.477	.49	.508	.531	.532
Adjusted R^2	0.963	0.948	0.943	0.939	0.939	0.940

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Sources: Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.18: Robustness: industry heterogeneity

	agriculture		manufacturing		services	
	(1)	(2)	(3)	(4)	(5)	(6)
Connection × access	-0.0188	-0.0194	0.0133*	0.0129*	0.00547	0.00642
	(0.0161)	(0.0163)	(0.00756)	(0.00739)	(0.0104)	(0.0107)
Town FE	×	×	×	×	×	×
Country × year FE	×	×	×	×	×	×
GSM coverage	×	×	×	×	×	×
Geography controls × connection		×		×		×
Observations	956,454	956,454	956,454	956,454	956,454	956,454
Countries	5	5	5	5	5	5
Regions	99	99	99	99	99	99
Share treated	.208	.208	.208	.208	.208	.208
Adjusted R^2	0.127	0.128	0.035	0.039	0.094	0.100

Notes: Employment shares are measured at the region level. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, IPUMS International, Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.19: Robustness: access point

	(1)	(2)
Connection × access point ∈ (0km, 10km]	0.147***	0.119***
	(0.0511)	(0.0385)
Connection × access point ∈ (10km, 30km]	0.0925	0.0863**
	(0.0606)	(0.0367)
Connection × access point ∈ (30km, 50km]	0.0489	0.0280
	(0.0545)	(0.0369)
Town FE	×	×
Country × year FE	×	×
GSM coverage	×	×
Geography controls × connection	×	×
Untreated controls		×
Observations	2,310	4,114
Countries	10	12
Towns	210	374
Share treated	.462	.27
Adjusted R^2	0.942	0.927

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.20: Robustness: distance threshold access points

	(1)	(2)	(3)	(4)	(5)
Threshold:	5km	7.5km	10km	12.5km	15km
Connection × access	0.0952**	0.107**	0.109***	0.0870**	0.0868**
	(0.0372)	(0.0426)	(0.0383)	(0.0410)	(0.0400)
Town FE	×	×	×	×	×
Country × year FE	×	×	×	×	×
GSM coverage	×	×	×	×	×
Geography controls × connection	×	×	×	×	×
Observations	1,936	2,156	2,310	2,387	2,398
Countries	9	10	10	10	10
Towns	176	196	210	217	218
Share treated	.415	.423	.462	.498	.518
Adjusted R^2	0.945	0.940	0.942	0.942	0.941

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Table A.21: Robustness: control group

	access point prior to				post-SMC years		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	2020	2018	2016	2014	20	14	8
Connection × access	0.109**	0.0879*	0.150**	0.146**	0.109**	0.0613	0.122**
	(0.0453)	(0.0503)	(0.0577)	(0.0647)	(0.0453)	(0.0487)	(0.0558)
Town FE	×	×	×	×	×	×	×
Country × year FE	×	×	×	×	×	×	×
GSM coverage	×	×	×	×	×	×	×
Geography controls × connection	×	×	×	×	×	×	×
Observations	2,310	2,101	1,496	1,177	2,310	2,079	1,320
Countries	10	9	8	6	10	10	8
Towns	210	191	136	107	210	189	120
Share treated	.459	.492	.522	.467	.459	.439	.592
Adjusted R^2	0.948	0.948	0.956	0.960	0.948	0.953	0.956

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

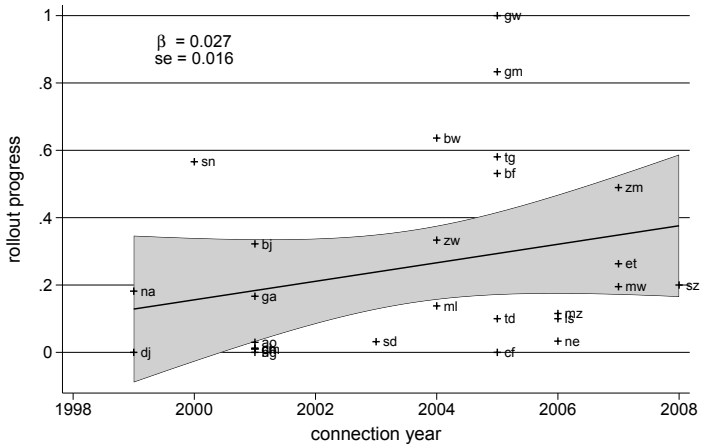
Table A.22: Robustness: lagged mobile coverage

	(1)	(2)	(3)	(4)	(5)	(6)
Connection × access	0.109***	0.110***	0.105***	0.106***	0.105***	0.102***
	(0.0383)	(0.0378)	(0.0384)	(0.0373)	(0.0373)	(0.0381)
GSM coverage	0.0539					
	(0.0380)					
GSM coverage (lag 1)		0.0758*				
		(0.0402)				
GSM coverage (lag 2)			-0.0161			
			(0.0399)			
GSM coverage (lag 3)				0.0510		
				(0.0327)		
GSM coverage (lag 4)					0.0518*	
					(0.0311)	
GSM coverage (lag 5)						0.0434
						(0.0335)
Town FE	×	×	×	×	×	×
Country × year FE	×	×	×	×	×	×
Geography controls × connection	×	×	×	×	×	×
Observations	2,310	2,310	2,310	2,310	2,310	2,310
Countries	10	10	10	10	10	10
Towns	210	210	210	210	210	210
Share treated	.462	.462	.462	.462	.462	.462
Adjusted R^2	0.942	0.942	0.942	0.942	0.942	0.942

Notes: NTL intensity is measured as the logarithmic sum of light intensities. Geography controls include indicators for local availability of and (logarithmic) distance to the capital, road, railroad, and port. Geography controls are constant over time and enter the model as interaction with the connection indicator. Robust standard errors clustered at the level of the closest access point are reported in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2020), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

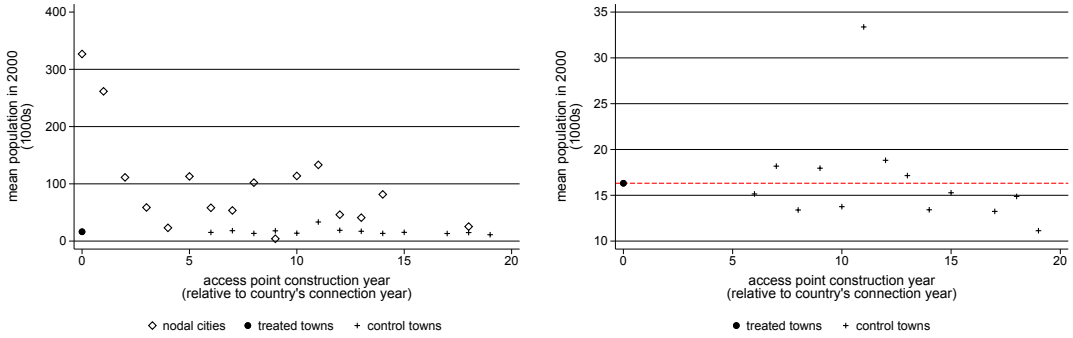
A.3 Figures

Figure A.1: SMC connection and national backbone rollout



Note: The figure plots rollout progress at the time of connection against connection year. Rollout progress is measured as share of access points in the connection year relative to the total number of access points in the most recent data year, 2020. Marker labels are ISO-2 country codes. Black line shows linear fit. The gray area represents 95% confidence intervals. β and 'se' refer to slope coefficient and standard error, respectively. *Sources:* Africa Bandwidth Maps, Submarine Cable Map.

Figure A.2: National backbone rollout



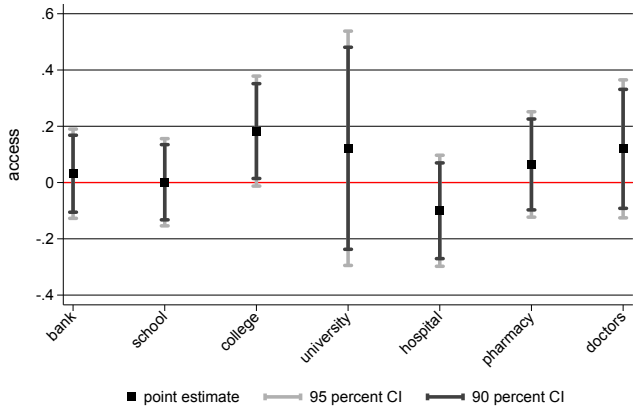
(a) with nodal cities **(b)** treatment and control towns

Note: The figure depicts the average population size of connected cities and towns by year relative to the connection year. On the left, the black dot in the lower left corner represents the treated towns, while the control towns are represented by the plus symbol and the nodal cities by a diamond. For treated towns and nodal cities that were connected in earlier years than the arrival of an SMC are shown in year zero as well for clarity. On the right, the treatment and control group are shown in more detail without nodal cities. *Sources:* Africa Bandwidth Maps, Africapolis, own calculations.

Figure A.3: Internet cafe in rural South Africa, 2009

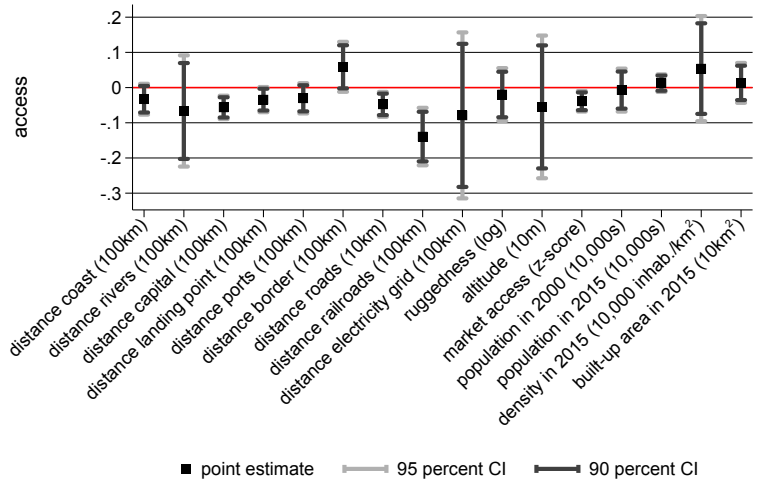
Source: Ossewa [CC BY-SA 4.0].

Figure A.4: Sample balance: POIs



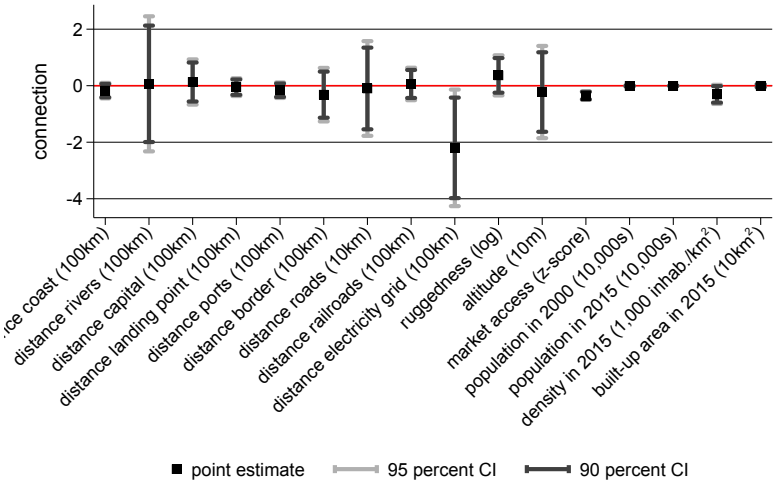
Note: Figure plots point estimates and confidence intervals for linear regressions of various points-of-interest on treatment group status. *Sources:* Africa Bandwidth Maps, Africapolis, Open Street Map, own calculations.

Figure A.5: Sample balance: national backbone rollout and geography



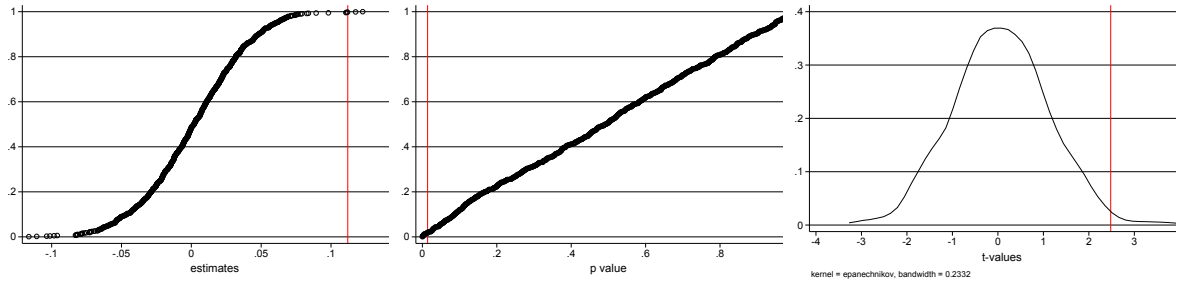
Note: Figure plots point estimates and confidence intervals for linear regressions of geodesic distance to various points-of-interest on treatment group status. *Sources:* Africa Bandwidth Maps, Africapolis, Open Street Map, own calculations.

Figure A.6: Sample balance: SMC connection and geography



Note: Figure plots point estimates and confidence intervals for linear regressions of geodesic distance to various points-of-interest on connection year, controlling for coastal country status. *Sources:* Africa Bandwidth Maps, Africapolis, Open Street Map, own calculations.

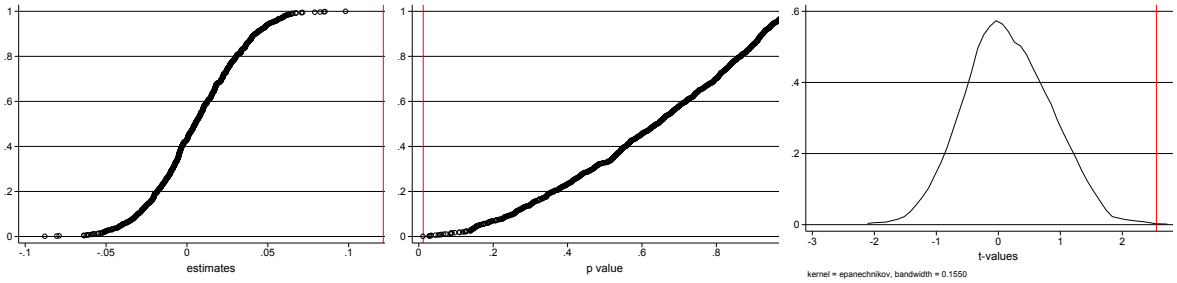
Figure A.7: Robustness: access placebo



(a) estimates **(b)** p-values **(c)** t-values

Note: Figure depicts different statistics of 1,000 permutations for our baseline estimation with randomly assigned treatment group status. Panel (a) plots coefficient estimates for our main effect and Panel (b) the respective p-values. Panel (c) depicts the kernel density estimate for the distribution of t-statistics. Values from the true regression are shown as vertical red lines. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2017), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

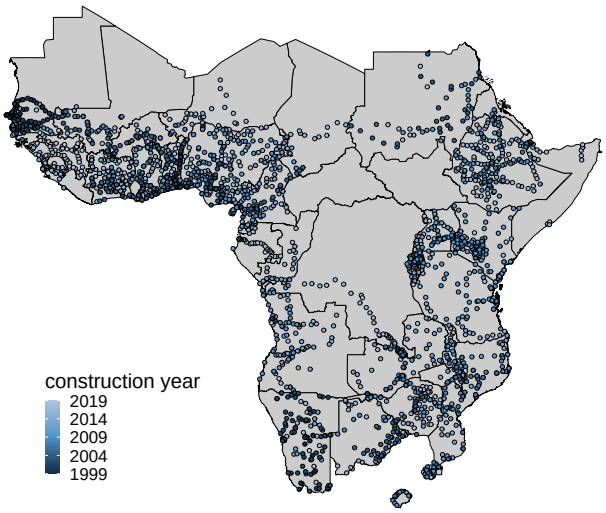
Figure A.8: Robustness: connection placebo



(a) estimates **(b)** p-values **(c)** t-values

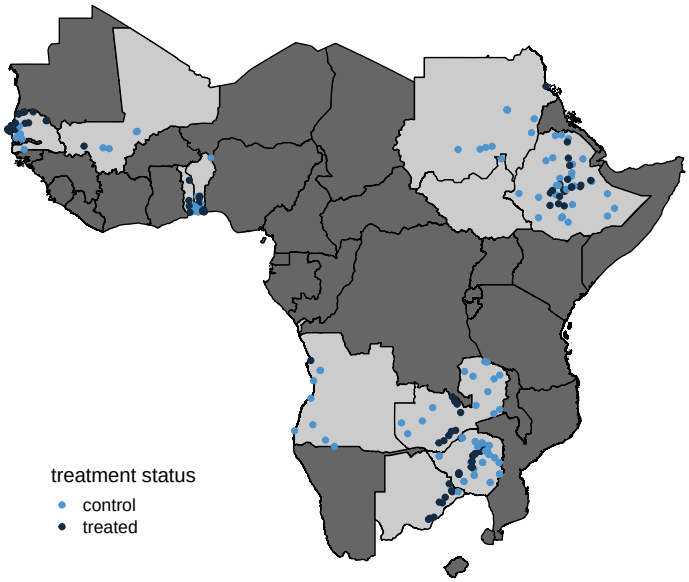
Note: Figure depicts different statistics of 1,000 permutations for our baseline estimation with randomly assigned treatment group status. Panel (a) plots coefficient estimates for our main effect and Panel (b) the respective p-values. Panel (c) depicts the kernel density estimate for the distribution of t-statistics. Values from the true regression are shown as vertical red lines. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2017), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Figure A.9: Access points



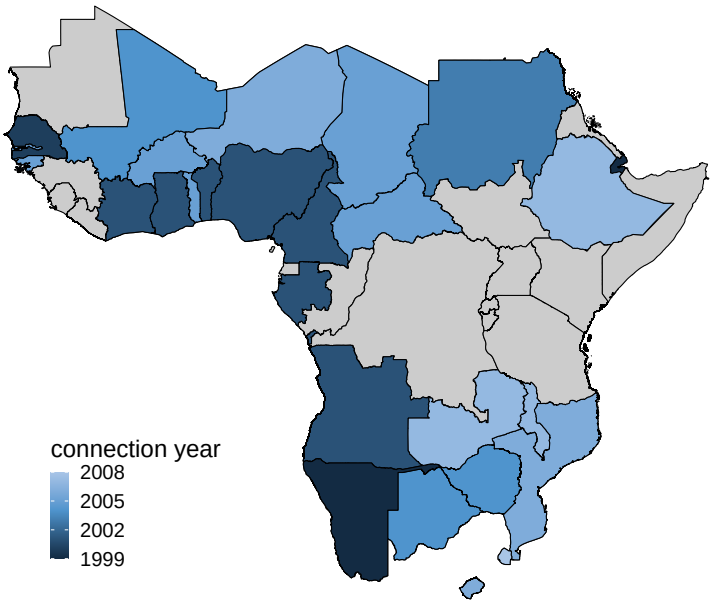
Note: Figure maps the location of all SSA access points. Blue coloring indicates contruction years with brighter blue corresponding to later years. *Sources:* Africa Bandwidth Maps, Table A.23.

Figure A.10: Sample: treatment and control towns



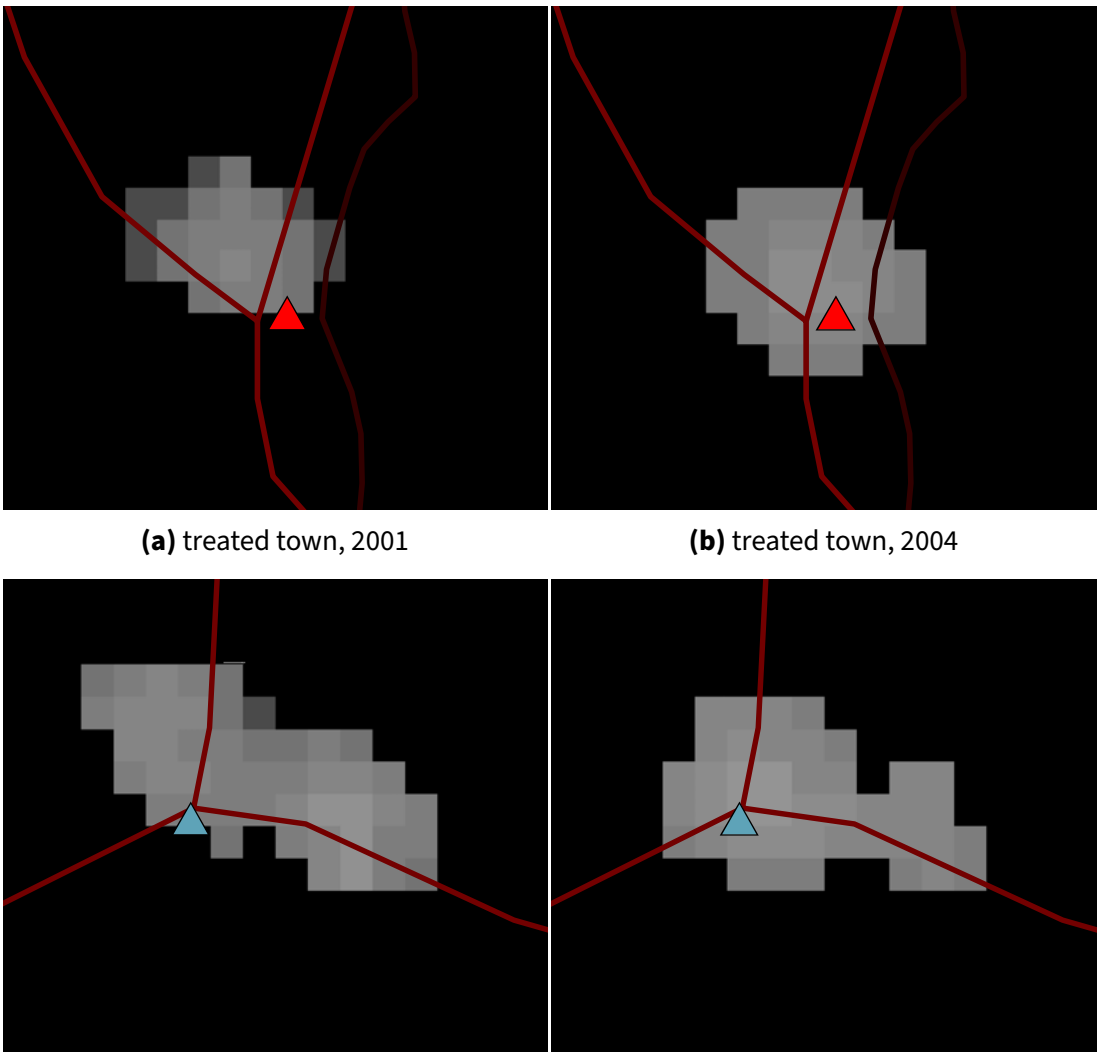
Note: Figure maps the countries in our main sample (brighter gray) and for each country the towns in the treatment and control group. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Africapolis, own calculations.

Figure A.11: SMC connection years



Note: Figure maps SSA countries and their country-wide connection years, with darker blues indicating earlier connection years. *Sources:* Submarine Cable Map.

Figure A.12: Data example treatment and control town, Benin

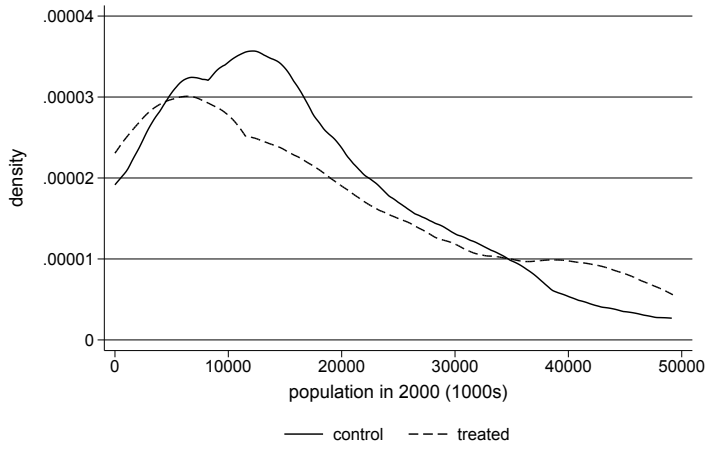


(a) treated town, 2001 **(b)** treated town, 2004

(c) control town, 2001 **(d)** control town, 2004

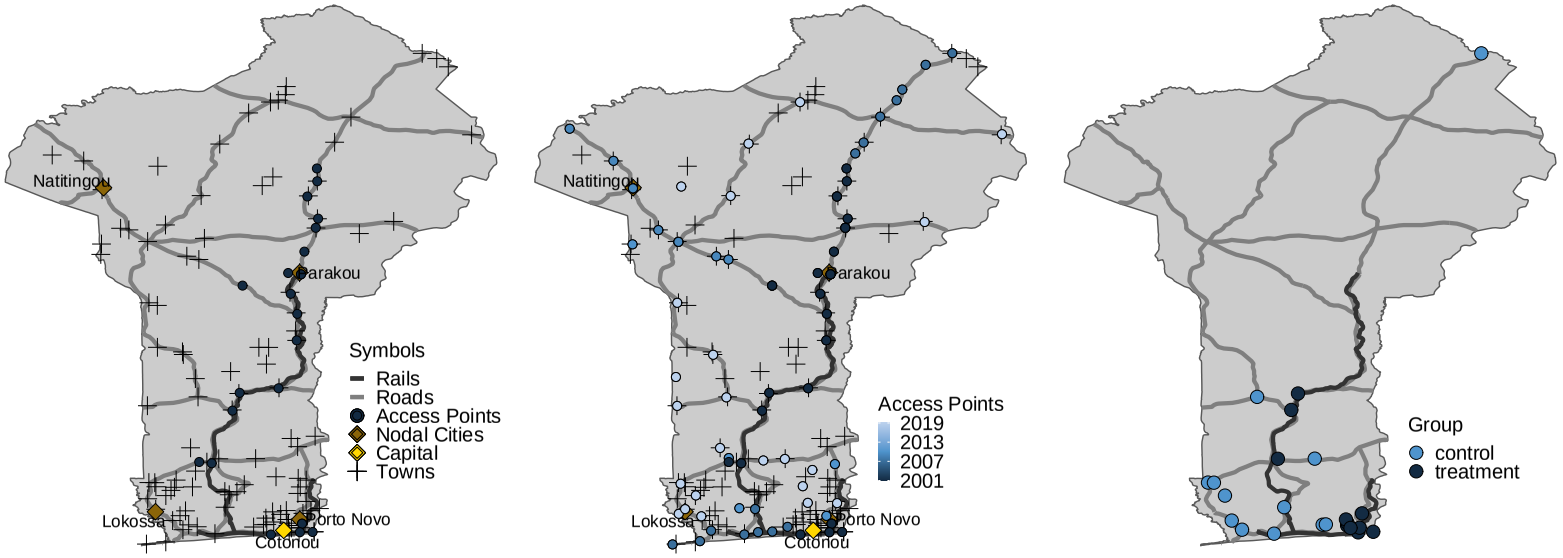
Note: The panels show a treatment and control group town from Benin, with gray NTLs pixels from 2001 and 2004. Access points are marked with a triangle (red if constructed until 2001 and blue if constructed afterward). The dark red line represents a major road connecting and the darker red line the railway. The black-to-white scale indicates light intensity, with brighter colors reflecting higher light intensities. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2017), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Figure A.13: Population distribution



Note: Figure plots kernel density estimates for the distribution of population size in 2000, separately for treated and control group towns. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Africapolis, own calculations.

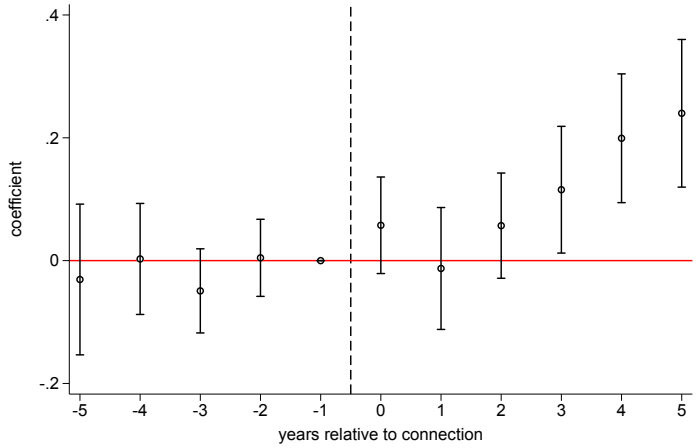
Figure A.14: Data example: national rollout in Benin



(a) Access points, 2001 **(b)** All access points **(c)** Treatment/control group

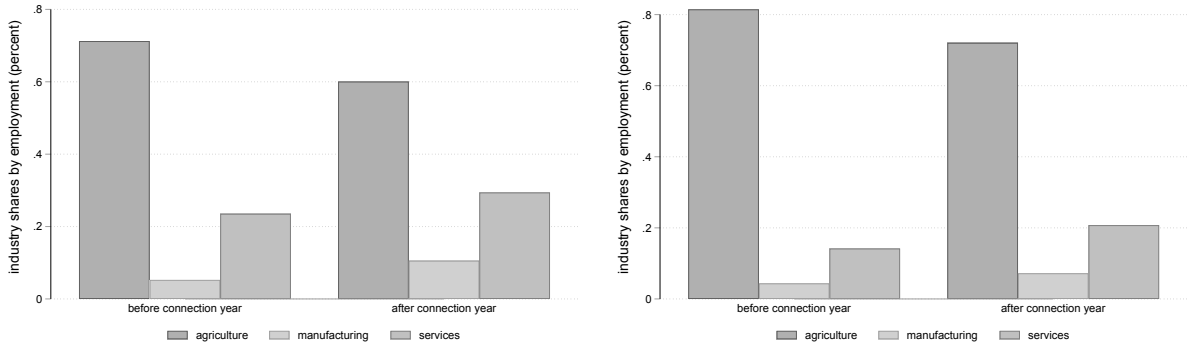
Note: Figure outlines the rollout of access points in Benin. Besides access points, the maps include the capital city, nodal cities, and all towns. Railroads and roads are included as well. In the left panel, the early rollout with access points being constructed until the arrival of the SMC in 2001 is shown. The middle panel depicts further access points and their respective construction years. The right panel shows the towns of your analysis divided into treatment and control group. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2017), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

Figure A.15: Event-study coefficients with 90%-level CIs



Note: The figure plots event study coefficients μ_{1j} based on Equation 1.2. The outcome is the logarithmic sum of light intensities. Bars represent 90% confidence intervals using robust standard errors clustered at the level of the closest access point. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, Li et al. (2017), Africapolis, Collins Bartholomew Mobile Coverage Maps, Open Street Map, own calculations.

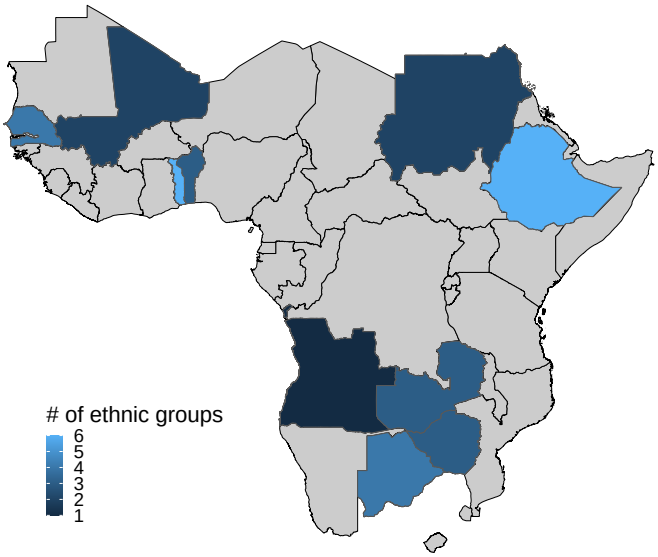
Figure A.16: Regional industry shares



(a) treatment regions **(b)** control regions

Note: Figure plots regional employment shares by industry for treated (Panel (a)) and control regions (Panel (b)), prior and after connection year. *Sources:* Africa Bandwidth Maps, Submarine Cable Map, IPUMS International, Africapolis, own calculations.

Figure A.17: Ethnic diversity



Note: Figure depicts the number of ethnic groups whose majority regions received at least one access point prior to the country-wide connection year. Brighter blues indicate a higher number of initially connected ethnic groups. *Sources:* Weidmann et al. (2010), Africa Bandwidth Maps, Submarine Cable Map, Africapolis, own calculations.

A.4 Early backbone deployment projects

Table A.23: Source register backbone deployment, pre-2009

Country	city/town	connection	URL source
Angola	Benguela	2009	https://www.linkedin.com/pulse/how-angola-got-its-first-workable-fiber-network-osvaldo-coelho
Angola	Cabinda	2009	https://www.linkedin.com/pulse/how-angola-got-its-first-workable-fiber-network-osvaldo-coelho
Angola	Dondo	2009	https://www.linkedin.com/pulse/how-angola-got-its-first-workable-fiber-network-osvaldo-coelho
Angola	N'dalatando	2009	https://www.linkedin.com/pulse/how-angola-got-its-first-workable-fiber-network-osvaldo-coelho
Angola	Sumbe	2009	https://www.linkedin.com/pulse/how-angola-got-its-first-workable-fiber-network-osvaldo-coelho
Angola	Chibia	2009	https://www.linkedin.com/pulse/how-angola-got-its-first-workable-fiber-network-osvaldo-coelho
Angola	Lubango	2009	https://www.linkedin.com/pulse/how-angola-got-its-first-workable-fiber-network-osvaldo-coelho
Angola	Luanda	2001	https://www.submarinenetworks.com/en/systems/euro-africa/sat-3
Angola	Malanje	2009	https://www.linkedin.com/pulse/how-angola-got-its-first-workable-fiber-network-osvaldo-coelho
Angola	Mocâmedes	2009	https://www.linkedin.com/pulse/how-angola-got-its-first-workable-fiber-network-osvaldo-coelho
Angola	Tômbua	2009	https://www.linkedin.com/pulse/how-angola-got-its-first-workable-fiber-network-osvaldo-coelho
Angola	N'zeto	2009	https://www.linkedin.com/pulse/how-angola-got-its-first-workable-fiber-network-osvaldo-coelho
Benin	Kandi	2007	http://www.infodev.org/infodev-files/resource/InfodevDocuments_421.pdf
Benin	Natitingou	2009	http://www.absucep.bj/fichiers/telechargeables/rapportFinal_SU_Volume1.pdf
Benin	Ouidah	2007	https://www.commsupdate.com/articles/2007/09/20/benin-and-togo-switch-on-sat-3-link/
Benin	Parakou	2001	https://researchictafrica.net/publications/Telecommunications_Sector_Performance_Reviews_2007/Benin%20Telecommunications%20Sector%20Performance%20Review%202007%20-%20English.pdf
Benin	Djougou	2009	http://www.absucep.bj/fichiers/telechargeables/rapportFinal_SU_Volume1.pdf
Benin	Cotonou	2001	https://www.submarinenetworks.com/en/systems/euro-africa/sat-3
Benin	Porto-Novo	2001	https://researchictafrica.net/publications/Telecommunications_Sector_Performance_Reviews_2007/Benin%20Telecommunications%20Sector%20Performance%20Review%202007%20-%20English.pdf
Benin	Abomey	2001	http://www.infodev.org/infodev-files/resource/InfodevDocuments_386.pdf
Botswana	Mahalapye	2004	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Palapye	2004	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Serowe	2005	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Nata	2008	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Ghanzi	2008	https://www.balancingact-africa.com/news/telecoms_en/4700/btc-launch-us323-million-trans-kalahari-fibre-project-in-botswana
Botswana	Mamuno	2008	https://www.balancingact-africa.com/news/telecoms_en/4700/btc-launch-us323-million-trans-kalahari-fibre-project-in-botswana
Botswana	Mochudi	2004	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Molepolole	2004	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Francistown	2004	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Maun	2008	https://www.balancingact-africa.com/news/telecoms_en/4700/btc-launch-us323-million-trans-kalahari-fibre-project-in-botswana
Botswana	Kasane	2008	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Ngoma	2008	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Gaborone	2004	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Lobatse	2004	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Kanye	2004	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Botswana	Jwaneng	2005	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
Burkina Faso	Banfora	2005	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf
Burkina Faso	Ouagadougou	2005	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf

Table continues on the next page.

Country	city/town	connection	URL source
Burkina Faso	Tenkodogo	2005	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf
Burkina Faso	Koupéla	2005	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf
Burkina Faso	Koudougou	2005	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf
Burkina Faso	Fada N'Gourma	2009	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf
Burkina Faso	Bobo Dioulasso	2005	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf
Burkina Faso	Orodara	2005	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf
Burkina Faso	Zorgho	2005	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf
Cameroon	Meiganga	2005	http://blog.gelgabon.net/2010/01/cameroun-fibre-optique-fibre-de_23.html
Cameroon	Bafia	2009	http://blog.gelgabon.net/2010/01/cameroun-fibre-optique-fibre-de_23.html
Cameroon	Yaounde	2005	https://www.researchictafrica.net/countries/cameroon/Sector_Strategy_for_Telecommunications_and_ICT_2005-2015.pdf
Cameroon	Mbalmayo	2009	http://blog.gelgabon.net/2010/01/cameroun-fibre-optique-fibre-de_23.html
Cameroon	Bélabo	2005	https://www.researchictafrica.net/countries/cameroon/Sector_Strategy_for_Telecommunications_and_ICT_2005-2015.pdf
Cameroon	Edéa	2005	https://www.researchictafrica.net/countries/cameroon/Sector_Strategy_for_Telecommunications_and_ICT_2005-2015.pdf
Cameroon	Douala	2001	https://www.submarinenetworks.com/en/systems/euro-africa/sat-3
Cameroon	Bamenda	2009	http://blog.gelgabon.net/2010/01/cameroun-fibre-optique-fibre-de_23.html
Cameroon	Kribi	2005	https://www.researchictafrica.net/countries/cameroon/Sector_Strategy_for_Telecommunications_and_ICT_2005-2015.pdf
Cameroon	Limbe	2009	http://blog.gelgabon.net/2010/01/cameroun-fibre-optique-fibre-de_23.html
Cameroon	Bafang	2009	http://blog.gelgabon.net/2010/01/cameroun-fibre-optique-fibre-de_23.html
Cameroon	Bafoussam	2009	http://blog.gelgabon.net/2010/01/cameroun-fibre-optique-fibre-de_23.html
Chad	Doba	2005	https://www.researchictafrica.net/countries/cameroon/Sector_Strategy_for_Telecommunications_and_ICT_2005-2015.pdf
Chad	Ndjamena	2009	http://blog.gelgabon.net/2010/01/cameroun-fibre-optique-fibre-de_23.html
Côte d'Ivoire	San-Pedro	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27295
Côte d'Ivoire	Sassandra	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27296
Côte d'Ivoire	Soubré	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27295
Côte d'Ivoire	Aboisso	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27295
Côte d'Ivoire	Gagnoa	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27295
Côte d'Ivoire	Divo	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27295
Côte d'Ivoire	Toumodi	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27296
Côte d'Ivoire	Yamoussoukro	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27295
Côte d'Ivoire	Dimbokro	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27295
Côte d'Ivoire	Abidjan	2001	https://www.submarinenetworks.com/en/systems/euro-africa/sat-3
Côte d'Ivoire	Man	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27295
Côte d'Ivoire	Guiglo	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27295
Côte d'Ivoire	Daloa	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27295
Côte d'Ivoire	Bouaflé	2005	https://idl-bnc-idrc.dspacedirect.org/handle/10625/27295
Côte d'Ivoire	Ferkessédougou	2008	http://malijet.com/a_la_une_du_mali/7721-mali-c_te_d_ivoire_interconnexion_de_la_fibre_optique.html
Côte d'Ivoire	Bouaké	2008	http://malijet.com/a_la_une_du_mali/7721-mali-c_te_d_ivoire_interconnexion_de_la_fibre_optique.html
Djibouti	Ali Sabieh	2007	https://www.submarinenetworks.com/en/systems/eurasia-terrestrial/renovation-of-the-djibouti-ethiopia-digital-corridor
Djibouti	Galafi	2007	https://www.submarinenetworks.com/en/systems/eurasia-terrestrial/renovation-of-the-djibouti-ethiopia-digital-corridor
Djibouti	Djibouti	1999	https://web.archive.org/web/20081222095315/http://www.heise.de/tp/r4/artikel/5/5245/1.html
Eritrea	Mendefera	2009	https://en.wikipedia.org/wiki/EASSy

Table continues on the next page.

Country	city/town	connection	URL source
Eritrea	Asmara	2009	https://en.wikipedia.org/wiki/EASSy
Eritrea	Massawa	2009	https://en.wikipedia.org/wiki/EASSy
Ethiopia	Addis Ababa	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Debre Birhan	2007	https://www.flickr.com/photos/ssong/7013508301/
Ethiopia	Debre Markos	2007	https://www.flickr.com/photos/ssong/7013508301/
Ethiopia	Dese	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Bahir Dar	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Gondar	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Asosa	2009	https://www.zte.com.cn/global/about/magazine/zte-technologies/2009/6/en_414/172517.html
Ethiopia	Dire Dawa	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Harar	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Asela	2009	https://www.zte.com.cn/global/about/magazine/zte-technologies/2009/6/en_414/172517.html
Ethiopia	Nazret	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Debre Zeyit	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Nekemte	2009	https://www.zte.com.cn/global/about/magazine/zte-technologies/2009/6/en_414/172517.html
Ethiopia	Gore	2009	https://www.zte.com.cn/global/about/magazine/zte-technologies/2009/6/en_414/172517.html
Ethiopia	Jima	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Shashemene	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Hagere Hiywet	2009	https://www.zte.com.cn/global/about/magazine/zte-technologies/2009/6/en_414/172517.html
Ethiopia	Gimbi	2009	https://www.zte.com.cn/global/about/magazine/zte-technologies/2009/6/en_414/172517.html
Ethiopia	Arba Minch	2009	https://www.zte.com.cn/global/about/magazine/zte-technologies/2009/6/en_414/172517.html
Ethiopia	Hosaina	2009	https://www.zte.com.cn/global/about/magazine/zte-technologies/2009/6/en_414/172517.html
Ethiopia	Awasa	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Sodo	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Jijiga	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Ethiopia	Aksum	2009	https://www.zte.com.cn/global/about/magazine/zte-technologies/2009/6/en_414/172517.html
Ethiopia	Adigrat	2009	https://www.zte.com.cn/global/about/magazine/zte-technologies/2009/6/en_414/172517.html
Ethiopia	Mekele	2007	https://www.lightwaveonline.com/network-design/article/16663413/zte-to-build-national-network-in-ethiopia
Gabon	Libreville	2001	https://www.submarinenetworks.com/en/systems/euro-africa/sat-3
Gambia	Banjul	2005	https://www.siemens.be/cmc/newsletters/index.aspx?id=13-574-16687
Gambia	Brikama	2005	https://www.siemens.be/cmc/newsletters/index.aspx?id=13-574-16687
Gambia	Basse Santa Su	2005	https://www.siemens.be/cmc/newsletters/index.aspx?id=13-574-16687
Gambia	Bansang	2005	https://www.siemens.be/cmc/newsletters/index.aspx?id=13-574-16688
Gambia	Georgetown	2005	https://www.siemens.be/cmc/newsletters/index.aspx?id=13-574-16689
Ghana	Kumasi	2004	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.195.150&rep=rep1&type=pdf
Ghana	Obuasi	2004	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.195.150&rep=rep1&type=pdf
Ghana	Sunyani	2007	https://wikileaks.org/plusd/cables/07ACCRA2162_a.html
Ghana	Cape Coast	2004	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.195.150&rep=rep1&type=pdf
Ghana	Winneba	2004	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.195.150&rep=rep1&type=pdf
Ghana	Koforidua	2004	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.195.150&rep=rep1&type=pdf
Ghana	Nkawkaw	2004	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.195.150&rep=rep1&type=pdf
Ghana	Accra	2001	https://www.submarinenetworks.com/en/systems/euro-africa/sat-3

Table continues on the next page.

Country	city/town	connection	URL source
Ghana	Tema	2004	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.195.150&rep=rep1&type=pdf
Ghana	Tamale	2007	https://wikileaks.org/plusd/cables/07ACCRA2162_a.html
Ghana	Ho	2008	https://www.moc.gov.gh/eastern-corridor-fiber-optic-backbone
Ghana	Sekondi	2004	http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.195.150&rep=rep1&type=pdf
Guinea-Bissau	Bissau	2005	https://www.siemens.be/cmc/newsletters/index.aspx?id=13-574-16687
Kenya	Bungoma	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Embu	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Garissa	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Kakamega	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Thika	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Kisumu	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Mwingi	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Nanyuki	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Machakos	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Meru	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Mombasa	2009	https://phys.org/news/2009-06-kenya-undersea-broadband-fibre-optic.html
Kenya	Nairobi	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Naivasha	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Nakuru	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Nyeri	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Voi	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Kenya	Eldoret	2009	https://www.nation.co.ke/kenya/business/teams-begins-laying-fibre-optic-cables-588868
Lesotho	Teyateyaneng	2009	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Lesotho.pdf
Lesotho	Butha-Buthe	2009	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Lesotho.pdf
Lesotho	Hlotse	2009	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Lesotho.pdf
Lesotho	Mafetang	2009	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Lesotho.pdf
Lesotho	Maseru	2006	https://researchictafrica.net/wp/wp-content/uploads/2018/01/2017_The-State-of-ICT-in-Lesotho_RIA_LCA.pdf
Lesotho	Mohales Hoek	2009	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Lesotho.pdf
Lesotho	Mokhotlong	2009	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Lesotho.pdf
Lesotho	Moyeni	2009	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Lesotho.pdf
Madagascar	Antananarivo	2009	https://www.lightwaveonline.com/network-design/article/16667160/orange-inaugurates-lion-submarine-cable-in-reunion
Madagascar	Toamasina	2009	https://www.lightwaveonline.com/network-design/article/16667160/orange-inaugurates-lion-submarine-cable-in-reunion
Madagascar	Mahajanga	2009	https://www.lightwaveonline.com/network-design/article/16667160/orange-inaugurates-lion-submarine-cable-in-reunion
Madagascar	Marovoay	2009	https://www.lightwaveonline.com/network-design/article/16667160/orange-inaugurates-lion-submarine-cable-in-reunion
Madagascar	Fianarantsoa	2009	https://www.lightwaveonline.com/network-design/article/16667160/orange-inaugurates-lion-submarine-cable-in-reunion
Madagascar	Ihosy	2009	https://www.lightwaveonline.com/network-design/article/16667160/orange-inaugurates-lion-submarine-cable-in-reunion
Madagascar	Antsirabe	2009	https://www.lightwaveonline.com/network-design/article/16667160/orange-inaugurates-lion-submarine-cable-in-reunion
Malawi	Lilongwe	2007	https://www.commsupdate.com/articles/2007/06/27/electric-board-begins-installing-fibre/
Malawi	Blantyre	2007	https://www.commsupdate.com/articles/2007/06/27/electric-board-begins-installing-fibre/
Malawi	Mwanza	2007	https://www.commsupdate.com/articles/2007/07/16/mtl-connects-network-to-mozambique/
Mali	Bamako	2004	https://journals.openedition.org/cea/944#ftn5

Table continues on the next page.

Country	city/town	connection	URL source
Mali	Bafoulabé	2004	https://journals.openedition.org/cea/944#ftn5
Mali	Kayes	2004	https://journals.openedition.org/cea/944#ftn5
Mali	Kita	2004	https://journals.openedition.org/cea/944#ftn5
Mali	Yélimané	2007	https://www.amrtp.ml/pdf/rapport_act/Rapport_2007.pdf
Mali	Kati	2004	https://journals.openedition.org/cea/944#ftn5
Mali	Koulikoro	2009	https://www.flickr.com/photos/ssong/6092447867/
Mali	Mopti	2009	https://www.afribone.com/?Inauguration-de-la-fibre-optique
Mali	Ségou	2009	https://www.flickr.com/photos/ssong/6092447867/
Mali	Bougouni	2005	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf
Mali	Koutiala	2005	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf
Mali	Sikasso	2005	https://www.itu.int/en/ITU-D/LDCs/Documents/2017/Country%20Profiles/Country%20Profile_Burkina%20Faso.pdf
Mozambique	Pemba	2008	https://macauhub.com.mo/2009/05/07/7018/
Mozambique	Xai-Xai	2006	https://www.fomsn.com/networks/fiber/fiber-optic-network-links-mozambique-and-south-africa/
Mozambique	Inhambane	2006	https://www.fomsn.com/networks/fiber/fiber-optic-network-links-mozambique-and-south-africa/
Mozambique	Maxixe	2009	https://farm7.static.flickr.com/6150/6035058808_7dc34bcf27_b.jpg
Mozambique	Vilanculos	2006	https://www.fomsn.com/networks/fiber/fiber-optic-network-links-mozambique-and-south-africa/
Mozambique	Chimoio	2008	https://macauhub.com.mo/2009/05/07/7018/
Mozambique	Manica	2009	https://farm7.static.flickr.com/6150/6035058808_7dc34bcf27_b.jpg
Mozambique	Maputo	2006	https://www.fomsn.com/networks/fiber/fiber-optic-network-links-mozambique-and-south-africa/
Mozambique	Nampula	2007	https://macauhub.com.mo/2009/05/07/7018/
Mozambique	Nacala	2009	https://farm7.static.flickr.com/6150/6035058808_7dc34bcf27_b.jpg
Mozambique	Lichinga	2008	https://macauhub.com.mo/2009/05/07/7018/
Mozambique	Cuamba	2007	https://macauhub.com.mo/2009/05/07/7018/
Mozambique	Beira	2006	https://www.fomsn.com/networks/fiber/fiber-optic-network-links-mozambique-and-south-africa/
Mozambique	Dondo	2007	https://www.commsupdate.com/articles/2007/07/09/tdm-lights-latest-link/
Mozambique	Tete	2008	https://macauhub.com.mo/2009/05/07/7018/
Mozambique	Nicuadala	2007	https://www.commsupdate.com/articles/2007/07/09/tdm-lights-latest-link/
Mozambique	Quelimane	2007	https://www.commsupdate.com/articles/2007/07/09/tdm-lights-latest-link/
Namibia	Karibib	2002	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Omaruru	2002	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Swakopmund	2002	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Walvis Bay	2002	http://home.intekom.com/intekom/clients/t/telecom_namibia/technology.stm
Namibia	Maltahöhe	2002	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Mariental	1999	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Rehoboth	1999	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Bethanie	2002	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Karasburg	2002	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Keetmanshoop	1999	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Lüderitz	2002	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Oranjemund	1999	https://www.oodaloop.com/documents/Legacy/CIA/factbook/geos/wa.html
Namibia	Rundu	2002	https://www.namibweb.com/namtel.htm
Namibia	Windhoek	1999	https://www.namibweb.com/namtel.htm
Namibia	Opuwo	2002	http://home.intekom.com/intekom/clients/t/telecom_namibia/technology.stm

Table continues on the next page.

Country	city/town	connection	URL source
Namibia	Oshikango	2002	http://home.intekom.com/intekom/clients/t/telecom_namibia/technology.stm
Namibia	Gobabis	2002	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Grootfontein	2002	https://www.namibweb.com/namtel.htm
Namibia	Otjiwarongo	2002	https://epublications.uef.fi/pub/URN_NBN_fi_joy-20090045/URN_NBN_fi_joy-20090045.pdf
Namibia	Katima Mulilo	2002	https://www.namibweb.com/namtel.htm
Niger	Dosso	2007	http://www.infodev.org/infodev-files/resource/InfodevDocuments_421.pdf
Niger	Gaya	2007	http://www.infodev.org/infodev-files/resource/InfodevDocuments_421.pdf
Niger	Niamey	2006	https://www.commsupdate.com/articles/2006/11/23/sonitel-fibre-optic-network-inaugurated/
Nigeria	Aba	2003	https://at.linkedin.com/in/josefweingand
Nigeria	Umuahia	2003	https://at.linkedin.com/in/josefweingand
Nigeria	Mubi	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Numan	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Yola	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Uyo	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Awka	2003	https://at.linkedin.com/in/josefweingand
Nigeria	Onitsha	2003	https://at.linkedin.com/in/josefweingand
Nigeria	Azare	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Bauchi	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Makurdi	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/
Nigeria	Oturkpo	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/
Nigeria	Bama	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Maiduguri	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Calabar	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Sapele	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/
Nigeria	Warri	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/
Nigeria	Benin City	2003	https://at.linkedin.com/in/josefweingand
Nigeria	Ado Ekiti	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Enugu	2003	https://at.linkedin.com/in/josefweingand
Nigeria	Abuja	2003	https://at.linkedin.com/in/josefweingand
Nigeria	Gombe	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Owerri	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/
Nigeria	Dutse	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Kaduna	2007	https://www.commsupdate.com/articles/2007/11/23/globacom-commissions-nationwide-fibre-optic-programme/
Nigeria	Zaria	2007	https://www.commsupdate.com/articles/2007/11/23/globacom-commissions-nationwide-fibre-optic-programme/
Nigeria	Kano	2007	https://www.commsupdate.com/articles/2007/11/23/globacom-commissions-nationwide-fibre-optic-programme/
Nigeria	Funtua	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Katsina	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Birnin Kebbi	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Lokoja	2003	https://at.linkedin.com/in/josefweingand
Nigeria	Ilorin	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/
Nigeria	Lagos	2001	https://www.submarinenetworks.com/en/systems/euro-africa/sat-3
Nigeria	Keffi	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/
Nigeria	Lafia	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/

Table continues on the next page.

Country	city/town	connection	URL source
Nigeria	Bida	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/
Nigeria	Minna	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/
Nigeria	Kontagora	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Abeokuta	2007	https://www.commsupdate.com/articles/2007/11/23/globacom-commissions-nationwide-fibre-optic-programme/
Nigeria	Ijebu Ode	2003	https://at.linkedin.com/in/josefweingand
Nigeria	Akure	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Ondo	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Owo	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Ife	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Oshogbo	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Ibadan	2003	https://at.linkedin.com/in/josefweingand
Nigeria	Ogbomosho	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/
Nigeria	Oyo	2008	https://www.commsupdate.com/articles/2008/07/16/globacom-in-ongoing-rollout/
Nigeria	Jos	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Port Harcourt	2003	https://at.linkedin.com/in/josefweingand
Nigeria	Sokoto	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Damaturu	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Potiskum	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Nigeria	Gusau	2009	http://documents.worldbank.org/curated/en/684121468010226781/pdf/536430PUB0Broa101Official0Use0Only1.pdf
Rwanda	Kibungo	2009	https://www.commsupdate.com/articles/2008/07/22/nationwide-backbone-to-be-completed-by-november-2009/
Rwanda	Kigali	2009	https://www.commsupdate.com/articles/2008/07/22/nationwide-backbone-to-be-completed-by-november-2009/
Rwanda	Byumba	2009	https://www.commsupdate.com/articles/2008/07/22/nationwide-backbone-to-be-completed-by-november-2009/
Rwanda	Butare	2009	https://www.commsupdate.com/articles/2008/07/22/nationwide-backbone-to-be-completed-by-november-2009/
Rwanda	Gitarama	2009	https://www.commsupdate.com/articles/2008/07/22/nationwide-backbone-to-be-completed-by-november-2009/
Rwanda	Nyanza	2009	https://www.commsupdate.com/articles/2008/07/22/nationwide-backbone-to-be-completed-by-november-2009/
Senegal	Dakar	2000	https://web.archive.org/web/20110927174252/http://www.convergedigest.com/Daily/v7/v7n092.htm
Senegal	Diourbel	2000	https://www.itu.int/osg/spu/wtpf/wtpf98/cases/Senegal/senegf1.pdf
Senegal	Fatick	2001	http://www.polis.sciencespobordeaux.fr/resultats/documents/externes/Guignard_DEA.pdf
Senegal	Kaolack	2001	http://www.polis.sciencespobordeaux.fr/resultats/documents/externes/Guignard_DEA.pdf
Senegal	Kolda	2001	http://www.polis.sciencespobordeaux.fr/resultats/documents/externes/Guignard_DEA.pdf
Senegal	Louga	2000	https://www.itu.int/osg/spu/wtpf/wtpf98/cases/Senegal/senegf1.pdf
Senegal	Dial	2000	https://www.itu.int/osg/spu/wtpf/wtpf98/cases/Senegal/senegf1.pdf
Senegal	Matam	2000	https://www.itu.int/osg/spu/wtpf/wtpf98/cases/Senegal/senegf1.pdf
Senegal	Saint-Louis	2000	https://www.itu.int/osg/spu/wtpf/wtpf98/cases/Senegal/senegf1.pdf
Senegal	Kidira	2000	https://www.itu.int/osg/spu/wtpf/wtpf98/cases/Senegal/senegf1.pdf
Senegal	Tambacounda	2001	http://www.polis.sciencespobordeaux.fr/resultats/documents/externes/Guignard_DEA.pdf
Senegal	Thiès	2000	https://www.itu.int/osg/spu/wtpf/wtpf98/cases/Senegal/senegf1.pdf
Senegal	Ziguinchor	2001	http://www.polis.sciencespobordeaux.fr/resultats/documents/externes/Guignard_DEA.pdf
South Africa	Bhisho	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	East London	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Graaff Reinet	2009	insufficient sources
South Africa	Grahamstown	1999	https://www.telkom.co.za/history/TelkomHistory/index.html

Table continues on the next page.

Country	city/town	connection	URL source
South Africa	Port Alfred	2009	insufficient sources
South Africa	Cradock	2009	insufficient sources
South Africa	Middelburg	2009	insufficient sources
South Africa	Queenstown	2009	insufficient sources
South Africa	Aliwal North	2009	insufficient sources
South Africa	Port Elizabeth	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Uitenhage	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Umtata	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Port St. Johns	2009	insufficient sources
South Africa	Johannesburg	1995	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Pretoria	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Springs	2009	insufficient sources
South Africa	Vereeniging	2009	insufficient sources
South Africa	Durban	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Port Shepstone	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Pietermaritzburg	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Ubomba	2009	insufficient sources
South Africa	Ladysmith	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Mtuzini	2002	https://www.submarinenetworks.com/en/systems/euro-africa/sat-3
South Africa	Ulundi	2009	insufficient sources
South Africa	Vryheid	2009	insufficient sources
South Africa	Lebowakgomo	2009	insufficient sources
South Africa	Polokwane	2004	https://www.commsupdate.com/articles/2004/08/25/tel-one-rolls-out-radio-link-to-south-africa/
South Africa	Tzaneen	2009	insufficient sources
South Africa	Musina	2004	https://www.commsupdate.com/articles/2004/08/25/tel-one-rolls-out-radio-link-to-south-africa/
South Africa	Thohoyandou	2009	insufficient sources
South Africa	Komatipoort	2006	https://www.fomsn.com/networks/fiber/fiber-optic-network-links-mozambique-and-south-africa/
South Africa	Mbombela	2006	https://www.fomsn.com/networks/fiber/fiber-optic-network-links-mozambique-and-south-africa/
South Africa	Bethal	2009	insufficient sources
South Africa	Standerton	2009	insufficient sources
South Africa	Volksrust	2009	insufficient sources
South Africa	Middelburg	2006	https://www.fomsn.com/networks/fiber/fiber-optic-network-links-mozambique-and-south-africa/
South Africa	Brits	2004	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
South Africa	Rustenburg	2004	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
South Africa	Klerksdorp	2009	insufficient sources
South Africa	Potchefstroom	2009	insufficient sources
South Africa	Bloemhof	2009	insufficient sources
South Africa	Vryburg	2009	insufficient sources
South Africa	Mmabatho	2004	https://researchictafrica.net/publications/Evidence_for_ICT_Policy_Action/Policy_Paper_1_-_Understanding_what_is_happening_in_ICT_in_Botswana.pdf
South Africa	Kimberley	2009	insufficient sources
South Africa	Poffader	2009	insufficient sources
South Africa	Springbok	1999	.https://www.oodaloop.com/documents/Legacy/CIA/factbook/geos/wa.html
South Africa	Alexander Bay	1999	.https://www.oodaloop.com/documents/Legacy/CIA/factbook/geos/wa.html
South Africa	Carnarvon	2009	insufficient sources

Table continues on the next page.

Country	city/town	connection	URL source
South Africa	Colesberg	2009	insufficient sources
South Africa	De Aar	2009	insufficient sources
South Africa	Prieska	2009	insufficient sources
South Africa	Kroonstad	2009	insufficient sources
South Africa	Welkom	2009	insufficient sources
South Africa	Bloemfontein	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Bethlehem	2009	insufficient sources
South Africa	Paarl	1995	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Worcester	1995	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Beaufort West	1995	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Melkbosstrand	1993	https://www.submarinenetworks.com/en/stations/africa/south-africa/melkbosstrand-cls
South Africa	Cape Town	1993	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	George	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Mossel Bay	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Oudtshoorn	2009	insufficient sources
South Africa	Bredasdorp	2009	insufficient sources
South Africa	Hermanus	2009	insufficient sources
South Africa	Swellendam	1999	https://www.telkom.co.za/history/TelkomHistory/index.html
South Africa	Saldanha	1999	.https://www.oodaloop.com/documents/Legacy/CIA/factbook/geos/wa.html
South Africa	Vanhynsdorp	1999	.https://www.oodaloop.com/documents/Legacy/CIA/factbook/geos/wa.html
Sudan	El Manaqil	2005	https://acir.yale.edu/sites/default/files/files/YaleLowensteinSudanReport.pdf
Sudan	Wad Madani	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Sudan	Ad Damazin	2005	https://acir.yale.edu/sites/default/files/files/YaleLowensteinSudanReport.pdf
Sudan	Gedaref	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Sudan	Kassala	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Sudan	Khartoum	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Sudan	Omdurman	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Sudan	El Fasher	2005	https://acir.yale.edu/sites/default/files/files/YaleLowensteinSudanReport.pdf
Sudan	El Obeid	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Sudan	Dongola	2004	https://www.sudantribune.com/spip.php?article6196
Sudan	Merowe	2004	https://www.sudantribune.com/spip.php?article6196
Sudan	Wadi Halfa	2004	https://www.sudantribune.com/spip.php?article6196

Table continues on the next page.

Country	city/town	connection	URL source
Sudan	Haiya	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Sudan	Port Sudan	2003	http://www.fiberatlantic.com/system/YE5Ln
Sudan	Atbara	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Sudan	Berber	2005	https://acir.yale.edu/sites/default/files/files/YaleLowensteinSudanReport.pdf
Sudan	Sennar	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Sudan	Nyala	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Sudan	Babanusa	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Sudan	Geneina	2005	https://acir.yale.edu/sites/default/files/files/YaleLowensteinSudanReport.pdf
Sudan	Muglad	2005	https://acir.yale.edu/sites/default/files/files/YaleLowensteinSudanReport.pdf
Sudan	Ed Dueim	2005	https://acir.yale.edu/sites/default/files/files/YaleLowensteinSudanReport.pdf
Sudan	Kosti	2004	https://books.google.de/books?id=WAs7lGNkVBkC&pg=PA217&lpg=PA217&dq=fiber+optic+cable+sudan+2004&source=bl&ots=cK8Dtz-0UR&sig=ACfU3U0ifVfHqC23ZJDM79R00SUWuBx3cQ&hl=de&sa=X&ved=2ahUKEwjzsYbnlJjqAhUHGuwKHbS1B2gQ6AEw-AHoECAkQAQ#v=onepage&q=fiber%20optic%20cable%20sudan%202004&f=false
Swaziland	Mbabane	2008	https://www.budde.com.au/Research/Eswatini-Swaziland-Telecoms-Mobile-and-Broadband-Statistics-and-Analyses
Swaziland	Piggs Peak	2009	https://books.google.de/books?id=oJutDwAAQBAJ&pg=PA196&lpg=PA196&dq=fiber+optic+network+swaziland+2008&source=bl&ots=F9C9z-OGFaG&sig=ACfU3U2RGIsygPIUq0exvMLmsMn2uq3pmg&hl=de&sa=X&ved=2ahUKEwiPj-TSnanqAhVF3KQKHf2jCDQQ6AEw-AHoECAcQAQ#v=onepage&q=fiber%20optic%20network%20swaziland%202008&f=false
Swaziland	Siteki	2009	https://books.google.de/books?id=oJutDwAAQBAJ&pg=PA196&lpg=PA196&dq=fiber+optic+network+swaziland+2008&source=bl&ots=F9C9z-OGFaG&sig=ACfU3U2RGIsygPIUq0exvMLmsMn2uq3pmg&hl=de&sa=X&ved=2ahUKEwiPj-TSnanqAhVF3KQKHf2jCDQQ6AEw-AHoECAcQAQ#v=onepage&q=fiber%20optic%20network%20swaziland%202008&f=false
Swaziland	Manzini	2009	https://books.google.de/books?id=oJutDwAAQBAJ&pg=PA196&lpg=PA196&dq=fiber+optic+network+swaziland+2008&source=bl&ots=F9C9z-OGFaG&sig=ACfU3U2RGIsygPIUq0exvMLmsMn2uq3pmg&hl=de&sa=X&ved=2ahUKEwiPj-TSnanqAhVF3KQKHf2jCDQQ6AEw-AHoECAcQAQ#v=onepage&q=fiber%20optic%20network%20swaziland%202008&f=false
Swaziland	Lavumisa	2009	https://books.google.de/books?id=oJutDwAAQBAJ&pg=PA196&lpg=PA196&dq=fiber+optic+network+swaziland+2008&source=bl&ots=F9C9z-OGFaG&sig=ACfU3U2RGIsygPIUq0exvMLmsMn2uq3pmg&hl=de&sa=X&ved=2ahUKEwiPj-TSnanqAhVF3KQKHf2jCDQQ6AEw-AHoECAcQAQ#v=onepage&q=fiber%20optic%20network%20swaziland%202008&f=false
Tanzania	Arusha	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Tanzania	Dar es Salaam	2009	https://www.zuj.edu.jo/conferences/ICIT11/PaperList/Papers/Computer%20Networks%20&%20Communications/627_Koloseni.pdf
Tanzania	Dodoma	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Tanzania	Moshi	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Tanzania	Same	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Tanzania	Lindi	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Tanzania	Babati	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Tanzania	Kilosa	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Tanzania	Morogoro	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf

Table continues on the next page.

Country	city/town	connection	URL source
Tanzania	Kibiti	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Tanzania	Manyoni	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Tanzania	Singida	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Tanzania	Tabora	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Tanzania	Tanga	2009	http://ijcir.mak.ac.ug/volume10-issue1/article6.pdf
Togo	Sokodé	2005	https://books.google.de/books?id=Hqt8eQNuRaoC&pg=PA41&lpg=PA41&dq=fiber+optic+cable+togo+burkina+faso+2006&source=bl&ots=A-zya4sPwq&sig=ACfU3U0Lv-bGFR3LrEJmvI7ir-7JI1NS3A&hl=de&sa=X&ved=2ahUKEwi6-4WpxIbqAhXKjqQKHcR0CbAQ6AEw-AHoECAcQAQ#v=onepage&q=fiber%20optic%20cable%20togo%20burkina%20faso%202006&f=false
Togo	Sotouboua	2005	https://books.google.de/books?id=Hqt8eQNuRaoC&pg=PA41&lpg=PA41&dq=fiber+optic+cable+togo+burkina+faso+2006&source=bl&ots=A-zya4sPwq&sig=ACfU3U0Lv-bGFR3LrEJmvI7ir-7JI1NS3A&hl=de&sa=X&ved=2ahUKEwi6-4WpxIbqAhXKjqQKHcR0CbAQ6AEw-AHoECAcQAQ#v=onepage&q=fiber%20optic%20cable%20togo%20burkina%20faso%202006&f=false
Togo	Lomé	2005	https://books.google.de/books?id=Hqt8eQNuRaoC&pg=PA41&lpg=PA41&dq=fiber+optic+cable+togo+burkina+faso+2006&source=bl&ots=A-zya4sPwq&sig=ACfU3U0Lv-bGFR3LrEJmvI7ir-7JI1NS3A&hl=de&sa=X&ved=2ahUKEwi6-4WpxIbqAhXKjqQKHcR0CbAQ6AEw-AHoECAcQAQ#v=onepage&q=fiber%20optic%20cable%20togo%20burkina%20faso%202006&f=false
Togo	Atakpamé	2005	https://books.google.de/books?id=Hqt8eQNuRaoC&pg=PA41&lpg=PA41&dq=fiber+optic+cable+togo+burkina+faso+2006&source=bl&ots=A-zya4sPwq&sig=ACfU3U0Lv-bGFR3LrEJmvI7ir-7JI1NS3A&hl=de&sa=X&ved=2ahUKEwi6-4WpxIbqAhXKjqQKHcR0CbAQ6AEw-AHoECAcQAQ#v=onepage&q=fiber%20optic%20cable%20togo%20burkina%20faso%202006&f=false
Togo	Mango	2005	https://books.google.de/books?id=Hqt8eQNuRaoC&pg=PA41&lpg=PA41&dq=fiber+optic+cable+togo+burkina+faso+2006&source=bl&ots=A-zya4sPwq&sig=ACfU3U0Lv-bGFR3LrEJmvI7ir-7JI1NS3A&hl=de&sa=X&ved=2ahUKEwi6-4WpxIbqAhXKjqQKHcR0CbAQ6AEw-AHoECAcQAQ#v=onepage&q=fiber%20optic%20cable%20togo%20burkina%20faso%202006&f=false
Uganda	Kampala	2009	https://www.commsupdate.com/articles/2006/06/13/mtn-uganda-extends-fibre-optic-network/
Uganda	Masaka	2009	https://www.commsupdate.com/articles/2006/06/13/mtn-uganda-extends-fibre-optic-network/
Uganda	Entebbe	2009	https://www.commsupdate.com/articles/2006/06/13/mtn-uganda-extends-fibre-optic-network/
Uganda	Busia	2009	https://www.commsupdate.com/articles/2009/07/07/kdn-builds-out-seacom-link/
Uganda	Jinja	2009	https://www.commsupdate.com/articles/2006/06/13/mtn-uganda-extends-fibre-optic-network/
Uganda	Tororo	2009	https://www.commsupdate.com/articles/2009/07/07/kdn-builds-out-seacom-link/
Uganda	Mbarara	2009	https://www.commsupdate.com/articles/2006/06/13/mtn-uganda-extends-fibre-optic-network/
Zambia	Kabwe	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zambia	Kapiri Mposhi	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zambia	Chililabombwe	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zambia	Chingola	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zambia	Kitwe	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zambia	Luanshya	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zambia	Mufulira	2009	https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.scribd.com%2Fdocument%2F274569230%2FZe-Scoop-Tic-Phase-i-i-20062012&psig=AOvVaw3VnkgSbaRyyBKkKOcAVBcN&ust=1593259795755000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCNCu1MC5n-oCFQAAAAAdAAAAABAY
Zambia	Ndola	2009	https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.scribd.com%2Fdocument%2F274569230%2FZe-Scoop-Tic-Phase-i-i-20062012&psig=AOvVaw3VnkgSbaRyyBKkKOcAVBcN&ust=1593259795755000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCNCu1MC5n-oCFQAAAAAdAAAAABAY
Zambia	Lusaka	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zambia	Solwezi	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zambia	Choma	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zambia	Kafue	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zambia	Livingstone	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zambia	Mazabuka	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html

Table continues on the next page.

Country	city/town	connection	URL source
Zambia	Sesheke	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zimbabwe	Bulawayo	2004	https://www.commsupdate.com/articles/2004/08/25/tel-one-rolls-out-radio-link-to-south-africa/
Zimbabwe	Harare	2004	https://www.commsupdate.com/articles/2004/08/25/tel-one-rolls-out-radio-link-to-south-africa/
Zimbabwe	Mutare	2009	https://ppiaf.org/documents/3160/download
Zimbabwe	Chinhoyi	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zimbabwe	Kadoma	2004	https://www.commsupdate.com/articles/2004/08/25/tel-one-rolls-out-radio-link-to-south-africa/
Zimbabwe	Kariba	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zimbabwe	Karoi	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zimbabwe	Victoria Falls	2007	https://www.networkworld.com/article/2278133/zesco-to-complete-zambia-fiber-backbone.html
Zimbabwe	Beitbridge	2004	https://www.commsupdate.com/articles/2004/08/25/tel-one-rolls-out-radio-link-to-south-africa/
Zimbabwe	Gweru	2004	https://www.commsupdate.com/articles/2004/08/25/tel-one-rolls-out-radio-link-to-south-africa/
Zimbabwe	Kwekwe	2004	https://www.commsupdate.com/articles/2004/08/25/tel-one-rolls-out-radio-link-to-south-africa/

Note: Source URLs last accessed in June and July, 2020. Extensive documentation and copies of primary sources available upon request. *Source:* own research.

B Supplementary Materials to Chapter 2

B.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).[1] 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.[2] 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*

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

[2] See Figure B.4 for the size distribution.

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 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 2.2. SCI data on the county-county level is taken from Bailey et al. (2018a)[3] and aggregated to economic-area level using the methodology suggested in Bailey et al. (2021):

$$\text{SCI}_{i,j} = \sum_{r_i \in R(i)} \sum_{r_j \in R(j)} \text{PopShare}_{r_i} * \text{PopShare}_{r_j} * \text{SCI}_{r_i, r_j} \tag{B.1}$$

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

$$\text{I} \rightarrow \frac{\text{I} - \min(\text{I})}{\max(\text{I}) - \min(\text{I})} * [\text{S}_{\max} - \text{S}_{\min}] + \text{S}_{\min} \tag{B.2}$$

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 B.1 used to aggregate the index to economic-area level from Bailey et al. (2021):

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

[3] Data retrieved online via: `data.humdata.org/dataset/social-connectedness-index`, last accessed 03/11/2023.

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

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

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

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 "Professional, Scientific, and Technical Services" (NAICS Rev. 2 code 54, "tech GDP") to the economic-area level using the crosswalk between counties and economic areas from Moretti (2021)[4] and take averages for the years 2014 to 2020.[5] From the *U.S. Census Bureau* I use county-level population estimates and apply the same aggregation procedure.[6] 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

[4] Retrieved at: `https://www.openicpsr.org/openicpsr/project/140581/version/V1/view;jsessionid=2BBE031DF440387A3F4EA8416E38D449`, last accessed 03/11/2023.
[5] Retrieved at: `https://www.bea.gov/data/gdp/gdp-county-metro-and-other-areas`, last accessed 03/11/2023.
[6] Retrieved at: `https://www.census.gov/data/datasets/time-series/demo/popest/2010s-counties-total.html`, last accessed 03/11/2023.

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).[7] Inventor latitude and longitude are assigned to economic areas using the economic area shape file by the *Bureau of Transportation Statistics*.[8] 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.

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

B.2 Tables

Table B.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 B.2: Sensitivity to colocation definition

Collaboration [log]	distance cutoff		
	(1)	(2)	(3)
	$= 0$ km	< 100 km	< 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 2.1, 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 B.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 2.1 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 B.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 2.1, 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 B.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 2.1, 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 B.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 2.1, 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 B.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 2.1, 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 B.8: Colocation and project types

Collaboration	(1)	users		commits		age	
		(2)	(3)	(4)	(5)	(6)	(7)
	baseline	≥ 3	< 3	$\geq$ median	$<$ median	$\geq$ median	$<$ 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 2.1, 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 B.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 2.1, 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 B.10: Colocation and economic-area characteristics

Collaboration		# local users		avg. firm size	
	(1)	(2)	(3)	(4)	(5)
	baseline	≥ median	Top 10	≥ median	≥ 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 2.1, defining colocation as indicator of being in the same economic area. Models (2) through (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 B.11: Colocation and strong versus weak ties

Collaboration		projects		commits			
				median		minimum	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	baseline	> 1	$= 1$	above	below	> 2	≤ 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 2.1, 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.

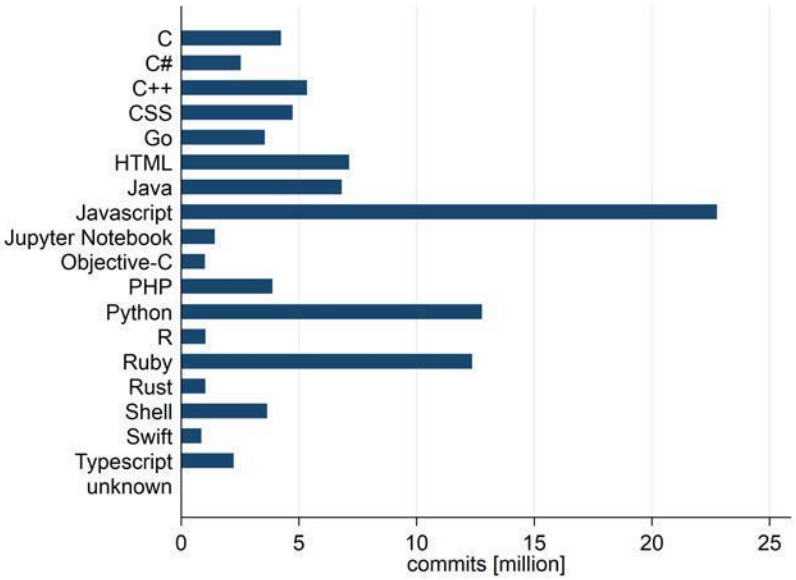
Table B.12: Colocation and collaboration intensity

Collaboration	(1)	counts		ratios	
		(2)	(3)	(4)	(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 2.1, 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.

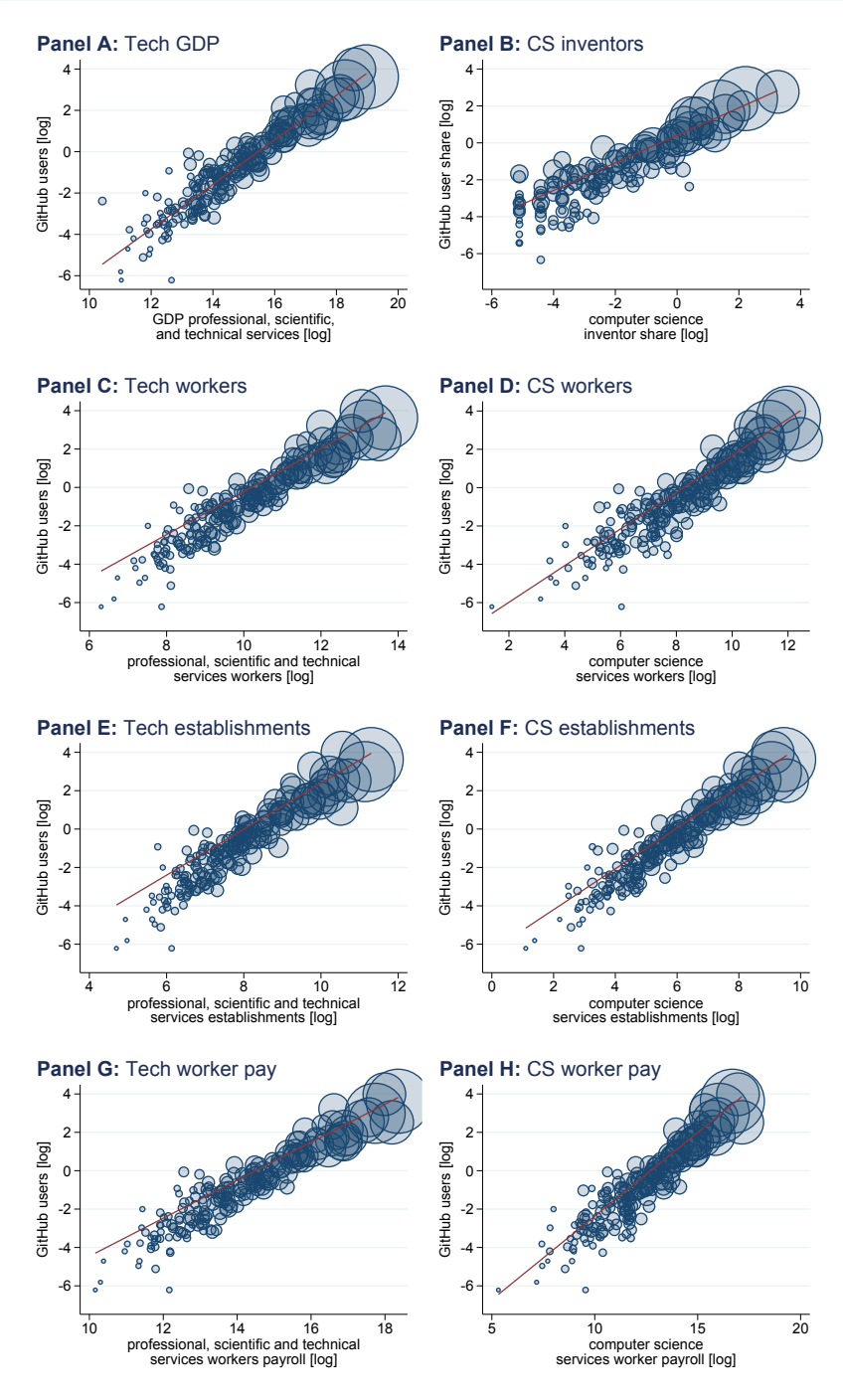
B.3 Figures

Figure B.1: Programming languages



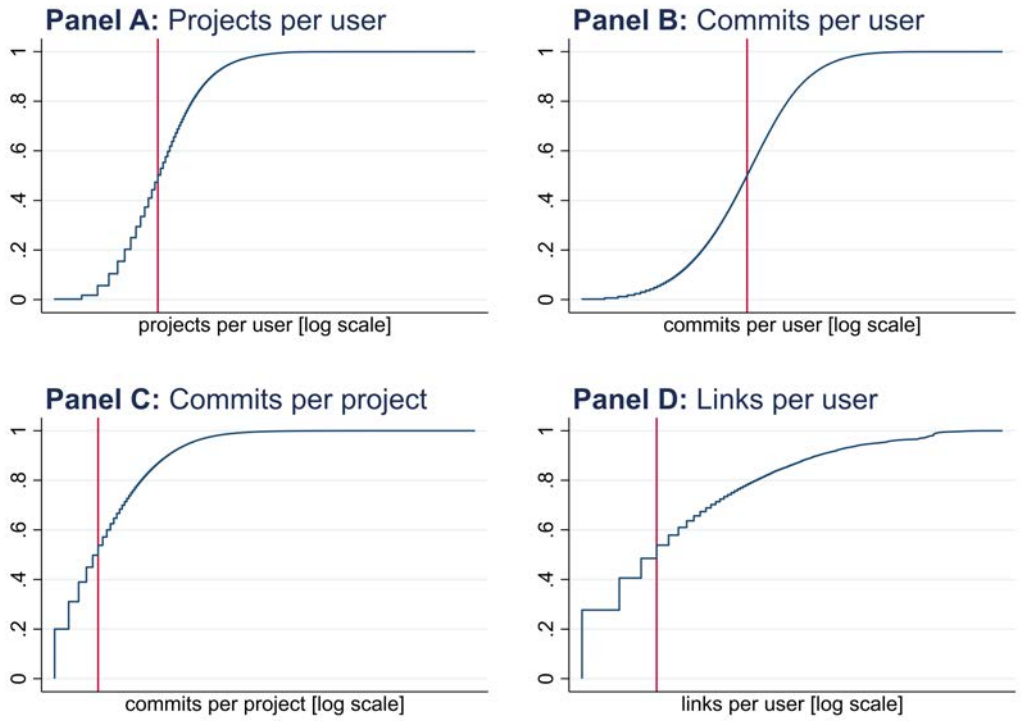
Note: Bars show the number of *commits* contributed to open-source projects by active, collaborating users in the United States in the observation period for each programming language. Unknown refers to *commits* that are not assigned to a programming language in the data. *Sources:* GHTorrent, own calculations.

Figure B.2: Representativness



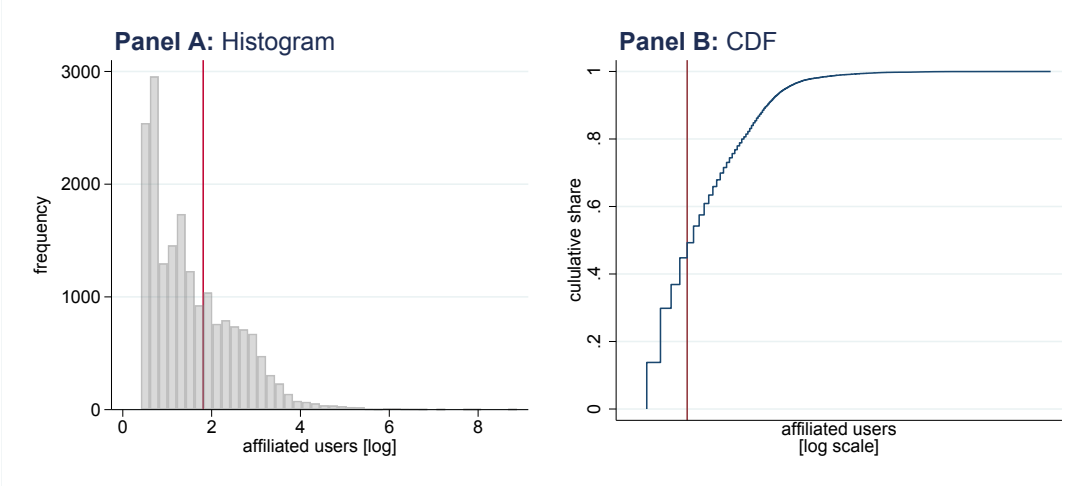
Note: Plots show the relationship between (the share of) users per economic area and economic-area level metrics related to software development after logarithmic transformation. Bubble size represents economic-area population size. Red lines are best linear fits from user-weighted log-log regressions.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

Figure B.3: CDFs of user activity



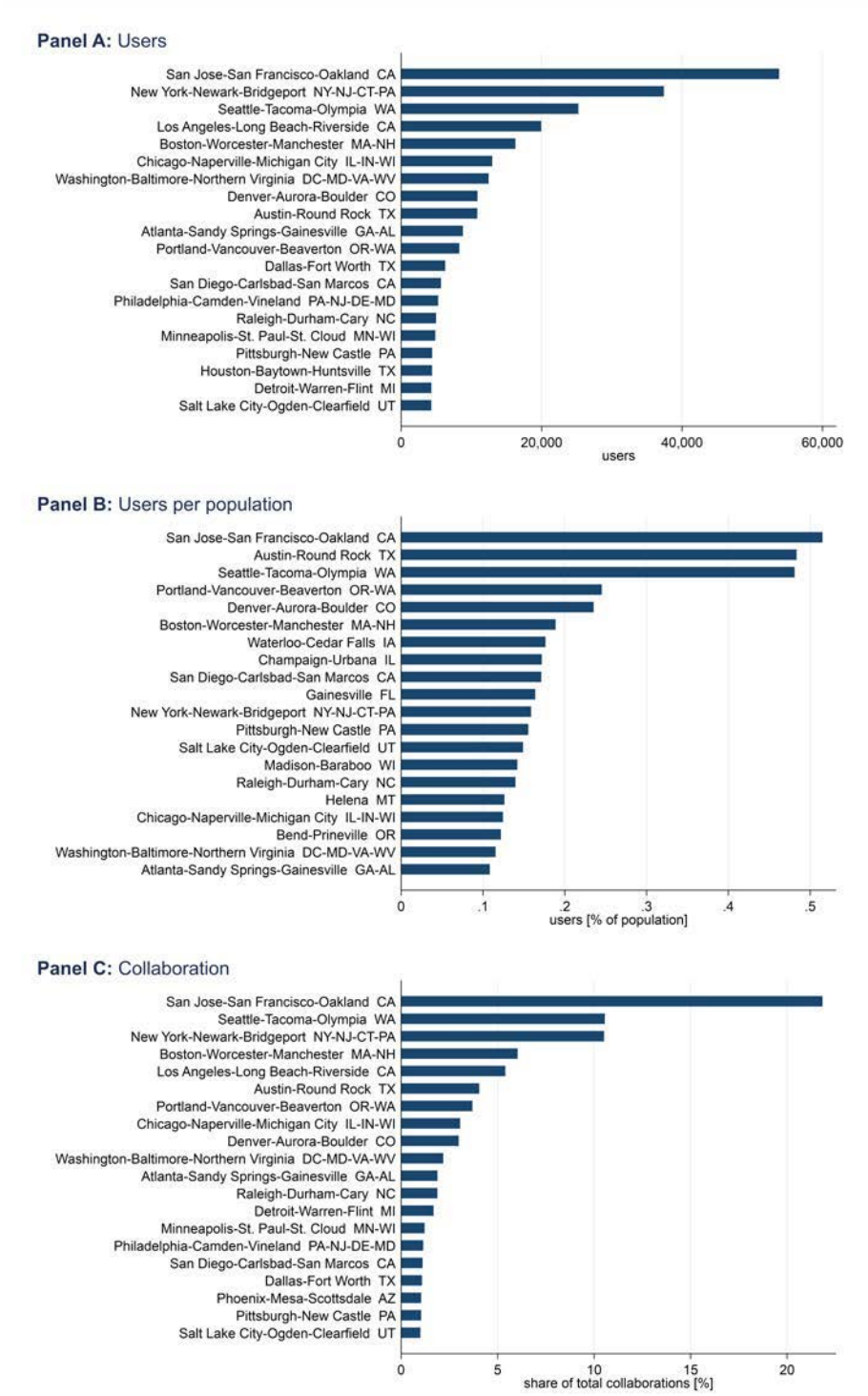
Note: Plots show cumulative density functions for different user activity metrics. Vertical red lines represent median values of each metric (i.e., projects per user: 14; *commits* per user: 156; *commits* per project: 7; links per user: 4). All x-axes are scaled logarithmically. The graph for *commits*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

Figure B.4: Organization size



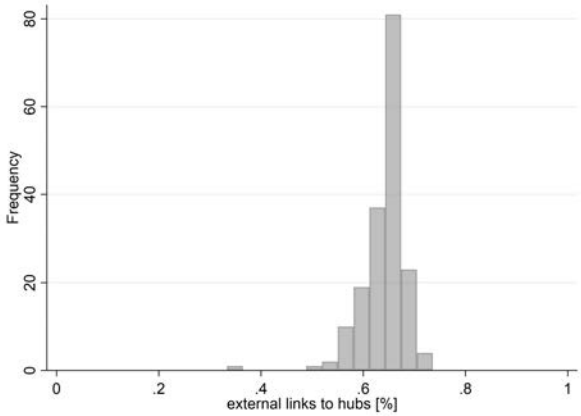
Notes: Plot shows the distribution of organization size as measured by number of affiliated users. Panel A shows a histogram and Panel B a cumulative distribution function. The horizontal red line indicates mean (6.1; histogram) and median (3.5; CDF) affiliated users. Organizations with only one affiliated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

Figure B.5: Concentration at the top



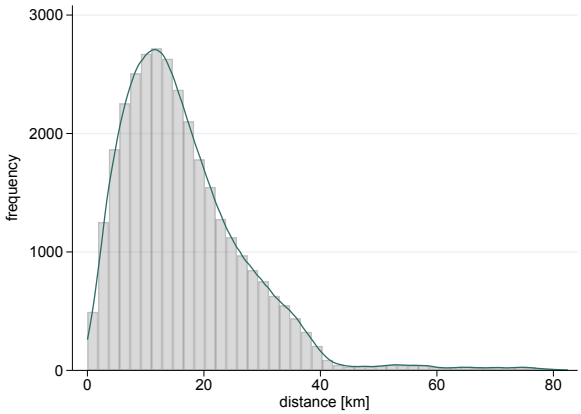
Notes: Plots show the values of different user and activity concentration metrics for the twenty largest economic areas in terms of respective metrics. *Sources:* GHTorrent, own calculations.

Figure B.6: Collaboration with hubs



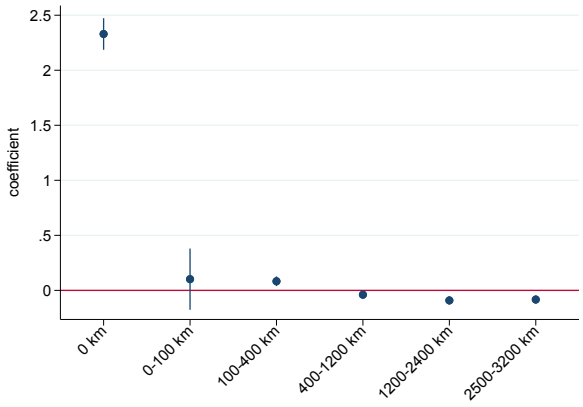
Notes: Plot shows the distribution of collaboration shares of each economic area with hubs, defined as the ten largest economic areas in terms of users. *Sources:* GHTorrent, Bureau of Economic Analysis, own calculations.

Figure B.7: Distance



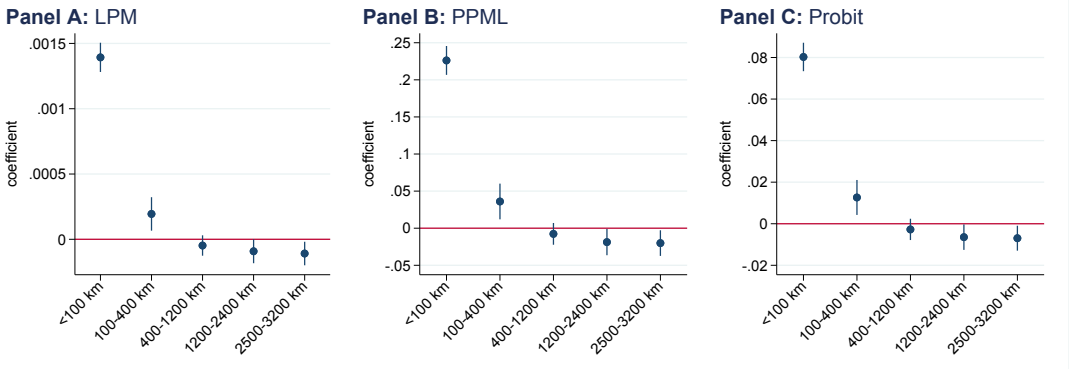
Notes: Plot shows the distribution of centroid-based geodesic distance between economic areas. The horizontal red line indicates the median distance of 1,439. The blue curve represents the Epanechnikov kernel density estimate. The right tail of the distribution starting approximately at distances greater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



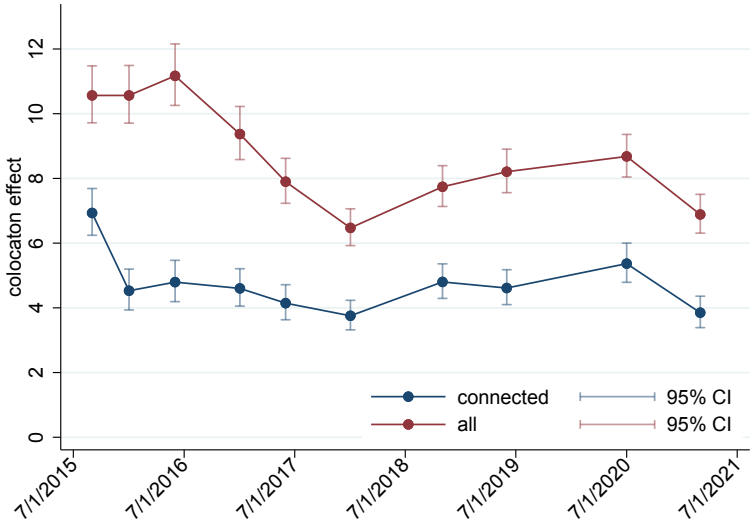
Notes: Plot shows coefficient point estimates and confidence intervals for the baseline fixed effects model specification with non-parametric distance. The indicator for distances above 3,200 km is omitted. Blue bars show 95% confidence intervals from robust standard errors. Collaborations with Anchorage,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

Figure B.9: Individual-level probability models



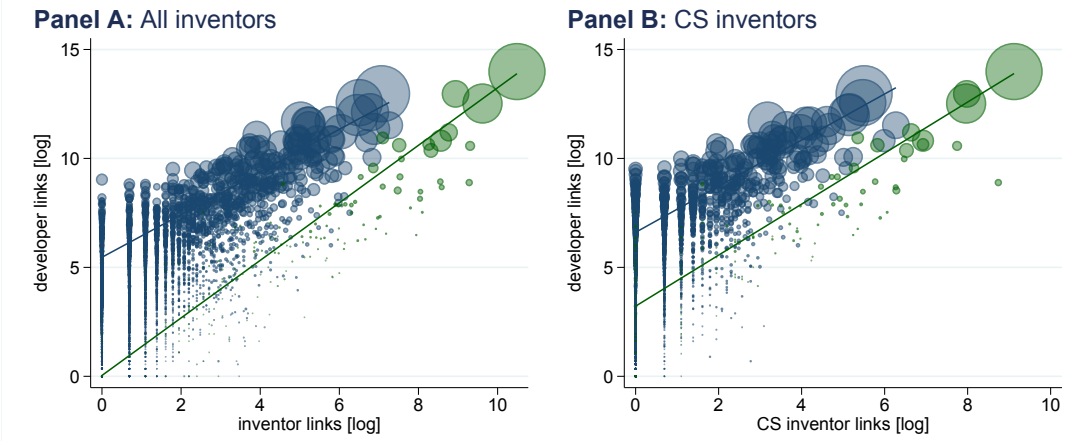
Notes: Plot shows coefficient point estimates and confidence intervals for the individual-level fixed effects model specification with non-parametric distance from Table B.4. The indicator for distances above 3,200 km is omitted. Blue bars show 95% confidence intervals from robust standard errors. Collaborations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

Figure B.10: Colocation dynamics



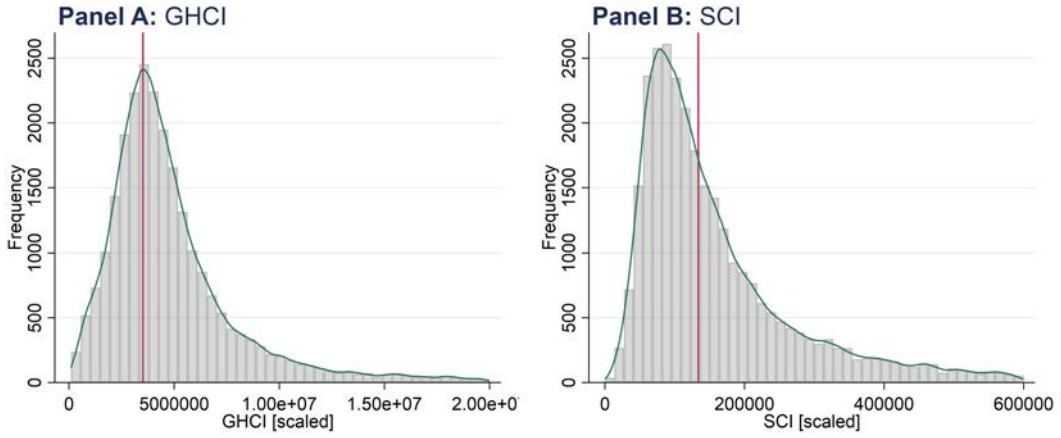
Notes: Plot shows coefficient point estimates and confidence intervals for the baseline fixed effects model specification with non-parametric distance. The indicator for distances above 3,200 km is omitted. Blue bars show 95% confidence intervals from robust standard errors. Collaborations with Anchorage,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

Figure B.11: Colocation effect relative to inventors



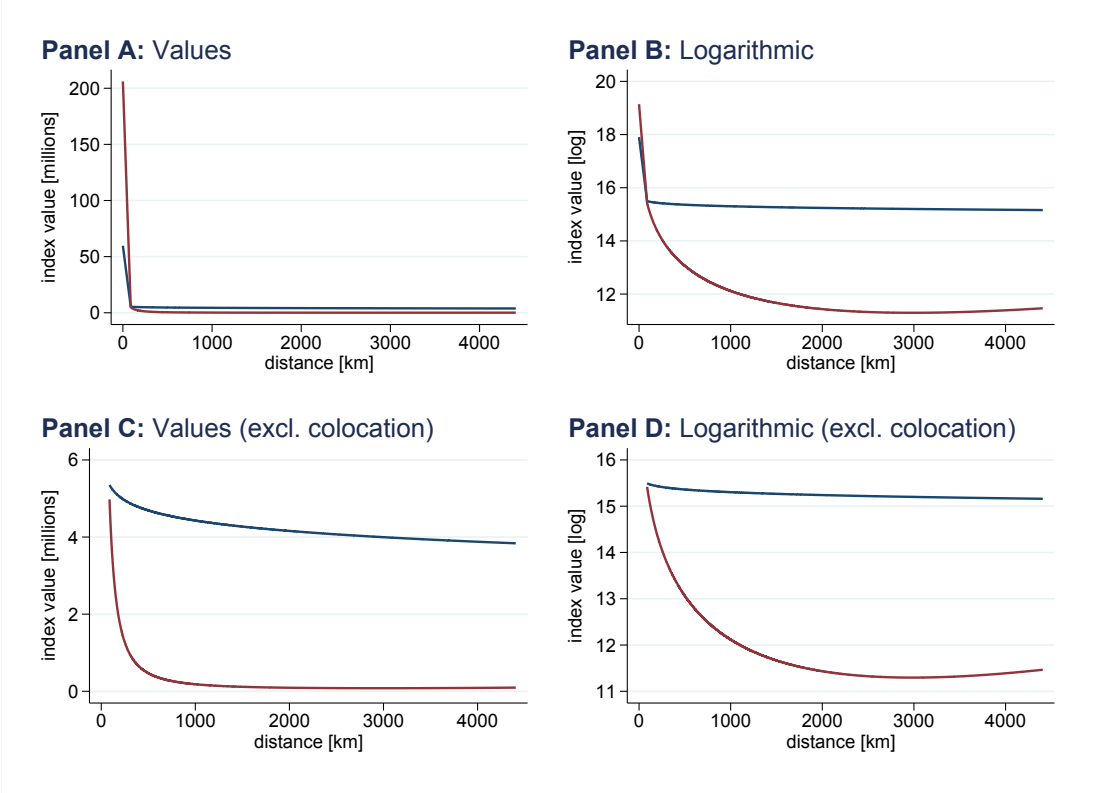
Note: Plots show the relationship between the number of collaborations between economic areas in the software developer and inventor network. Panel A compares software developer collaborations to all collaborations in collaborative patents and Panel B to collaborative computer science patents. Collaborations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



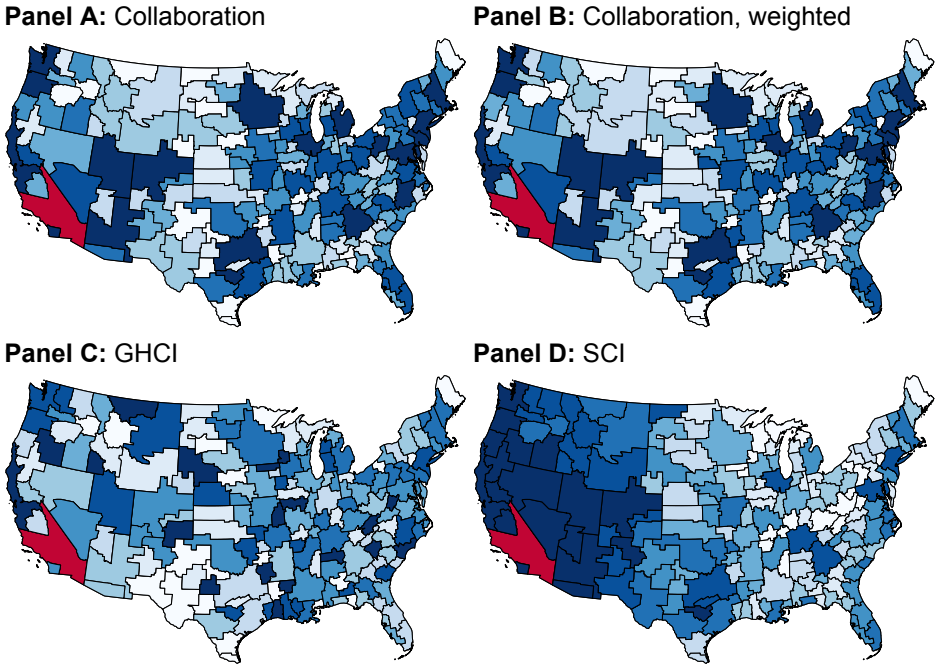
Note: Plots show the distribution of scaled GHCI and SCI regional connectedness indices. The horizontal red lines indicate medians of 133,753 for the GHCI and 3,518,538 for the SCI. The blue curves represent the Epanechnikov kernel density estimates. Both indices are scaled between 1 and 1,000,000,000. Scaled SCI from Bailey et al. (2018a) 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. (2018a), Bureau of Economic Analysis, own calculations.

Figure B.13: Spatial decay



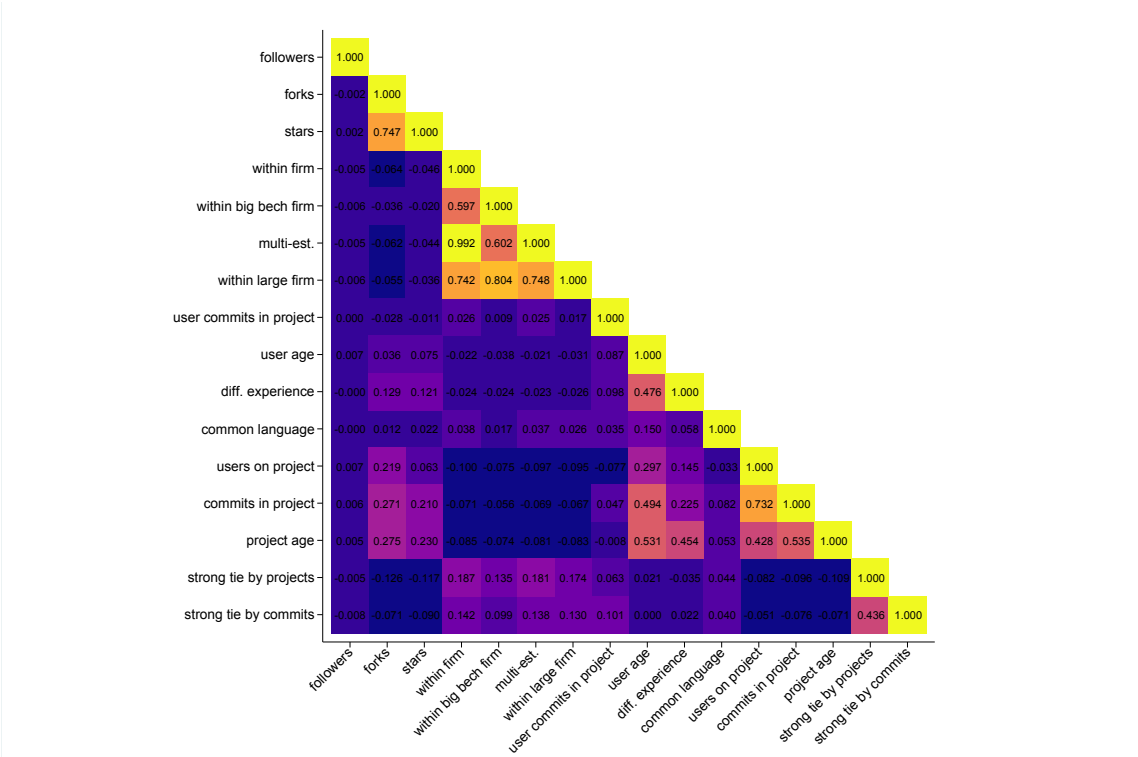
Note: Plot shows spatial decay as predicted per fractional polynomial model with (Panels A and B) and without (Panels C and D) the colocation effect and in values (Panels A and C) and logarithmically (Panels B and D). *Sources:* GHTorrent, Bailey et al. (2018a), Bureau of Economic Analysis, own calculations.

Figure B.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. (2018a), own calculations.

Figure B.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.

C Supplementary Materials to Chapter 3

C.1 Tables

Table C.1: Users by country

ISO2	Country	Users	Share
UK	United Kingdom	32,914	22.8%
FR	France	23,516	16.3%
DE	Germany	21,211	14.7%
PL	Poland	10,293	7.1%
NL	Netherlands	9,371	6.5%
ES	Spain	7,104	4.9%
IT	Italy	5,167	3.6%
CZ	Czech Republic	3,701	2.6%
SE	Sweden	3,692	2.6%
FI	Finland	3,660	2.5%
DK	Denmark	3,227	2.2%
AT	Austria	3,021	2.1%
CH	Switzerland	2,637	1.8%
BE	Belgium	2,136	1.5%
NO	Norway	1,897	1.3%
RO	Romania	1,863	1.3%
EL	Greece	1,682	1.2%
PT	Portugal	1,534	1.1%
HR	Croatia	965	0.7%
RS	Serbia	740	0.5%
	Other	3,790	2.6%
	Total	144,121	100%

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

Table C.2: Border effect and country size

Collaboration	(1)	(2)	(3)
Cross-border	-0.180***	-0.133***	-0.269***
	(0.014)	(0.014)	(0.022)
Cross-border × small involved		-0.155***	
		(0.012)	
Cross-border × both small			0.034
			(0.022)
Cross-border × both large			0.129***
			(0.020)
Colocation	0.862***	0.879***	0.888***
	(0.068)	(0.068)	(0.068)
Distance [log]	-0.129***	-0.119***	-0.120***
	(0.007)	(0.007)	(0.007)
Origin FE	×	×	×
Destination FE	×	×	×
Observations	84,100	84,100	84,100
Adj. R^2	0.922	0.922	0.922

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. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, own calculations.

Table C.3: Collaboration and interests

Collaboration	(1)	(2)	(3)
Cross-border	-0.414***	-0.212***	-0.004
	(0.011)	(0.013)	(0.032)
Colocation		1.132***	1.436***
		(0.067)	(0.070)
Distance [log]		-0.084***	-0.025***
		(0.007)	(0.008)
Business and Industry			0.918**
			(0.409)
Education			0.000
			(0.164)
Family and Relationships			-0.700***
			(0.185)
Fitness and Wellness			1.704***
			(0.552)
Food and Drink			1.153**
			(0.473)
Hobbies and Activities			2.089***
			(0.372)
Lifestyle and Culture			3.788***
			(0.427)
News and Entertainment			6.952***
			(0.795)
Non-local Business			-17.013***
			(2.024)
People			0.287***
			(0.068)
Shopping and Fashion			0.595
			(0.435)
Sports and Outdoors			0.152
			(0.163)
Technology			1.035***
			(0.299)
Travel, Places and Events			1.074***
			(0.266)
Other			-1.000
			(0.737)
Origin FE	×	×	×
Destination FE	×	×	×
Observations	77,284	77,284	77,284
Adj. R^2	0.929	0.932	0.933

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. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, Obradovich et al. (2022), own calculations.

Table C.4: Collaboration and preferences

Collaboration	(1)	(2)	(3)
Cross-border	-0.361***	-0.229***	-0.158***
	(0.011)	(0.012)	(0.017)
Colocation		1.310***	1.360***
		(0.066)	(0.068)
Distance [log]		-0.044***	-0.033***
		(0.007)	(0.007)
Patience			-0.118***
			(0.017)
Risk taking			-0.036
			(0.049)
Positive reciprocity			-0.094***
			(0.034)
Negative reciprocity			-0.040**
			(0.017)
Altruism			-0.033
			(0.027)
Trust			-0.015
			(0.020)
Origin FE	×	×	×
Destination FE	×	×	×
Observations	48,888	48,888	48,888
Adj. R^2	0.951	0.954	0.955

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. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, Falk et al. (2018), CEPII, own calculations.

Table C.5: Collaboration and cultural dimensions

Collaboration	(1)	(2)	(3)
Cross-border	-0.396***	-0.248***	-0.221***
	(0.011)	(0.012)	(0.016)
Colocation		1.312***	1.317***
		(0.066)	(0.067)
Distance [log]		-0.048***	-0.047***
		(0.006)	(0.007)
Power distance			-0.034***
			(0.006)
Individualism			-0.022*
			(0.012)
Achievement and success			0.002
			(0.004)
Uncertainty avoidance			0.010*
			(0.006)
Long-term orientation			-0.001
			(0.006)
Indulgence			0.001
			(0.006)
Origin FE	×	×	×
Destination FE	×	×	×
Observations	67,828	67,828	67,828
Adj. R^2	0.939	0.941	0.941

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. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, Hofstede (2011), own calculations.

Table C.6: Border effect in the United States

Collaboration	(1)	(2)	(3)	(4)
Cross-border	-0.527***	-0.429***	-0.502***	-0.100***
	(0.098)	(0.041)	(0.037)	(0.033)
Users, multiplied [log]		0.750***		
		(0.004)		
Colocation				2.191***
				(0.073)
Distance [log]				-0.060***
				(0.011)
Origin FE			×	×
Destination FE			×	×
Observations	32,041	32,041	32,041	32,041
Adj. R^2	0.002	0.856	0.917	0.922
Border effect	-41.0%	-34.9%	-39.4%	-9.5%
Δ(Europe – USA)	-18.6 p.p.	+3.9 p.p.	+3.4 p.p.	-6.9 p.p.
BE_{USA} / BE_{Europe}	0.69	1.13	1.09	0.58

Notes: The outcome variable is the natural logarithm of collaborations between two economic areas plus one. Colocation indicates collaboration between users in the same economic area. Distance is scaled in 100km. Users, GDPs, and Populations refers to the respective variables for both origin and destination. Users, multiplied, is the multiplication of the number of users in origin and destination. Collaboration with Anchorage, AK, and Honolulu, HI, are excluded. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, Bureau of Economic Analysis, Goldbeck (2023), own calculations.

Table C.7: Collaboration and history

Collaboration	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Cross-border	0.000	0.032	-0.010	-0.010	-0.008	-0.006	0.048
	(0.037)	(0.042)	(0.041)	(0.041)	(0.041)	(0.037)	(0.043)
Colocation	1.469***	1.441***	1.447***	1.447***	1.473***	1.465***	1.490***
	(0.069)	(0.070)	(0.069)	(0.069)	(0.069)	(0.069)	(0.069)
Distance [log]	-0.007	-0.011	-0.008	-0.008	-0.006	-0.007	-0.002
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Cultural distance	-0.068***	-0.073***	-0.059***	-0.059***	-0.065***	-0.065***	-0.064***
	(0.017)	(0.017)	(0.018)	(0.018)	(0.018)	(0.017)	(0.017)
Genetic distance	-0.001*	-0.001	-0.001*	-0.001*	-0.001*	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Common language	0.069**	0.078**	0.103***	0.103***	0.073**	0.066**	0.071**
	(0.033)	(0.033)	(0.034)	(0.034)	(0.034)	(0.033)	(0.033)
Religious distance	-0.000	0.002	0.016	0.016	-0.001	0.004	-0.001
	(0.020)	(0.020)	(0.021)	(0.021)	(0.021)	(0.020)	(0.020)
Same country history	-0.081***	-0.078***	-0.072***	-0.072***	-0.080***	-0.116***	-0.091***
	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
Colonial history	0.001	0.011	0.023	0.023	0.001	0.005	0.007
	(0.015)	(0.016)	(0.017)	(0.017)	(0.015)	(0.015)	(0.015)
Social connectedness	0.016***	0.017***	0.019***	0.019***	0.016***	0.015***	0.018***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Contiguity		-0.020*					
		(0.010)					
Common legal origin			-0.037***				
			(0.009)				
Common legal origin (post-transformation)				-0.037***			
				(0.009)			
Common legal origin (pre-transformation)					-0.003		
					(0.011)		
Communist history						0.141***	
						(0.041)	
Iron curtain							0.059**
							(0.027)
Origin FE	×	×	×	×	×	×	×
Destination FE	×	×	×	×	×	×	×
Observations	54,702	54,702	54,630	54,630	54,630	54,702	54,702
Adj. R^2	0.949	0.949	0.949	0.949	0.949	0.949	0.949

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. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, Obradovich et al. (2022), Creanza et al. (2015), Bailey et al. (2018b), CEPII, own calculations.

Table C.8: Collaboration and political systems

Collaboration	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Cross-border	0.013	0.008	0.013	0.047	-0.003	0.008	0.003	0.000
	(0.038)	(0.038)	(0.038)	(0.044)	(0.044)	(0.037)	(0.037)	(0.037)
Colocation	1.472***	1.464***	1.471***	1.462***	1.472***	1.449***	1.469***	1.469***
	(0.070)	(0.070)	(0.070)	(0.070)	(0.070)	(0.070)	(0.069)	(0.069)
Distance [log]	-0.009	-0.010	-0.009	-0.010	-0.009	-0.014*	-0.006	-0.007
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Cultural distance	-0.080***	-0.081***	-0.080***	-0.076***	-0.081***	-0.077***	-0.068***	-0.068***
	(0.017)	(0.017)	(0.017)	(0.019)	(0.018)	(0.017)	(0.017)	(0.017)
Genetic distance	-0.001*	-0.001**	-0.001*	-0.001*	-0.001*	-0.001**	-0.001*	-0.001*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Common language	0.062*	0.055	0.062*	0.066*	0.061*	0.070**	0.068**	0.069**
	(0.034)	(0.035)	(0.034)	(0.034)	(0.034)	(0.034)	(0.033)	(0.033)
Religious distance	-0.007	-0.005	-0.007	0.001	-0.007	0.003	-0.002	-0.001
	(0.020)	(0.021)	(0.020)	(0.021)	(0.020)	(0.021)	(0.020)	(0.020)
Same country history	-0.078***	-0.079***	-0.078***	-0.076***	-0.080***	-0.073***	-0.081***	-0.081***
	(0.028)	(0.028)	(0.028)	(0.028)	(0.029)	(0.028)	(0.028)	(0.028)
Colonial history	0.001	0.002	0.001	0.003	0.017	0.004	0.001	0.001
	(0.016)	(0.016)	(0.016)	(0.016)	(0.033)	(0.016)	(0.015)	(0.015)
Social connectedness	0.013***	0.013***	0.013***	0.012***	0.013***	0.011**	0.017***	0.016***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Diplomatic disagreement		0.017						
		(0.018)						
EU			-0.020					
			(0.048)					
RTA				-0.044***				
				(0.013)				
Hegemon					-0.019			
					(0.033)			
Monarchies						-0.045***		
						(0.015)		
Δ economic freedom							-0.008	
							(0.018)	
Δ political rights								0.007
								(0.037)
Origin FE	×	×	×	×	×	×	×	×
Destination FE	×	×	×	×	×	×	×	×
Observations	55,169	55,169	55,169	55,097	55,169	55,169	54,702	54,702
Adj. R^2	0.947	0.947	0.947	0.947	0.947	0.947	0.949	0.949

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. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, Obradovich et al. (2022), Creanza et al. (2015), Bailey et al. (2018b), Graafland and de Jong (2022), CEPII, own calculations.

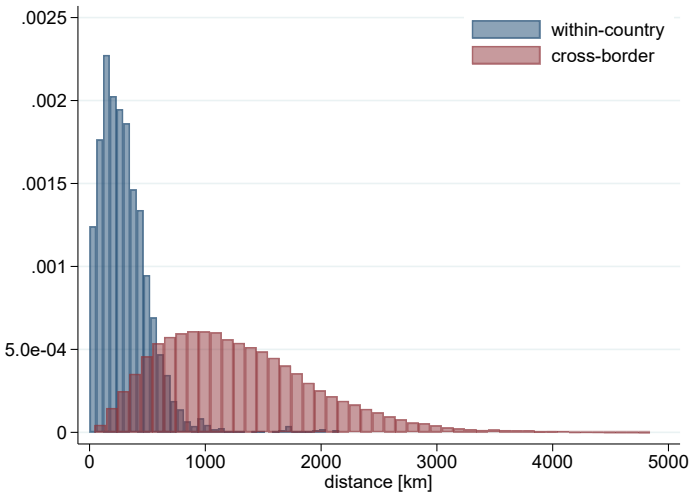
Table C.9: Collaboration, language, and religion

Collaboration	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Cross-border	0.013	0.027	0.023	0.033	0.024	0.024	0.021
	(0.038)	(0.037)	(0.043)	(0.048)	(0.037)	(0.037)	(0.040)
Colocation	1.472***	1.460***	1.461***	1.462***	1.462***	1.463***	1.477***
	(0.070)	(0.070)	(0.070)	(0.070)	(0.070)	(0.070)	(0.070)
Distance [log]	-0.009	-0.010	-0.010	-0.009	-0.010	-0.010	-0.008
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.009)
Cultural distance	-0.080***	-0.090***	-0.092***	-0.092***	-0.089***	-0.089***	-0.079***
	(0.017)	(0.018)	(0.018)	(0.018)	(0.017)	(0.017)	(0.017)
Genetic distance	-0.001*	-0.001***	-0.001**	-0.001**	-0.001**	-0.001**	-0.001*
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Same country history	-0.078***	-0.081***	-0.080***	-0.081***	-0.081***	-0.081***	-0.077***
	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)	(0.028)
Colonial history	0.001	-0.000	0.001	0.003	0.002	0.002	0.003
	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)	(0.016)
Social connectedness	0.013***	0.014***	0.014***	0.014***	0.014***	0.014***	0.012***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Common spoken language	0.062*						0.064*
	(0.034)						(0.035)
Common native language		0.013					
		(0.025)					
Linguistic proximity (Tree)			0.001				
			(0.003)				
Linguistic proximity (ASJP)				0.002			
				(0.004)			
Common Language Index [log]					0.018		
					(0.028)		
Common Language Index [level]						0.019	
						(0.028)	
Religious distance	-0.007	-0.009	-0.012	-0.013	-0.011	-0.011	
	(0.020)	(0.020)	(0.021)	(0.021)	(0.020)	(0.020)	
Religious proximity [Fearon weighted]							0.003
							(0.008)
Origin FE	×	×	×	×	×	×	×
Destination FE	×	×	×	×	×	×	×
Observations	55,169	55,169	55,097	55,097	55,169	55,169	54,702
Adj. R^2	0.947	0.947	0.947	0.947	0.947	0.947	0.947

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. Robust standard errors are reported in parenthesis. *** p<0.01, ** p<0.05, * p<0.1. *Sources:* GHTorrent, Obradovich et al. (2022), Creanza et al. (2015), Bailey et al. (2018b), CEPII, own calculations.

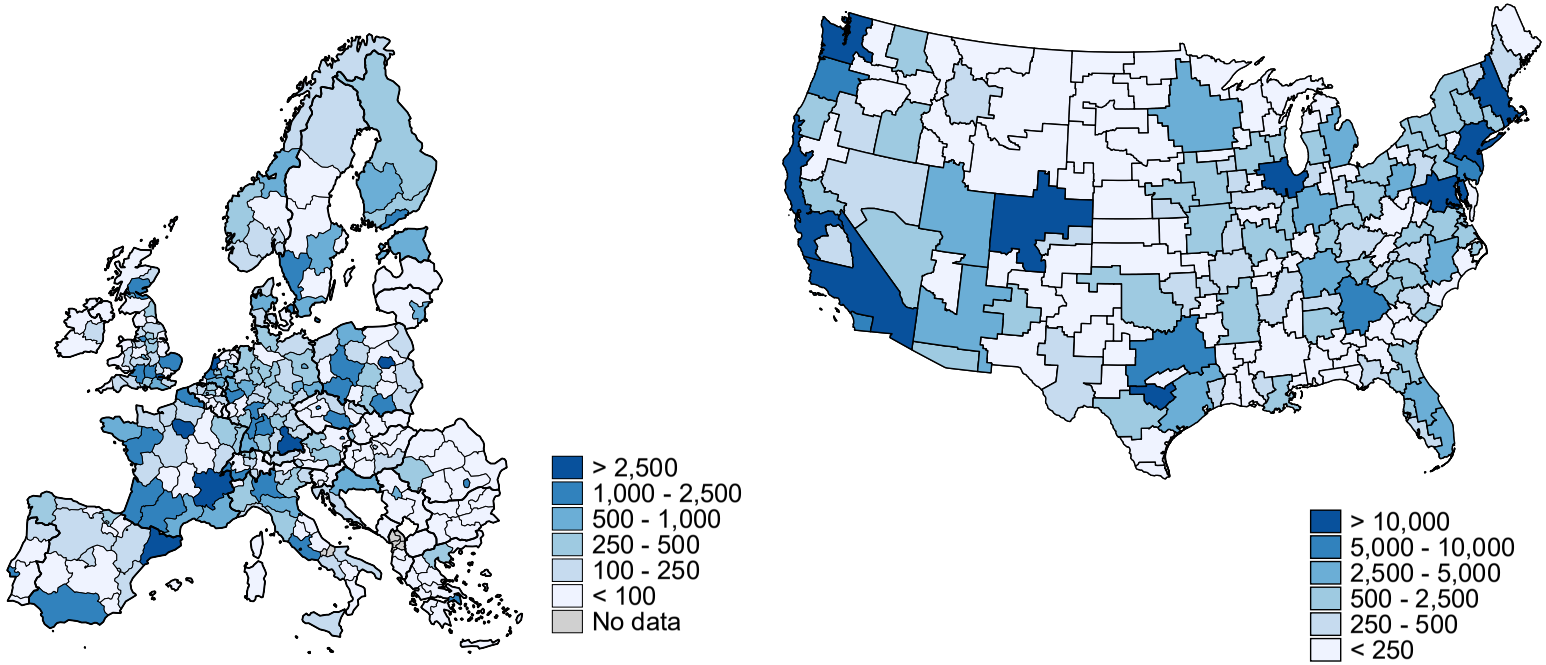
C.2 Figures

Figure C.1: Distance histogram



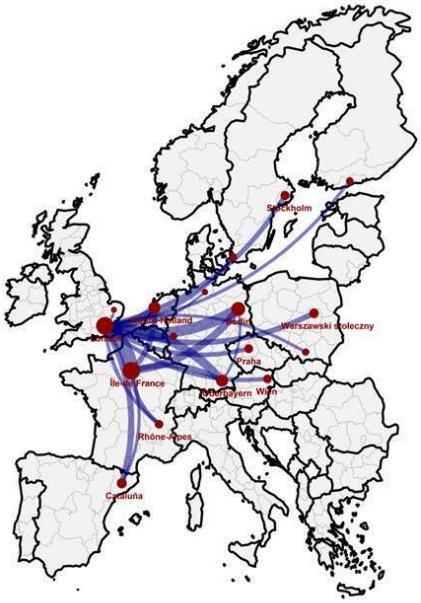
Notes: Figure shows histograms of within-country and cross-border distances based on NUTS2 centroids, respectively. *Sources:* GHTorrent, own calculations.

Figure C.2: Geographic user distribution

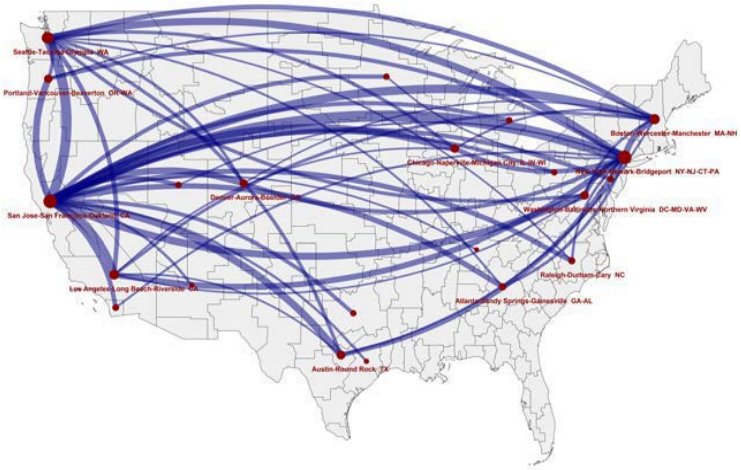


Notes: Maps show the number of (in-sample) users per NUTS2 region and economic area, respectively. The remote economic areas Anchorage, AK, and Honolulu, HI, as well as Ireland are not shown. *Sources:* GHTorrent, Bureau of Economic Analysis, Goldbeck (2023), own calculations.

Figure C.3: Inter-regional collaboration



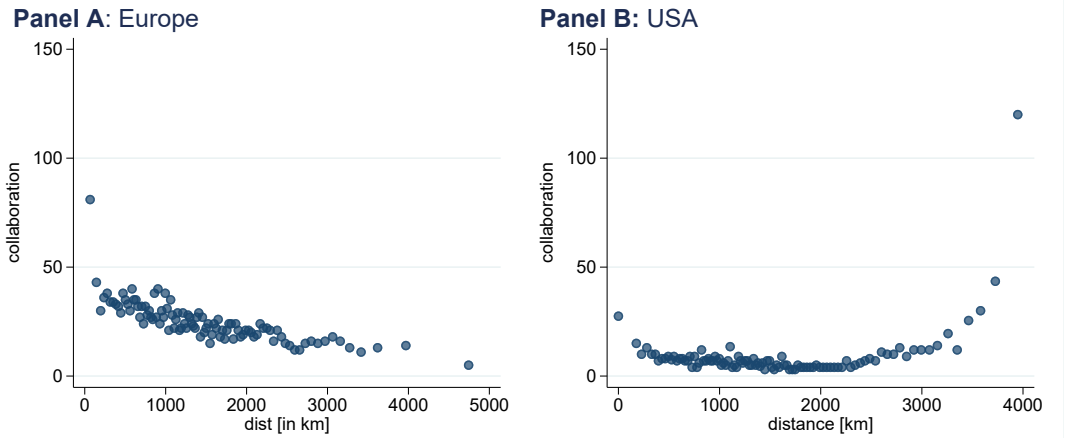
(a) Europe



(b) USA

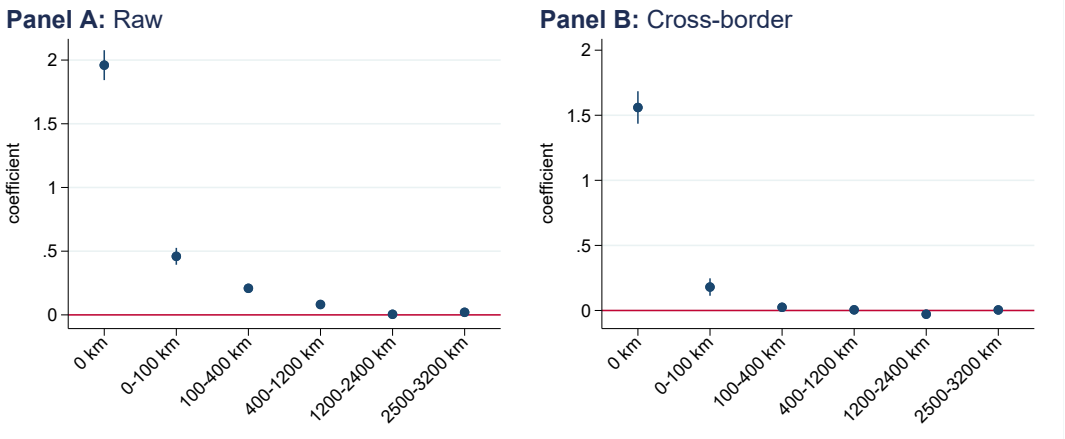
Notes: Maps show the structure of the European and US software developer collaboration networks, respectively. Important edges of the network, defined as links between economic areas above 25,000 connections, are shown in blue and scaled by the logarithm of the number of links. Regions are shown in gray with their centroids as nodes in red, scaled by overall links to other economic areas. The remote economic areas Anchorage, AK, and Honolulu, HI, as well as Ireland are not shown. *Sources:* GHTorrent, Bureau of Economic Analysis, Goldbeck (2023), own calculations.

Figure C.4: Collaboration and distance



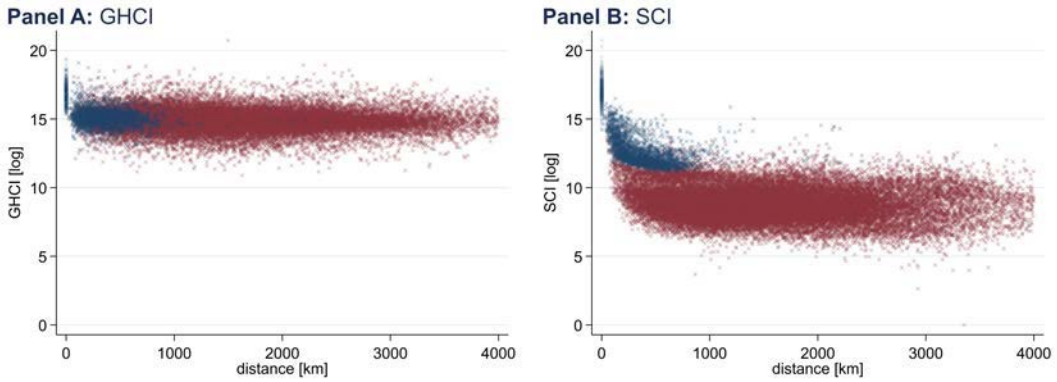
Notes: Panels A and B show binned scatter plots of the median number of collaborations and the geographic distance between economic-area pairs in Europe and the US, respectively. The number of bins is 100, i.e., each point represents one percentile of economic-area pairs. *Sources:* GHTorrent, own calculations.

Figure C.5: Non-parametric distance



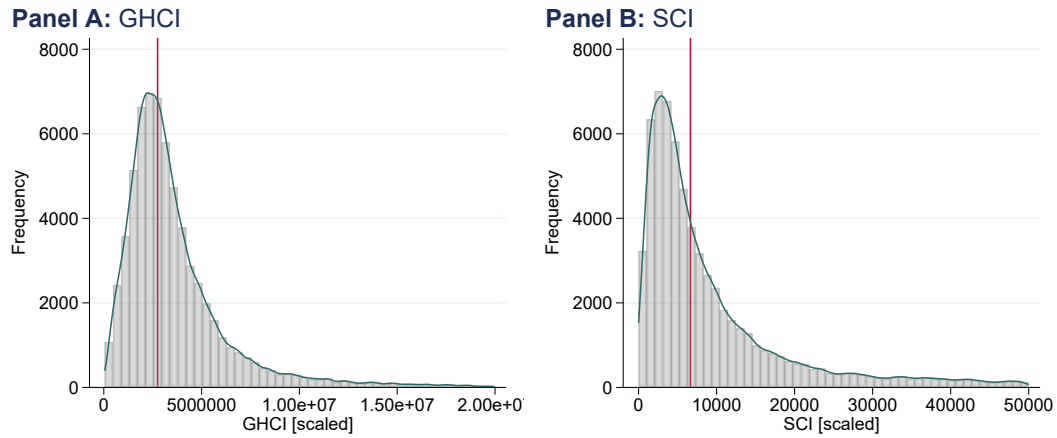
Notes: Plot shows coefficient point estimates and confidence intervals for the baseline fixed effects model specification with non-parametric distance. Panel A (Panel B) shows results from a specification without (with) cross-border indicator. The indicator for distances above 3,200 km is omitted. Blue bars show 95% confidence intervals from robust standard errors. *Sources:* GHTorrent, own calculations.

Figure C.6: Border effect



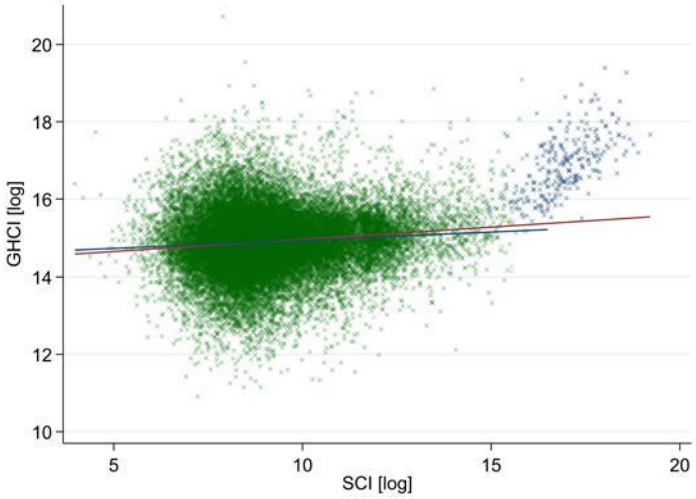
Notes: Figure shows scattered values of scaled GHCI (Panel A) and scaled SCI (Panel B) after logarithmic transformation. Both indices are scaled between 1 and 1,000,000,000. Scaled SCI from Bailey et al. (2018b) is mean-aggregated from county-county level weighted by multiplied populations of each county-pair and rescaled between 1 and 1,000,000,000. Within-country (cross-border) observations are shown in blue (red). *Sources:* GHTorrent, Bailey et al. (2018b), own calculations.

Figure C.7: Distribution of connectedness indices



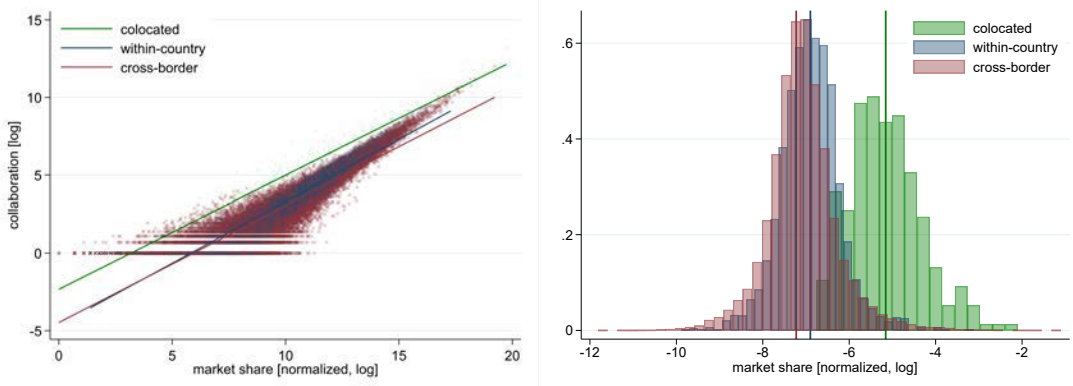
Notes: Plots show the distribution of scaled GHCI and SCI regional connectedness indices. The horizontal red lines indicate medians of 6,650 for the SCI and 2,750,304 for the GHCI. The blue curves represent the Epanechnikov kernel density estimates. Both indices are scaled between 1 and 1,000,000,000. 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, we restrict the y-axes to maximum values of 20,000,000 for GHCI and 50,000 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), own calculations.

Figure C.8: Relatedness GHCI and SCI



Notes: Figure shows the correlation between scaled GHCI and SCI after logarithmic transformation with within-regional collaborations excluded. Colocated collaborations are colored blue. *Sources:* GHTorrent, Bailey et al. (2018b), own calculations.

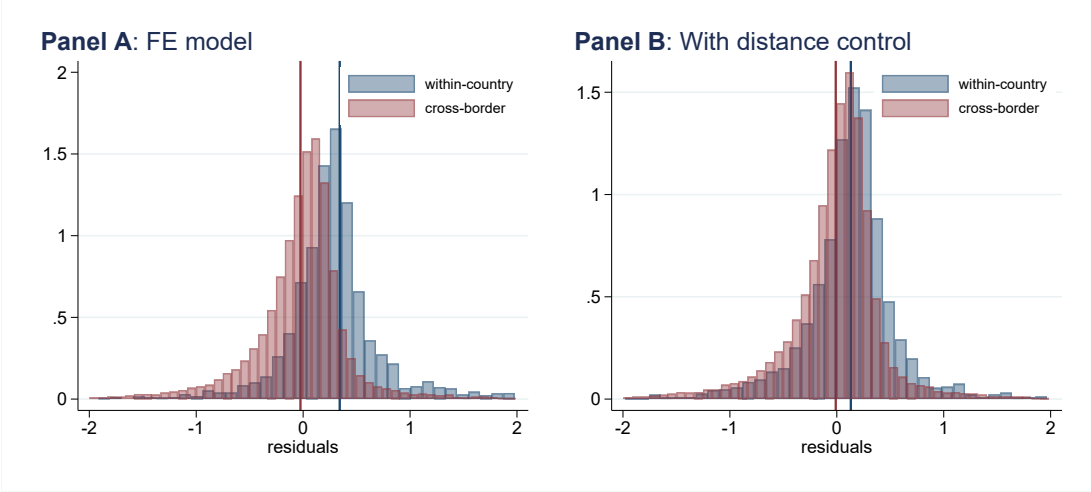
Figure C.9: Independence benchmark



(a) market share **(b)** distribution

Note: Figure shows the independence benchmark following Santamaría et al. (2023b) for colocated (green) within-country (blue) and cross-border (red) collaboration, respectively. *Sources:* GHTorrent, own calculations.

Figure C.10: Fixed-effect model residuals



Notes: Figure shows residual histograms for within-country and cross-border collaboration, respectively. Panel A (Panel B) depicts residuals from the baseline fixed-effects model without (with) controls. *Sources:* GHTorrent, own calculations.

D Supplementary Materials to Chapter 4

D.1 Tables

Table D.1: Sample selection

Median	All users	Movers	Δ
Activity			
Commits	6.00	170.00	164.00
commits single projects	*2.00*	*73.00*	*71.00*
commits team projects	*1.00*	*65.00*	*64.00*
Experience	34.00	39.00	5.00
Collaboration			
Projects	2.00	15.00	13.00
single projects	*2.00*	*9.00*	*7.00*
team projects	*2.00*	*5.00*	*3.00*
Quality			
Followers	0.00	5.00	5.00
Stars	0.00	1.30	1.30
stars single projects	*0.00*	*0.10*	*0.10*
Forks	0.00	0.76	0.76
forks single projects	*0.00*	*0.00*	*0.00*

Notes: Experience is measured as tenure on the platform in months since the first commit at the move date. Column Δ reports the absolute difference in median between movers in our sample and all users in the ten *GHTorrent* snapshots we utilize (N = 28,802,543). Column %Δ sets this difference in relation to other movers' median. *Sources:* GHTorrent, own calculations.

Table D.2: Affiliation and job transitions

Affiliation	all movers	job movers	other movers	Δ
Largest 100 firms	28.9 %	28.9 %	27.2 %	+1.7 p.p.
Big tech	*7.2 %*	*7.3 %*	*4.9 %*	*+2.4 p.p.*
Academic	8.9 %	9.0 %	6.3 %	+2.7 p.p.
Other	55.1 %	54.8 %	61.6 %	-6.8 p.p.
Job transitions	anytime	origin	destination	Δ
Largest 100 firms	28.9 %	20.3 %	26.8 %	+6.5 p.p.
Big tech	*7.2 %*	*2.0 %*	*7.1 %*	*+5.1 p.p.*
Academic	8.9 %	9.1 %	7.2 %	-2.0 p.p.
Other	55.1 %	68.6 %	58.9 %	-9.6 p.p.

Notes: Table reports affiliations and job transitions by organization type in shares of the respective sample. Column Δ reports the percentage point difference between job and other movers. *Sources:* GHTorrent, own calculations.

Table D.3: Top origin and destination cities

Origin	**Users**	**Share**	**Destination**	**Users**	**Share**
New York, USA	650	2.84 %	San Francisco, USA	1,307	5.71 %
San Francisco, USA	618	2.70 %	New York, USA	936	4.09 %
London, UK	421	1.84 %	London, UK	763	3.33 %
Bangalore, India	325	1.42 %	Seattle, USA	708	3.09 %
Chicago, USA	311	1.36 %	Bangalore, India	559	2.44 %
Boston, USA	305	1.33 %	Los Angeles, USA	379	1.66 %
Los Angeles, USA	305	1.33 %	Austin, USA	345	1.51 %
Moscow, Russia	305	1.33 %	Toronto, Canada	331	1.45 %
Seattle, USA	273	1.19 %	Chicago, USA	318	1.39 %
Paris, France	247	1.08 %	Boston, USA	315	1.38 %
Cumulative share		15.09 %	Cumulative share		26.05 %

Notes: Table reports the ten largest origin and destination cities in terms of the number of users in our sample. *Sources:* GHTorrent, own calculations.

Table D.4: Domestic moves

Country	Users	Share	
		all	domestic
United States	10,348	45.20 %	63.49 %
India	1,219	5.32 %	7.48 %
United Kingdom	638	2.79 %	3.91 %
Canada	620	2.71 %	3.80 %
China	522	2.28 %	3.20 %
France	436	1.90 %	2.68 %
Germany	417	1.82 %	2.56 %
Russia	375	1.64 %	2.30 %
Poland	195	0.85 %	1.20 %
Australia	194	0.85 %	1.19 %
		65.36 %	91.81 %

Notes: Table reports the ten largest countries in terms of the number of domestic movers in our sample. Shares reported in the third and fourth columns refer to all and to domestic movers, respectively. *Sources:* GHTorrent, own calculations.

Table D.5: Top origin and destination countries

International movers					
Origin	*Users*	*Share*	*Destination*	*Users*	*Share*
United States	1,831	0.28	United States	2,011	0.30
India	817	0.12	United Kingdom	774	0.12
United Kingdom	491	0.07	Canada	506	0.08
Russia	386	0.06	Germany	319	0.05
Canada	384	0.06	Russia	306	0.05
France	267	0.04	Netherlands	290	0.04
Australia	186	0.03	Australia	240	0.04
Italy	165	0.03	Poland	228	0.03
Brazil	163	0.02	France	182	0.03
Germany	151	0.02	Brazil	169	0.03
Inter-continental movers					
Origin	*Users*	*Share*	*Destination*	*Users*	*Share*
United States	1,453	0.34	United States	1,583	0.37
India	793	0.18	United Kingdom	428	0.10
United Kingdom	284	0.07	Russia	287	0.07
Russia	203	0.05	Canada	275	0.06
Australia	180	0.04	Australia	229	0.05
France	144	0.03	Germany	177	0.04
China	130	0.03	Poland	159	0.04
Canada	105	0.02	France	116	0.03
Italy	72	0.02	Netherlands	111	0.03
Poland	72	0.02	Italy	96	0.02

Notes: Table reports the ten largest origin and destination countries in terms of the number of international and inter-continental movers in our sample. *Sources:* GHTorrent, own calculations.

Table D.6: Top origin and destination affiliations

Origin	Share	Destination	Share
Student	0.92 %	Microsoft	2.08 %
Microsoft	0.72 %	Google	2.00 %
University of Washington	0.62 %	Amazon	1.37 %
Freelancer	0.51 %	Facebook	1.00 %
IBM	0.41 %	Red Hat	0.64 %
New York University	0.41 %	Shopify	0.44 %
University of California	0.41 %	IBM	0.37 %
University of Florida	0.41 %	Stanford University	0.31 %
University of Oxford	0.41 %	LinkedIn	0.28 %
Amazon	0.31 %	Apple	0.26 %
	5.13 %		8.75 %

Notes: Table reports the ten most frequently stated affiliations as a percentage of all users with non-empty affiliation information. *Sources:* GHTorrent, own calculations.

Table D.7: Classification of programming languages

Classification	programming language	share	
		lang.	class.
App development	Ruby	5.68 %	
	Go	4.06 %	
	Swift	1.09 %	
	Objective-C	0.65 %	11.48 %
Data engineering	Python	13.03 %	
	R	1.22 %	
	Jupyter Notebook	1.18 %	
	Scala	0.89 %	16.32 %
Low-level programming	C++	5.37 %	
	C	3.33 %	
	C#	2.30 %	
	Rust	1.40 %	
	Assembly	0.08 %	12.48 %
Program routine	Shell	3.16 %	
	PowerShell	0.22 %	3.38 %
Web development	JavaScript	20.91 %	
	HTML	6.65 %	
	Java	6.19 %	
	PHP	4.36 %	
	CSS	4.28 %	
	TypeScript	3.21 %	42.39 %
Other			10.74 %

Notes: The 27 most-used programming languages in terms of commits in the *GHTorrent* are classified, 21 of which are represented in our sample. Classified programming languages account for 89.26% of commits in our sample. *Sources:* GHTorrent, own calculations.

Table D.8: Top-paying programming languages

Classification	programming language	share		median pay	
		lang.	class. cumul.	lang.	class. avg.
Top 30 top-paying languages	Zig	0.009 %		$103,611	
	Erlang	0.145 %		$99,492	
	F#	0.091 %		$99,311	
	Ruby	5.749 %		$98,522	
	Clojure	0.399 %		$96,381	
	Elixir	0.383 %		$96,381	
	Scala	0.894 %		$96,381	
	Perl	0.491 %		$94,540	
	Go	4.087 %		$92,760	
	OCaml	0.365 %		$91,026	
	Objective-C	0.646 %		$90,000	
	Rust	1.365 %		$87,012	
	Swift	1.041 %		$86,897	
	Groovy	0.202 %		$86,271	
	Shell	3.347 %		$85,672	
	Haskell	0.771 %		$85,672	
	Apex	0.015 %		$81,552	
	PowerShell	0.23 %		$81,311	
	SAS	0.002 %		$81,000	
	Lua	0.312 %		$80,690	
	Nim	0.016 %		$80,000	
	Raku	0.001 %		$79,448	
	Python	12.933 %		$78,331	
	Kotlin	0.438 %		$78,207	
	APL	0 %		$77,500	
	Crystal	0.041 %		$77,104	
	TypeScript	3.074 %		$77,104	
	Assembly	0.078 %		$77,010	
	Fortran	0.132 %		$76,104	
	Cobol	0.001 %		$76,000	
	C#	2.314 %	39.572 %	$74,963	$86,008
Other top-paying languages	C++	5.516 %		$74,963	
	Julia	0.416 %		$74,963	
	R	1.217 %		$74,963	
	SQL	0.12 %		$74,963	
	C	3.438 %		$74,351	
	JavaScript	20.381 %		$74,034	
	Solidity	0.007 %		$72,701	
	Ada	0.013 %		$72,656	
	HTML	6.653 %		$71,500	
	CSS	4.264 %		$70,148	
	Prolog	0.018 %		$70,000	
	Delphi	0 %		$69,608	
	GDScript	0.021 %		$69,608	
	VBA	0.002 %		$65,698	
	Visual Basic	0.096 %		$65,000	
	Matlab	0.215 %		$61,735	
	PHP	4.375 %		$58,899	
	Dart	0.221 %	46.973 %	$55,862	$69,536
Not listed			13.455 %		

Notes: Table reports programming languages on the *StackOverflow* list of top-paying technologies. We further distinguish between the top 30 and other listed programming languages. Classified programming languages account for 86.54% of commits in our sample. *Sources:* GHTorrent, StackOverflow, own calculations.

Table D.9: Keywords

Cluster	keywords	% projects
Code	adventofcode; algorithm; algorithms; android; api; app; application; apps; c; class; framework; functions; game; hacktoberfest; ios; javascript; library; module; nodejs; plugin; python; react; server; software; template; testing; tictactoe; tool; ui	7.06
Website	blog; personal; personalwebsite; portfolio; resume; site; website	2.11
File	collection; docs; document; documentation; dotfiles; file; files; githubslideshow; presentation; presentations; scripts	1.17
Education	course; coursera; example; examples; exercise; exercises; freecodecamp; helloworld; homework; learning; nowgithubstarter; programmingassignment; repdata; peerassessment; test	0.85
Data	data; database	0.48
Other		13.06

Notes: Table reports keywords assigned to project type clusters. Projects may be assigned to multiple clusters. Keywords search is conducted in project descriptions; 24.73% of projects feature non-empty project descriptions. *Sources:* GHTorrent, own calculations.

Table D.10: Model specification

Model class:	OLS				LPM	NB	PPML
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent variable:	log	ihs	ihs	ihs	dummy	count	count
Sample:	full	full	geo	change	full	full	full
Job mover × job search	0.1326***	0.1646***	0.1654***	0.1384**	0.0711***	0.4983***	0.1358***
	(0.0119)	(0.0141)	(0.0142)	(0.0548)	(0.0039)	(0.0280)	(0.0521)
Job mover × post move	-0.0851***	-0.1036***	-0.1021***	-0.2804***	-0.0307***	-0.2690***	-0.1707**
	(0.0159)	(0.0190)	(0.0190)	(0.0849)	(0.0056)	(0.0453)	(0.0670)
User FE	×	×	×	×	×	×	×
Month FE	×	×	×	×	×	×	×
Experience FE	×	×	×	×	×	×	×
Adjusted R^2	0.35803	0.35945	0.35958	0.43305	0.34000		
Observations	1,946,413	1,946,413	1,941,317	76,797	1,946,413	1,630,215	1,630,215
# User FE	22,896	22,896	22,838	885	22,896	22,896	22,896

Notes: Results from estimation of Equation 4.2 for different model classes, outcome transformations, and sample definitions. Experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$. *Sources:* GHTorrent, own calculations.

Table D.11: Project ownership and initial forks

IHS(single commits)	project owner		
	(1) own	(2) non-own	(3) no initial forks
Job mover × job search	0.1310***	0.0428***	0.1440***
	(0.0138)	(0.0080)	(0.0149)
Job mover × post move	-0.1157***	0.0024	-0.1088***
	(0.0182)	(0.0097)	(0.0194)
User FE	×	×	×
Month FE	×	×	×
Experience FE	×	×	×
Adjusted R^2	0.33534	0.32483	0.32440
Observations	1,946,413	1,946,413	1,946,413
Users	22,896	22,896	22,896

Notes: Results from estimation of Equation 4.2 by repository ownership and without initial fork projects. Experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$. *Sources:* GHTorrent, own calculations.

Table D.12: Heterogeneity by project types (keywords)

IHS(single commits)	(1) education	(2) data	(3) website	(4) code	(5) files	(6) other
Job mover × job search	0.0030	0.0000	0.0135***	0.0307***	0.0154***	0.1097***
	(0.0024)	(0.0027)	(0.0043)	(0.0069)	(0.0037)	(0.0123)
Job mover × post move	-0.0091***	-0.0089***	-0.0049	-0.0335***	-0.0044	-0.0641***
	(0.0035)	(0.0031)	(0.0056)	(0.0090)	(0.0045)	(0.0163)
User FE	×	×	×	×	×	×
Month FE	×	×	×	×	×	×
Experience FE	×	×	×	×	×	×
Adjusted R^2	0.09276	0.10952	0.15628	0.16827	0.21257	0.31379
Observations	1,946,413	1,946,413	1,946,413	1,946,413	1,946,413	1,946,413
Users	22,896	22,896	22,896	22,896	22,896	22,896

Notes: Results from estimation of Equation 4.2 for different project types, according to keyword-based method. Experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$. *Sources:* GHTorrent, own calculations.

Table D.13: Event study coefficients

IHS(single commits)	(1)	(2)	(3)
Job mover × event_time = -21	0.0126	0.0025	0.0017
	(0.0206)	(0.0217)	(0.0215)
Job mover × event_time = -20	0.0178	0.0283	0.0326
	(0.0208)	(0.0217)	(0.0215)
Job mover × event_time = -19	-0.0397**	-0.0016	0.0042
	(0.0196)	(0.0208)	(0.0207)
Job mover × event_time = -18	-0.0555***	-0.0084	-0.0066
	(0.0193)	(0.0204)	(0.0203)
Job mover × event_time = -17	-0.0328*	-0.0167	-0.0145
	(0.0168)	(0.0178)	(0.0176)
Job mover × event_time = -15	0.1771***	0.0574***	0.0557***
	(0.0188)	(0.0197)	(0.0196)
Job mover × event_time = -14	0.5110***	0.1608***	0.1596***
	(0.0239)	(0.0251)	(0.0252)
Job mover × event_time = -13	0.5415***	0.1787***	0.1807***
	(0.0243)	(0.0251)	(0.0251)
Job mover × event_time = -12	0.6329***	0.2443***	0.2455***
	(0.0277)	(0.0282)	(0.0282)
Job mover × event_time = -11	0.5882***	0.1942***	0.1996***
	(0.0271)	(0.0276)	(0.0278)
Job mover × event_time = -10	0.5708***	0.1640***	0.1675***
	(0.0268)	(0.0272)	(0.0273)
Job mover × event_time = -9	0.4677***	0.1141***	0.1221***
	(0.0264)	(0.0270)	(0.0269)
Job mover × event_time = -8	0.4538***	0.1290***	0.1377***
	(0.0273)	(0.0278)	(0.0277)
Job mover × event_time = -7	0.4278***	0.1339***	0.1475***
	(0.0273)	(0.0278)	(0.0278)
Job mover × event_time = -6	0.4627***	0.1440***	0.1630***
	(0.0287)	(0.0293)	(0.0295)
Job mover × event_time = -5	0.4658***	0.1158***	0.1318***
	(0.0278)	(0.0284)	(0.0285)
Job mover × event_time = -4	0.3806***	0.0759***	0.0967***
	(0.0274)	(0.0276)	(0.0278)
Job mover × event_time = -3	0.3846***	0.0388	0.0654**
	(0.0265)	(0.0272)	(0.0272)
Job mover × event_time = -2	0.3617***	0.0416	0.0690**
	(0.0264)	(0.0271)	(0.0273)
Job mover × event_time = -1	0.4193***	0.0331	0.0738***
	(0.0275)	(0.0283)	(0.0285)
Job mover × event_time = 0	-0.0184	-0.1128***	-0.0799***
	(0.0225)	(0.0237)	(0.0242)
Job mover × event_time = 1	0.1672***	-0.2069***	-0.0380
	(0.0357)	(0.0363)	(0.0360)
Job mover × event_time = 2	0.1323***	-0.2101***	-0.0355
	(0.0391)	(0.0397)	(0.0394)
Job mover × event_time = 3	-0.0117	-0.3078***	-0.1291***
	(0.0379)	(0.0383)	(0.0380)
Job mover × event_time = 4	-0.0196	-0.2641***	-0.0780**
	(0.0338)	(0.0342)	(0.0340)
Job mover × event_time = 5	-0.0234	-0.2527***	-0.0621*
	(0.0364)	(0.0371)	(0.0367)
Job mover × event_time = 6	0.0134	-0.2151***	-0.0197
	(0.0386)	(0.0386)	(0.0381)
Job mover × event_time = 7	-0.3461***	-0.2582***	-0.0785***
	(0.0311)	(0.0309)	(0.0303)
Job mover × event_time = 8	-0.3202***	-0.2582***	-0.0671**
	(0.0302)	(0.0298)	(0.0295)
Job mover × event_time = 9	-0.2907***	-0.2614***	-0.0634**
	(0.0320)	(0.0316)	(0.0313)
Job mover × event_time = 10	-0.3573***	-0.2762***	-0.0725**
	(0.0312)	(0.0310)	(0.0307)
User FE	×	×	×
Month FE		×	×
Experience FE			×
Adjusted R^2	0.28992	0.30870	0.35963
Observations	1,946,413	1,946,413	1,946,413
Users	22,896	22,896	22,896

Notes: Estimates for $t_j \times \text{JobChanger}_i$ based on Equation 1.2 with user and calendar month fixed effects. The outcome is IHS-transformed commits to single-authored projects. The reference month is $t = -16$. Bars show 95% confidence intervals. Standard errors are clustered at the user level. * p > 0.01, ** p > 0.05, and *** p > 0.1. *Sources:* GHTorrent, own calculations.

Table D.14: Job search period

Job search period:	(1) [−15, −9]	(2) [−15, −6]	(3) [−15, −3]	(4) [−15, 0]
Job mover × job search	0.1947***	0.1836***	0.1768***	0.1646***
	(0.0154)	(0.0144)	(0.0141)	(0.0141)
Job mover × uncertain	0.1423***	0.1308***	0.0925***	
	(0.0161)	(0.0178)	(0.0203)	
Job mover × post move	-0.1099***	-0.1099***	-0.1100***	-0.1036***
	(0.0184)	(0.0184)	(0.0184)	(0.0190)
User FE	×	×	×	×
Month FE	×	×	×	×
Experience FE	×	×	×	×
Adjusted R^2	0.35946	0.35946	0.35946	0.35945
Observations	1,946,413	1,946,413	1,946,413	1,946,413
Users	22,896	22,896	22,896	22,896
Relation to baseline	+3.01 p.p.	+1.90 p.p.	+1.22 p.p.	baseline
	+18.3 %	+11.5 %	+7.4 %	baseline

Notes: Results from estimation of Equation 4.2 for different definitions of the job serach period. Experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$. *Sources:* GHTorrent, own calculations.

Table D.15: International movers

IHS(single commits)	international		inter-continental	
	(1) yes	(2) no	(3) yes	(4) no
Job mover × job search	0.2027***	0.1474***	0.2335***	0.1483***
	(0.0263)	(0.0167)	(0.0336)	(0.0155)
Job mover × post move	-0.0812**	-0.1124***	-0.1057**	-0.1031***
	(0.0342)	(0.0228)	(0.0435)	(0.0211)
User FE	×	×	×	×
Month FE	×	×	×	×
Experience FE	×	×	×	×
Adjusted R^2	0.36811	0.35640	0.36273	0.35907
Observations	562,982	1,383,431	366,271	1,580,142
Users	6,598	16,298	4,305	18,591

Notes: Results from estimation of Equation 4.2 with IHS-transformed number of commits to (non-)international and (non-)inter-continental single-authored projects. Upward income group moves are defined as moves from developing to developed countries. Upward moves in GDP per capita are based on current 2021 PPP USD. Experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$.
Sources: GHTorrent, own calculations.

Table D.16: Upward movers

IHS(single commits)	GDP p. c.		income class	
	(1) other	(2) up	(3) other	(4) up
Job mover × job search	0.1622***	0.1821***	0.1610***	0.2381***
	(0.0149)	(0.0437)	(0.0146)	(0.0512)
Job mover × post move	-0.1034***	-0.1038*	-0.1038***	-0.0949
	(0.0199)	(0.0627)	(0.0195)	(0.0755)
User FE	×	×	×	×
Month FE	×	×	×	×
Experience FE	×	×	×	×
Adjusted R^2	0.36073	0.34025	0.35980	0.33293
Observations	1,776,167	170,246	1,854,956	91,457
Users	20,829	2,067	21,763	1,133

Notes: Results from estimation of Equation 4.2 with IHS-transformed number of commits to (non-)upward single-authored projects in terms of GDP p.c. and income class, respectively. Upward income group moves are defined as moves from developing to developed countries. Upward moves in GDP per capita are based on current 2021 PPP USD. Experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$. *Sources:* GHTorrent, own calculations.

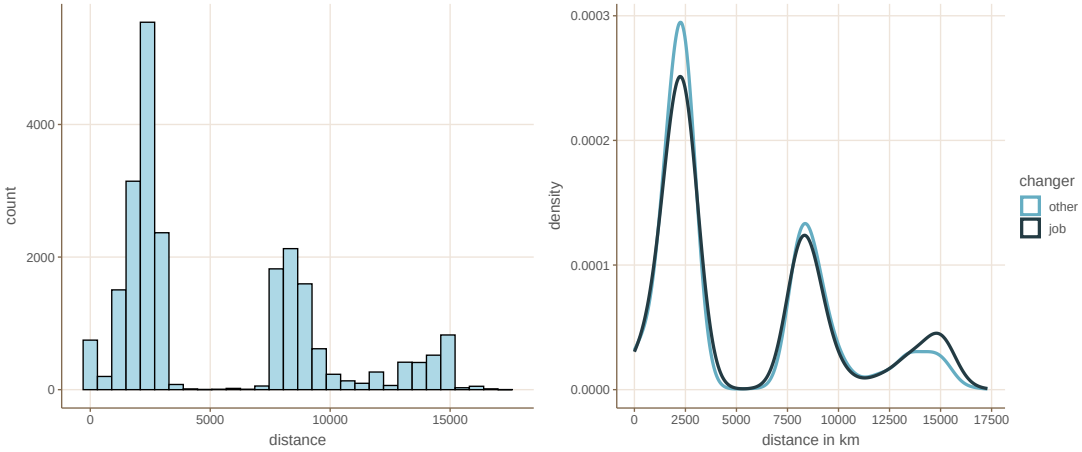
Table D.17: Affiliation

IHS(single commits)	destination						origin			
	median		big tech		academia		median		adademia	
	(1) below	(2) above	(3) no	(4) yes	(5) no	(6) yes	(7) below	(8) above	(9) no	(10) yes
Job mover × job search	0.1770***	0.0068	0.1740***	0.1556***	0.1535***	0.3158***	0.1636***	0.2339**	0.1565***	0.1511***
	(0.0145)	(0.0526)	(0.0212)	(0.0176)	(0.0146)	(0.0459)	(0.0142)	(0.1052)	(0.0519)	(0.0472)
Job mover × post move	-0.0955***	-0.1556**	-0.1001***	-0.0895***	-0.1199***	0.1547**	-0.1038***	-0.0610	-0.1755**	-0.1593**
	(0.0195)	(0.0668)	(0.0296)	(0.0232)	(0.0194)	(0.0717)	(0.0191)	(0.1320)	(0.0758)	(0.0721)
User FE	×	×	×	×	×	×	×	×	×	×
Month FE	×	×	×	×	×	×	×	×	×	×
Experience FE	×	×	×	×	×	×	×	×	×	×
Adjusted R^2	0.35927	0.36002	0.36084	0.35832	0.35823	0.36154	0.35933	0.35989	0.35999	0.36103
Observations	1,900,195	1,369,596	1,553,857	1,715,934	1,900,917	1,368,874	1,935,568	1,334,223	1,361,217	1,368,330
Users	22,387	16,194	18,374	20,207	22,378	16,203	22,767	15,814	16,130	16,212

Notes: Results from estimation of Equation 4.2 with IHS-transformed number of commits to single-authored projects. Median split refers to median size of affiliation in terms of users in the full *GHTorrent* sample. Big tech refers to Google, Amazon, Meta, Apple and Microsoft. Academia refers to students and university affiliations. Destination (origin) refers to users' affiliation before (after) the affiliation change. Experience is measured as months since the first commit at move month. Robust standard errors are clustered at the user level. * $p > 0.01$, ** $p > 0.05$, and *** $p > 0.1$. *Sources:* GHTorrent, own calculations.

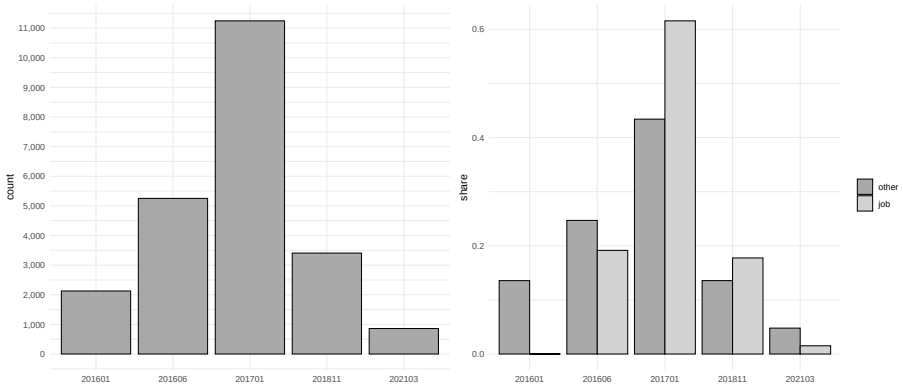
D.2 Figures

Figure D.1: Distribution of move distances



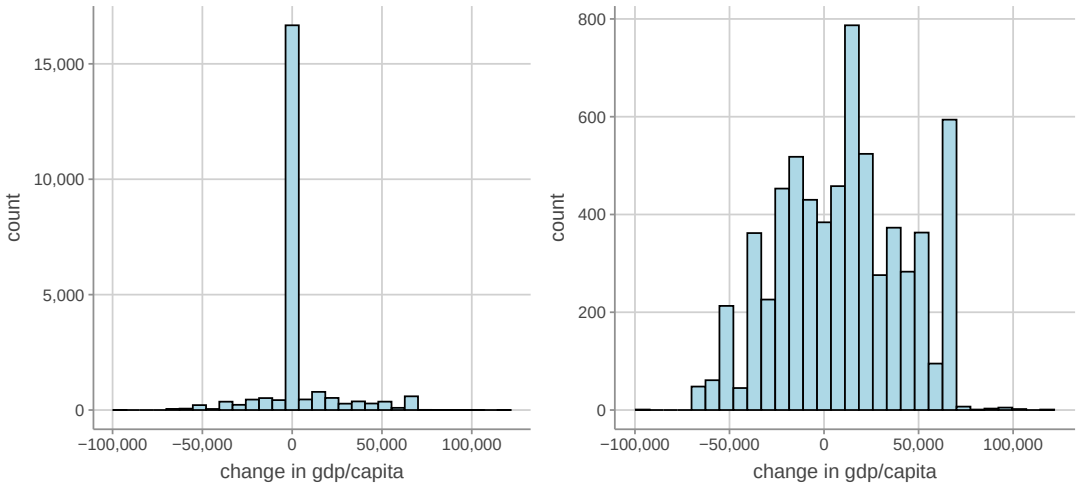
Notes: Histogram on the left shows the distribution of move distances. Estimates on the right show kernel densities for job movers and other movers. *Sources:* GHTorrent, own calculations.

Figure D.2: Distribution of moves across time



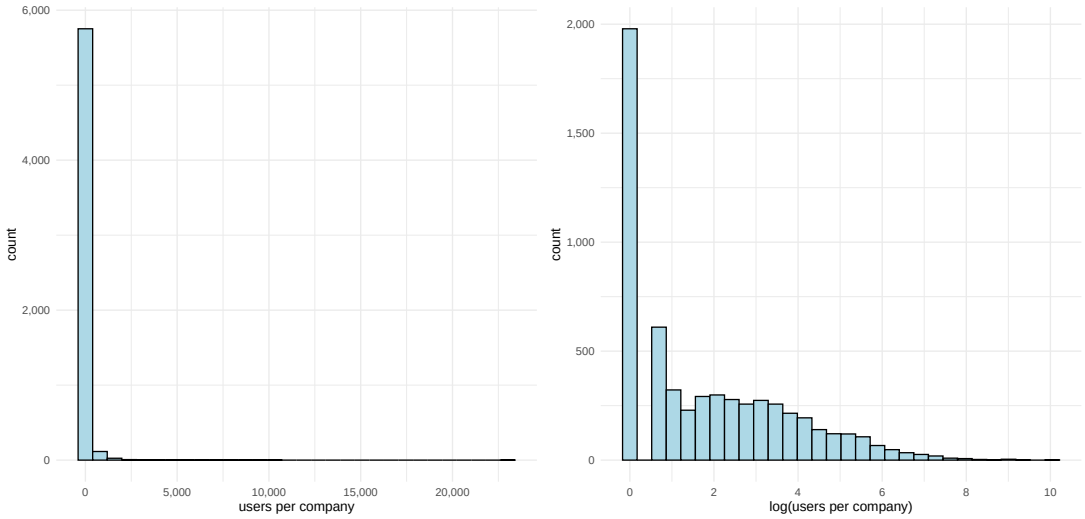
Notes: Histogram on the left shows the distribution of moves across data snapshots. Shares on the right depict the distribution of moves across data snapshots for job movers (dark gray) and other movers (light gray). *Sources:* GHTorrent, own calculations.

Figure D.3: Distribution of income changes



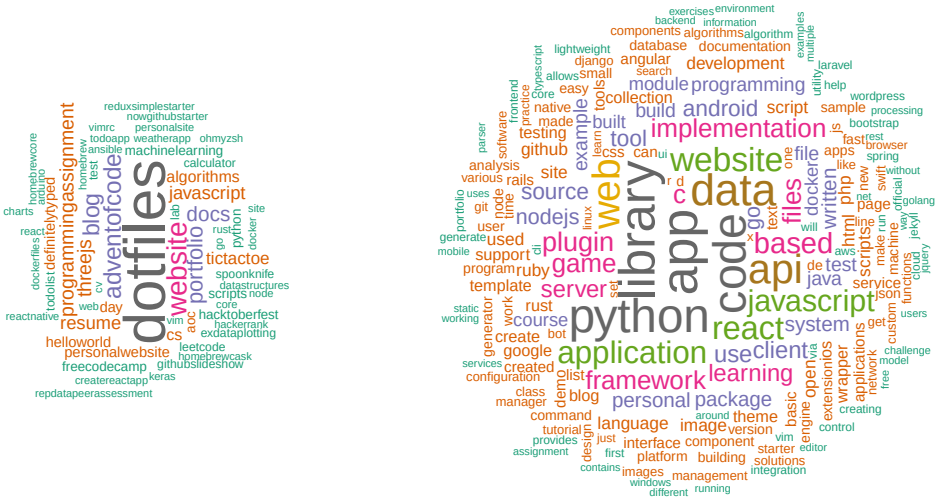
Notes: Histograms depict the distribution of national per capita GDP changes of movers in the full sample (left) and the international sample (right). GDP is measured in current 2021 PPP USD. *Sources:* GHTorrent, World Development Indicators, own calculations.

Figure D.4: Distribution of affiliation size



Notes: Histograms depict the distribution of affiliations with respect to the number of affiliated users in the full *GHTorrent* sample as counts (left) and after logarithmic transformation (right). Note that string-based merging of affiliations is likely imperfect, especially for small firms, which leads to a downward bias of firm size. *Sources:* GHTorrent, own calculations.

Figure D.5: Frequent words in project names and descriptions



Notes: Word clouds show frequently occurring words in single projects of movers. Word size and color represent word frequency in project titles (left) and descriptions (right). Frequency limits are set at 50 (titles) and 100 (descriptions). We remove English stop words, numbers, punctuation, URLs, white space, and the words *project*, *repository/repo*, *simple*, and *using*. *Sources:* GHTorrent, own calculations.

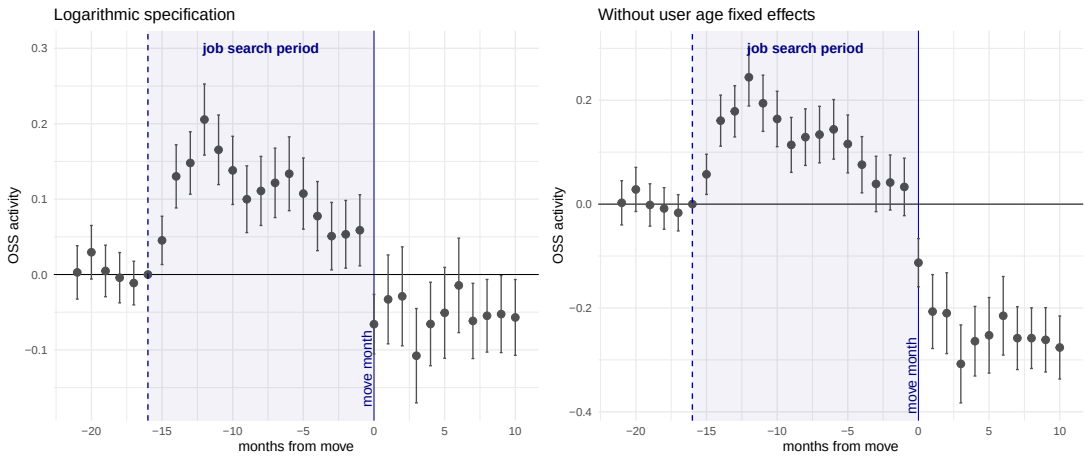
Figure D.6: Heterogeneity by user popularity

Notes: Estimates for $t_j \times \text{JobChanger}_i$ based on Equation 1.2 with user, experience and calendar month fixed effects. The outcome is IHS-transformed commits to single-authored projects in the respective follower quartile (1st quartile: green; 2nd quartile: orange; 3rd quartile: blue and 4th quartile: purple.). The reference month is $t = -16$. Bars show 95% confidence intervals. Standard errors are clustered at the user level. *Sources:* GHTorrent, own calculations.

Figure D.7: Heterogeneity by project age

Notes: Estimates for $t_j \times \text{JobChanger}_i$ based on Equation 1.2 with user and calendar month fixed effects. The outcome is IHS-transformed commits to single-authored new (orange) and old (green) projects. New projects are defined as projects with the date of the first commit in the month under consideration. The reference month is $t = -16$. Bars show 95% confidence intervals. Standard errors are clustered at the user level. *Sources:* GHTorrent, own calculations.

Figure D.8: Event study model robustness



Notes: Estimates for $t_j \times \text{JobChanger}_i$ based on Equation 1.2 with user and calendar month fixed effects. The outcome is logarithmically transformed using $ln(y+1)$ in the left panel and IHS-transformed commits to single-authored projects in the right panel. The reference month is $t = -16$. Bars show 95% confidence intervals. Standard errors are clustered at the user level. *Sources:* GHTorrent, own calculations.

Bibliography

Abou El-Komboz, L. and Fackler, T. (2022). Productivity spillovers among knowledge workers in agglomerations: Evidence from github. *Working Paper*.

Abrahams, A., Oram, C., and Lozano-Gracia, N. (2018). Deblurring DMSP nighttime lights: A new method using gaussian filters and frequencies of illumination. *Remote Sensing of Environment*, 210:242–258.

Abreu, M., Faggian, A., and McCann, P. (2015). Migration and inter-industry mobility of uk graduates. *Journal of Economic Geography*, 15(2):353–385.

Acemoglu, D. and Autor, D. (2011). Skills, tasks and technologies: Implications for employment and earnings. In *Handbook of Labor Economics*, volume 4, pages 1043–1171. Elsevier.

Acemoglu, D., Johnson, S., and Robinson, J. A. (2005). Institutions as a fundamental cause of long-run growth. *Handbook of Economic Growth*, 1:385–472.

Acikalin, U. U., Caskurlu, T., Hoberg, G., and Phillips, G. M. (2022). Intellectual property protection lost and competition: An examination using machine learning. *Working Paper*.

Adams, J. D., Black, G. C., Clemmons, J. R., and Stephan, P. E. (2005). Scientific teams and institutional collaborations: Evidence from us universities, 1981–1999. *Research Policy*, 34(3):259–285.

Adrian, F., Miriam, F., and Michael, P. S. (2017). *The Human Face of Global Mobility: A Research Agenda*. Routledge.

Aggarwal, S. (2018). Do rural roads create pathways out of poverty? evidence from india. *Journal of Development Economics*, 133:375–395.

Aghion, P. and Howitt, P. W. (2008). *The Economics of Growth*. MIT Press.

Agrawal, A., Catalini, C., and Goldfarb, A. (2015). Crowdfunding: Geography, social networks, and the timing of investment decisions. *Journal of Economics & Management Strategy*, 24(2):253–274.

Agrawal, A., Cockburn, I., and McHale, J. (2006). Gone but not forgotten: Knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography*, 6(5):571–591.

Agrawal, A. and Goldfarb, A. (2008). Restructuring research: Communication costs and the democratization of university innovation. *American Economic Review*, 98(4):1578–1590.

Agrawal, A., Lacetera, N., and Lyons, E. (2016). Does standardized information in online markets disproportionately benefit job applicants from less developed countries? *Journal of International Economics*, 103:1–12.

Agyeman, O. T. (2007). Survey of ict and education in africa: Benin country report. *World Bank Report*.

Ahoyo, C. P. (2006). Rapport d'analyse de l'enquête sur l'utilisation des tic au bénin. In *Pour Une Vraie Économie Numérique Au Bénin*. International Telecommunication Union (ITU).

Aihounton, G. and Henningsen, A. (2021). Units of measurement and the inverse hyperbolic sine transformation. *The Econometrics Journal*, 24(2):334–351.

Akcigit, U., Caicedo, S., Miguelez, E., Stantcheva, S., and Sterzi, V. (2018). Dancing with the stars: Innovation through interactions. *NBER Working Paper*.

Aker, J. C. (2010). Information from Markets Near and Far: Mobile Phones and Agricultural Markets in Niger. *American Economic Journal: Applied Economics*, 2(3):46–59.

Aker, J. C. and Mbiti, I. M. (2010). Mobile phones and economic development in africa. *Journal of Economic Perspectives*, 24(3):207–232.

Akerman, A., Gaarder, I., and Mogstad, M. (2015). The skill complementarity of broadband internet. *Quarterly Journal of Economics*, 130(4):1781–1824.

Akerman, A., Leuven, E., and Mogstad, M. (2022). Information frictions, internet, and the relationship between distance and trade. *American Economic Journal: Applied Economics*, 14(1):133–163.

Al-Ani, B. and Edwards, H. K. (2008). A comparative empirical study of communication in distributed and collocated development teams. *IEEE International Conference on Global Software Engineering (ICSSP)*, pages 35–44.

Albouy, D. (2016). What are cities worth? land rents, local productivity, and the total value of amenities. *The Review of Economics and Statistics*, 98(3):477–487.

Alesina, A. and Dollar, D. (2000). Who gives foreign aid to whom and why? *Journal of Economic Growth*, 5:33–63.

Alesina, A. and Giuliano, P. (2015). Culture and institutions. *Journal of Economic Literature*, 53(4):898–944.

Alipour, J.-V., Falck, O., and Schüller, S. (2023). Germany's capacity to work from home. *European Economic Review*, 151:104354.

Amior, M. (2015). Why are higher skilled workers more mobile geographically?: The role of the job surplus. *CEPR Working Paper*.

Amior, M. (2019). Education and geographical mobility: The role of the job surplus. *CEPR Working Paper*.

Anderson, B. and O'Connor, A. C. (2020). Economic impact of 2africa. *RTI International Report*.

Anderson, J. and van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review*, 93(1):170–192.

Anderson, J. E. (1979). A theoretical foundation for the gravity equation. *American Economic Review*, 69(1):106–116.

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

Andreessen, M. (2011). Why software is eating the world. *Wall Street Journal*, 20(2011):C2.

Angrist, J., Azoulay, P., Ellison, G., Hill, R., and Lu, S. F. (2017). Economic research evolves: Fields and styles. *American Economic Review*, 107(5):293–297.

Antràs, P., Garicano, L., and Rossi-Hansberg, E. (2006). Offshoring in a knowledge economy. *Quarterly Journal of Economics*, 121(1):31–77.

Arkolakis, C., Huneeus, F., and Miyauchi, Y. (2023). Spatial production networks. *NBER Working Paper*.

Arrow, K. J. (1962). The economic implications of learning by doing. *Review of Economic Studies*, 29(3):155–173.

Arrow, K. J. (1974). *The Limits of Organization*. WW Norton & Company.

Asher, S., Lunt, T., Matsuura, R., and Novosad, P. (2021). Development research at high geographic resolution: An analysis of night-lights, firms, and poverty in india using the SHRUG open data platform. *The World Bank Economic Review*, 35(4):845–871.

Asher, S. and Novosad, P. (2020). Rural roads and local economic development. *American Economic Review*, 110(3):797–823.

Athey, S. and Ellison, G. (2014). Dynamics of open source movements. *Journal of Economics & Management Strategy*, 23(2):294–316.

Atkin, D., Chen, M. K., and Popov, A. (2022). The returns to face-to-face interactions: Knowledge spillovers in silicon valley. *NBER Working Paper*.

Autor, D. H., Dorn, D., and Hanson, G. H. (2015). Untangling trade and technology: Evidence from local labour markets. *Economic Journal*, 125(584):621–646.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2006). The polarization of the us labor market. *American Economic Review*, 96(2):189–194.

Azoulay, P., Graff Zivin, J. S., and Wang, J. (2010). Superstar extinction. *Quarterly Journal of Economics*, 125(2):549–589.

Badashian, A. S., Esteki, A., Gholipour, A., Hindle, A., and Stroulia, E. (2014). Involvement, contribution and influence in github and stack overflow. *International Conference on Computer Science and Software Engineering (CASCON)*, 14:19–33.

Bagozzi, R. P. and Dholakia, U. M. (2006). Open source software user communities: A study of participation in linux user groups. *Management Science*, 52(7):1099–1115.

Bahar, D., Choudhury, P., Kim, D. Y., and Koo, W. W. (2022). Innovation on wings: Nonstop flights and firm innovation in the global context. *Management Science*.

Bahia, K., Castells, P., Cruz, G., Masaki, T., Rodriguez-Castelan, C., and Sanfelice, V. (2021). Mobile broadband internet, poverty and labor outcomes in tanzania. *IZA Discussion Paper*.

Bahia, K., Pedrós, X., Rodríguez-castelán, C., Winkler, H., Pfutze, T., and Rodríguez-castelán, C. (2020). The Welfare Effects of Mobile Broadband Internet: Evidence from Nigeria. *World Bank Policy Research Working Paper*.

Baier, S. L. and Bergstrand, J. H. (2007). Do free trade agreements actually increase members' international trade? *Journal of International Economics*, 71(1):72–95.

Bailey, M., Cao, R., Kuchler, T., and Stroebel, J. (2018a). The economic effects of social networks: Evidence from the housing market. *Journal of Political Economy*, 126(6):2224–2276.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., and Wong, A. (2018b). Social connectedness: Measurement, determinants, and effects. *Journal of Economic Perspectives*, 32(3):259–80.

Bailey, M., Farrell, P., Kuchler, T., and Stroebel, J. (2020a). Social connectedness in urban areas. *Journal of Urban Economics*, 118:103264.

Bailey, M., Gupta, A., Hillenbrand, S., Kuchler, T., Richmond, R., and Stroebel, J. (2021). International trade and social connectedness. *Journal of International Economics*, 129:103418.

Bailey, M., Johnston, D., Kuchler, T., Russel, D., Stroebel, J., et al. (2020b). The determinants of social connectedness in europe. *International Conference on Social Informatics (SocInfo)*, pages 1–14.

Baily, M. N. (2002). The new economy: Post mortem or second wind? *Journal of Economic Perspectives*, 16(2):3–22.

Baily, M. N. and Lawrence, R. Z. (2001). Do we have a new e-conomy? *American Economic Review*, 91(2):308–312.

Baldwin, R. (2017). *The Great Convergence: Information Technology and the New Globalization*. Harvard University Press.

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

Baldwin, R. and Dingel, J. I. (2022). Telemigration and development: On the offshorability of teleworkable jobs. In *Robots and AI*, pages 150–179. Routledge.

Baldwin, R. and Forslid, R. (2023). Globotics and development: When manufacturing is jobless and services are tradeable. *World Trade Review*, 22(3-4):302–311.

Balgova, M. (2020). Leaping into the unknown? the role of job search in migration decisions. *Working Paper*.

Balland, P.-A., Jara-Figueroa, C., Petralia, S. G., Steijn, M. P. A., Rigby, D. L., and Hidalgo, C. A. (2020). Complex economic activities concentrate in large cities. *Nature Human Behaviour*, 4(3):248–254.

Banerjee, A., Duflo, E., and Qian, N. (2020). On the Road: Access to Transportation Infrastructure and Economic Growth in China. *Journal of Development Economics*, 145(January):102442.

Baragwanath, K., Goldblatt, R., Hanson, G., and Khandelwal, A. K. (2021). Detecting urban markets with satellite imagery: An application to india. *Journal of Urban Economics*, 125:103173.

Barbosa, A. V., Casagrande, D., Maier, P., and Trevisan, G. (2021). Changing the pyramids: The impact of broadband internet on firm employment. *Working Paper*.

Bartelme, D. and Ziv, O. (2024). The internal geography of firms. *Journal of International Economics*, 148:103889.

Battiston, D., Blanes i Vidal, J., and Kirchmaier, T. (2021). Face-to-face communication in organizations. *The Review of Economic Studies*, 88(2):574–609.

Baum-Snow, N., Gendron-Carrier, N., and Pavan, R. (2020). Local productivity spillovers. *Working Paper*.

Baumol, W. J. (1967). Macroeconomics of unbalanced growth: The anatomy of urban crisis. *American Economic Review*, 57(3):415–426.

Bellégo, C., Benatia, D., and Pape, L. (2022). Dealing with logs and zeros in regression models. *Working Paper*.

Bellemare, M. F. and Wichman, C. J. (2020). Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics*, 82(1):50–61.

BenYishay, A., Rotberg, R., Wells, J., Lv, Z., Goodman, S., Kovacevic, L., and Runfola, D. (2017). Geocoding afrobarometer rounds 1-6: Methodology & data quality. *AidData*.

Bercovitz, J. and Feldman, M. (2011). The mechanisms of collaboration in inventive teams: Composition, social networks, and geography. *Research Policy*, 40(1):81–93.

Berge, L., Krantz, S., and McDermott, G. (2023). fixest: Fast and user-friendly fixed-effects estimation. *CRAN*.

Bergstrand, J. H. (1985). The gravity equation in international trade: Some microeconomic foundations and empirical evidence. *The Review of Economics and Statistics*, pages 474–481.

Bergstrom, T., Blume, L., and Varian, H. (1986). On the private provision of public goods. *Journal of Public Economics*, 29(1):25–49.

Bertschek, I. and Niebel, T. (2016). Mobile and more productive? firm-level evidence on the productivity effects of mobile internet use. *Telecommunications Policy*, 40(9):888–898.

Best, R. and Burke, P. J. (2018). Electricity availability: A precondition for faster economic growth? *Energy Economics*, 74:321–329.

Bilodeau, M. and Slivinski, A. (1996). Toilet cleaning and department chairing: Volunteering a public service. *Journal of Public Economics*, 59(2):299–308.

Bird, C., Nagappan, N., Devanbu, P., Gall, H., and Murphy, B. (2009). Does distributed development affect software quality? an empirical case study of windows vista. *IEEE International Conference on Software Engineering (ICSE)*, pages 518–528.

Bitzer, J. and Geishecker, I. (2010). Who contributes voluntarily to oss? an investigation among german it employees. *Research Policy*, 39(1):165–172.

Bitzer, J., Schrettl, W., and Schröder, P. J. H. (2007). Intrinsic motivation in open source software development. *Journal of Comparative Economics*, 35(1):160–169.

Bitzer, J. and Schröder, P. J. H. (2005). Bug-fixing and code-writing: The private provision of open source software. *Information Economics and Policy*, 17(3):389–406.

Bitzer, J. and Schröder, P. J. H. (2007). Open source software, competition and innovation. *Industry and Innovation*, 14(5):461–476.

Blau, G. (1994). Testing a two-dimensional measure of job search behavior. *Organizational Behavior and Human Decision Processes*, 59(2):288–312.

Blincoe, K., Sheoran, J., Goggins, S., Petakovic, E., and Damian, D. (2016). Understanding the popular users: Following, affiliation influence and leadership on github. *Information and Software Technology*, 70:30–39.

Bliss, C. and Nalebuff, B. (1984). Dragon-slaying and ballroom dancing: The private supply of a public good. *Journal of Public Economics*, 25(1-2):1–12.

Bloom, N., Han, R., and Liang, J. (2022). How hybrid working from home works out. *NBER Working Paper*.

Bloom, N., Jones, C. I., van Reenen, J., and Webb, M. (2020). Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144.

Bloom, N., Liang, J., Roberts, J., and Ying, Z. J. (2015). Does working from home work? evidence from a chinese experiment. *Quarterly Journal of Economics*, 130(1):165–218.

Bluhm, R. and Krause, M. (2022). Top lights: Bright cities and their contribution to economic development. *Journal of Development Economics*, 157:102880.

Bluhm, R. and McCord, G. C. (2022). What can we learn from nighttime lights for small geographies? measurement errors and heterogeneous elasticities. *Remote Sensing*, 14(5):1–25.

Blum, B. and Goldfarb, A. (2006). Does the internet defy the law of gravity? *Journal of International Economics*, 70(2):384–405.

Boisso, D. and Ferrantino, M. (1997). Economic distance, cultural distance, and openness in international trade: Empirical puzzles. *Journal of Economic Integration*, pages 456–484.

Boopen, S. (2006). Transport infrastructure and economic growth: Evidence from africa using dynamic panel estimates. *Empirical Economics Letters*, 5(1):37–52.

Bove, V. and Gokmen, G. (2018). Genetic distance, trade, and the diffusion of development. *Journal of Applied Econometrics*, 33(4):617–623.

Breschi, S. and Lissoni, F. (2009). Mobility of skilled workers and co-invention networks: An anatomy of localized knowledge flows. *Journal of Economic Geography*, 9(4):439–468.

Bresnahan, T. F., Brynjolfsson, E., and Hitt, L. M. (2002). Information technology, workplace organization, and the demand for skilled labor: Firm-level evidence. *Quarterly Journal of Economics*, 117(1):339–376.

Briglauer, W., Dürr, N. S., Falck, O., and Hüschelrath, K. (2019). Does state aid for broadband deployment in rural areas close the digital and economic divide? *Information Economics and Policy*, 46:68–85.

Brucks, M. S. and Levav, J. (2022). Virtual communication curbs creative idea generation. *Nature*, 605(7908):108–112.

Brülhart, M., Desmet, K., and Klinke, G.-P. (2020). The shrinking advantage of market potential. *Journal of Development Economics*, 147:102529.

Brun, J.-F., Carrère, C., Guillaumont, P., and De Melo, J. (2005). Has distance died? evidence from a panel gravity model. *The World Bank Economic Review*, 19(1):99–120.

Burbidge, J. B., Magee, L., and Robb, A. L. (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association*, 83(401):123–127.

Burlig, F. and Preonas, L. (2016). Out of the darkness and into the light? development effects of rural electrification. *Energy Institute at Haas Working Paper*, 268:26.

Butler, S., Gamalielsson, J., Lundell, B., Brax, C., Sjöberg, J., Mattsson, A., Gustavsson, T., Feist, J., and Lönroth, E. (2019). On company contributions to community open source software projects. *IEEE Transactions on Software Engineering*, 47(7):1381–1401.

Buys, P., Dasgupta, S., Thomas, T. S., and Wheeler, D. (2009). Determinants of a digital divide in sub-saharan africa: A spatial econometric analysis of cell phone coverage. *World Development*, 37(9):1494–1505.

Byanyuma, M., Kalolo, S., Mrutu, S. I., Nyakyi, C., and Sam, A. (2013). Affordable broadband connectivity for rural areas. *IEEE Pan-African International Conference on Information Science, Computing and Telecommunications (PACT)*, pages 62–65.

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

Cao, Y., Sickles, R. C., Triebs, T. P., and Tumlinson, J. (2024). Linguistic distance to english impedes research performance. *Research Policy*, 53(4):104971.

Carlino, G. and Kerr, W. R. (2015). *Agglomeration and Innovation*, volume 5. Elsevier B.V., 1 edition.

Catalini, C. (2018). Microgeography and the direction of inventive activity. *Management Science*, 64(9):4348–4364.

Catalini, C., Fons-Rosen, C., and Gaulé, P. (2020). How do travel costs shape collaboration? *Management Science*, 66(8):3340–3360.

Chabossou, A. F. (2007). *2007 Benin Telecommunications Sector Performance Review: A Supply-side Analysis of Policy Outcomes*. Research ICT Africa.

Chakravorty, U., Pelli, M., and Marchand, B. U. (2014). Does the quality of electricity matter? evidence from rural india. *Journal of Economic Behavior & Organization*, 107:228–247.

Chamberlain, A. (2015). Why is hiring taking longer. *Glassdoor Policy Brief*.

Chamberlin, J. (1974). Provision of collective goods as a function of group size. *American Political Science Review*, 68(2):707–716.

Chan, K., Covrig, V., and Ng, L. (2005). What determines the domestic bias and foreign bias? evidence from mutual fund equity allocations worldwide. *The Journal of Finance*, 60(3):1495–1534.

Chattergoon, B. and Kerr, W. R. (2022). Winner takes all? tech clusters, population centers, and the spatial transformation of us invention. *Research Policy*, 51(2):104418.

Chaurey, R. and Le, D. T. (2022). Infrastructure maintenance and rural economic activity: Evidence from india. *Journal of Public Economics*, 214:104725.

Chavula, J., Feamster, N., Bagula, A., and Suleman, H. (2015). Quantifying the effects of circuitous routes on the latency of intra-africa internet traffic: A study of research and education networks. In *E-Infrastructure and E-Services for Developing Countries*, pages 64–73. Springer International Publishing.

Chen, C., Frey, C. B., and Presidente, G. (2022). Disrupting science. *Working Paper*.

Chen, J. and Roth, J. (2023). Logs with zeros? some problems and solutions. *Working Paper*.

Chen, N. (2004). Intra-national versus international trade in the european union: Why do national borders matter? *Journal of International Economics*, 63(1):93–118.

Chen, X. and Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 108(21):8589–8594.

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R. B., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., et al. (2022a). Social capital I: Measurement and associations with economic mobility. *Nature*, 608(7921):108–121.

Chetty, R., Jackson, M. O., Kuchler, T., Stroebel, J., Hendren, N., Fluegge, R. B., Gong, S., Gonzalez, F., Grondin, A., Jacob, M., et al. (2022b). Social capital II: Determinants of economic connectedness. *Nature*, 608(7921):122–134.

Chetty, R., Looney, A., and Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, 99(4):1145–1177.

Chevalier, J. and Ellison, G. (1999). Career concerns of mutual fund managers. *Quarterly Journal of Economics*, 114(2):389–432.

Choudhury, P. (2020). Our work-from-anywhere future. *Harvard Business Review*, 28.

Choudhury, P., Foroughi, C., and Larson, B. (2021). Work-from-anywhere: The productivity effects of geographic flexibility. *Strategic Management Journal*, 42(4):655–683.

Christiaensen, L. and Kanbur, R. (2017). Secondary Towns and Poverty Reduction: Refocusing the Urbanization Agenda. *Annual Review of Resource Economics*, 9:405–419.

Christiaensen, L. and Todo, Y. (2014). Poverty Reduction During the Rural-urban Transformation: The Role of the Missing Middle. *World Development*, 63:43–58.

Ciriaci, D. (2014). Does university quality influence the interregional mobility of students and graduates? the case of italy. *Regional Studies*, 48(10):1592–1608.

Clark, T. N., Lloyd, R., Wong, K. K., and Jain, P. (2002). Amenities Drive Urban Growth. *Journal of Urban Affairs*, 24(5):493–515.

Cohn, J. B., Liu, Z., and Wardlaw, M. I. (2022). Count (and count-like) data in finance. *Journal of Financial Economics*, 146(2):529–551.

Colombo, M. G., Piva, E., and Rossi-Lamastra, C. (2014). Open innovation and within-industry diversification in small and medium enterprises: The case of open source software firms. *Research Policy*, 43(5):891–902.

Conte, M., Cotterlaz, P., and Mayer, T. (2022). *The CEPII Gravity Database*. CEPII.

Conti, A., Gupta, V., Guzman, J., and Roche, M. P. (2023). Incentivizing innovation in open source: Evidence from the github sponsor program. *Working Paper*.

Conti, A., Peukert, C., and Roche, M. P. (2021). Beefing it up for your investor? open sourcing and startup funding: Evidence from github. *Working Paper*.

Correia, S. (2019). reghdfe: Stata module to perform linear or instrumental-variable regression absorbing any number of high-dimensional fixed effects. *Statistical Software Components*.

Correia, S., Guimarães, P., and Zylkin, T. (2019). ppmlhdfe: Fast poisson estimation with high-dimensional fixed effects. *Working Paper*.

Coscia, M., Neffke, F. M. H., and Hausmann, R. (2020). Knowledge diffusion in the network of international business travel. *Nature Human Behaviour*, 4(10):1011–1020.

Cowan, R., Jonard, N., and Zimmermann, J.-B. (2007). Bilateral collaboration and the emergence of innovation networks. *Management Science*, 53(7):1051–1067.

Cramton, C. D. (2001). The mutual knowledge problem and its consequences for dispersed collaboration. *Organization Science*, 12(3):346–371.

Creanza, N., Ruhlen, M., Pemberton, T. J., Rosenberg, N. A., Feldman, M. W., and Ramachandran, S. (2015). A somparison of worldwide phonemic and genetic variation in human populations. *Proceedings of the National Academy of Sciences*, 112(5):1265–1272.

Cristea, A. D. (2011). Buyer-seller relationships in international trade: Evidence from us states' exports and business-class travel. *Journal of International Economics*, 84(2):207–220.

Cummings, J. N. and Kiesler, S. (2007). Coordination costs and project outcomes in multi-university collaborations. *Research Policy*, 36(10):1620–1634.

Curiel, R. P., Heinrigs, P., and Heo, I. (2017). Cities and spatial interactions in west africa. *OECD West African Papers*.

Czernich, N., Falck, O., Kretschmer, T., and Woessmann, L. (2011). Broadband Infrastructure and Economic Growth. *Economic Journal*, 121(552):505–532.

Daniel, S., Agarwal, R., and Stewart, K. J. (2013). The effects of diversity in global, distributed collectives: A study of open source project success. *Information Systems Research*, 24(2):312–333.

Dauth, W., Findeisen, S., Moretti, E., and Suedekum, J. (2022). Matching in cities. *Journal of the European Economic Association*, 20(4):1478–1521.

De La Roca, J. and Puga, D. (2017). Learning by working in big cities. *The Review of Economic Studies*, 84(1):106–142.

Deardorff, A. (1998). Determinants of bilateral trade: Does gravity work in a neoclassical world? In *The Regionalization of the World Economy*, pages 7–32. University of Chicago Press.

del Rio-Chanona, M., Laurentsyeva, N., and Wachs, J. (2023). Are large language models a threat to digital public goods? evidence from activity on stack overflow. *arXiv Preprint*.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Management Unit Working Paper*.

Deller, S. C., Lledo, V., and Marcouiller, D. W. (2008). Modeling Regional Economic Growth with a Focus on Amenities. *Review of Urban and Regional Development Studies*, 20(1):1–21.

Deming, D. J. and Noray, K. (2020). Earnings dynamics, changing job skills, and stem careers. *Quarterly Journal of Economics*, 135(4):1965–2005.

Dey, M., Frazis, H., Loewenstein, M. A., and Sun, H. (2020). 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*.

Diemer, A. and Regan, T. (2022). No inventor is an island: Social connectedness and the geography of knowledge flows in the us. *Research Policy*, 51(2):104416.

Dijkstra, L., Hamilton, E., Lall, S., and Wahba, S. (2020). How do we define cities, towns, and rural areas? *World Bank Blog: Sustainable Cities*.

Ding, W. W., Levin, S. G., Stephan, P. E., and Winkler, A. E. (2010). The impact of information technology on academic scientists' productivity and collaboration patterns. *Management Science*, 56(9):1439–1461.

Dingel, J. I. and Neiman, B. (2020). How many jobs can be done at home? *Journal of Public Economics*, 189:104235.

Dinkelman, T. (2011). The Effects of Rural Electrification on Employment: New Evidence from South Africa. *American Economic Review*, 101(7):3078–3108.

Disdier, A.-C. and Head, K. (2008). The puzzling persistence of the distance effect on bilateral trade. *The Review of Economics and Statistics*, 90(1):37–48.

Disdier, A.-C. and Mayer, T. (2007). Je t'aime, moi non plus: Bilateral opinions and international trade. *European Journal of Political Economy*, 23(4):1140–1159.

Donaldson, D. (2018). Railroads of the Raj: Estimating the Impact of Transportation Infrastructure. *American Economic Review*, 108(4-5):899–934.

Donaldson, D. and Hornbeck, R. (2016). Railroads and american economic growth: A "market access" approach. *Quarterly Journal of Economics*, 131(2):799–858.

Donaldson, D. and Storeygard, A. (2016). The view from above: Applications of satellite data in economics. *Journal of Economic Perspectives*, 30(4):171–198.

Draca, M., Sadun, R., and van Reenen, J. (2007). Ict and productivity: A review of the evidence. *Handbook of Information and Communication Technologies*.

Drucker, P. (1969). *The Age of Discontinuity: Guidelines to Our Changing Society*. Taylor Francis.

Duede, E., Teplitskiy, M., Lakhani, K., and Evans, J. (2024). Being together in place as a catalyst for scientific advance. *Research Policy*, 53(2):104911.

Dugoua, E., Kennedy, R., and Urpelainen, J. (2018). Satellite data for the social sciences: Measuring rural electrification with night-time lights. *International Journal of Remote Sensing*, 39(9):2690–2701.

Dutta, S., Armanios, D. E., and Desai, J. D. (2022). Beyond spatial proximity: The impact of enhanced spatial connectedness from new bridges on entrepreneurship. *Organization Science*, 33(4):1620–1644.

D'Mello, M. and Sahay, S. (2007). 'i am kind of a nomad where i have to go places and places...': Understanding mobility, place and identity in global software work from india. *Information and Organization*, 17(3):162–192.

Eaton, J. and Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, 70(5):1741–1779.

Eckert, F., Hejlesen, M., and Walsh, C. (2022). The return to big-city experience: Evidence from refugees in denmark. *Journal of Urban Economics*, 130:103454.

Ellison, G., Glaeser, E. L., and Kerr, W. R. (2010). What causes industry agglomeration? evidence from coagglomeration patterns. *American Economic Review*, 100(3):1195–1213.

Emanuel, N., Harrington, E., and Pallais, A. (2023). The power of proximity: Training of tomorrow or productivity today? *Working Paper*.

Faber, B. (2014). Trade Integration, Market Size, and Industrialization: Evidence from China's National Trunk Highway System. *Review of Economic Studies*, 81(3):1046–1070.

Fackler, T. and Laurentsyeva, N. (2020). Gravity in online collaborations: Evidence from github. *CESifo Forum*, 21(03):15–20.

Fadeev, E. (2023). Creative construction: Knowledge sharing and cooperation between firms. *Working Paper*.

Falchetta, G., Pachauri, S., Byers, E., Danylo, O., and Parkinson, S. C. (2020). Satellite observations reveal inequalities in the progress and effectiveness of recent electrification in sub-saharan africa. *One Earth*, 2(4):364–379.

Falck, O., Heblich, S., Lameli, A., and Südekum, J. (2012). Dialects, cultural identity, and economic exchange. *Journal of Urban Economics*, 72(2-3):225–239.

Falck, O., Heimisch-Roecker, A., and Wiederhold, S. (2021). Returns to ict skills. *Research Policy*, 50(7):104064.

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2018). Global evidence on economic preferences. *Quarterly Journal of Economics*, 133(4):1645–1692.

Felbermayr, G. and Toubal, F. (2010). Cultural proximity and trade. *European Economic Review*, 54(2):279–293.

Fernandes, A. M., Mattoo, A., Nguyen, H., and Schiffbauer, M. (2019). The internet and chinese exports in the pre-ali baba era. *Journal of Development Economics*, 138:57–76.

Fershtman, C. and Gandal, N. (2004). The determinants of output per contributor in open source projects: An empirical examination. *CEPR Working Paper*.

Fetzer, T., Henderson, V., Nigmatulina, D., and Shanghavi, A. (2016). What Happens to Cities When Countries Become Democratic? *Working Paper*.

Fidrmuc, J. and Fidrmuc, J. (2014). Foreign languages and trade. *CESifo Working Paper*.

Fiol, C. M. and O'Connor, E. J. (2005). Identification in face-to-face, hybrid, and pure virtual teams: Untangling the contradictions. *Organization Science*, 16(1):19–32.

Firaz, A. (2022). How long do software engineers stay at a job? *LinkedIn Blog*.

Forman, C., Goldfarb, A., and Greenstein, S. (2016). Agglomeration of invention in the bay area: Not just ict. *American Economic Review*, 106(5):146–51.

Forman, C., Goldfarb, A., and Greenstein, S. (2018). How geography shapes – and is shaped by – the internet. In *The New Oxford Handbook of Economic Geography*, volume 269. Oxford University Press Oxford, UK.

Forman, C. and van Zeebroeck, N. (2019). Digital technology adoption and knowledge flows within firms: Can the internet overcome geographic and technological distance? *Research Policy*, 48(8):103697.

Forman, C. and Zeebroeck, N. v. (2012). From wires to partners: How the internet has fostered r&d collaborations within firms. *Management Science*, 58(8):1549–1568.

Frankel, J. and Rose, A. (2002). An estimate of the effect of common currencies on trade and income. *Quarterly Journal of Economics*, 117(2):437–466.

French, K. R. and Poterba, J. M. (1991). Investor diversification and international equity markets. *American Economic Review*, 81(2):222.

Freund, C. L. and Weinhold, D. (2004). The effect of the internet on international trade. *Journal of International Economics*, 62(1):171–189.

Frey, W. H. and Zimmer, Z. (2001). Defining the city. *Handbook of Urban Studies*, 1:14–35.

Fukui, R., Arderne, C. J., and Kelly, T. (2019). Africa's connectivity gap: Can a map tell the story? *World Bank Blog*.

Fuller, J., Langer, C., and Sigelman, M. (2022). Skills-based hiring is on the rise. *Harvard Business Review*, 11.

Garmendia, A., Llano, C., Minondo, A., and Requena, F. (2012). Networks and the disappearance of the intranational home bias. *Economics Letters*, 116(2):178–182.

Gaspar, J. and Glaeser, E. L. (1998). Information technology and the future of cities. *Journal of Urban Economics*, 43(1):136–156.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

Gerosa, M., Wiese, I., Trinkenreich, B., Link, G., Robles, G., Treude, C., Steinmacher, I., and Sarma, A. (2021). The shifting sands of motivation: Revisiting what drives contributors in open source. *IEEE International Conference on Software Engineering (ICSE)*, pages 1046–1058.

Ghani, E., Goswami, A. G., and Kerr, W. R. (2016). Highway to Success: The Impact of the Golden Quadrilateral Project for the Location and Performance of Indian Manufacturing. *Economic Journal*, 126(591):317–357.

Gibbs, M., Mengel, F., and Siemroth, C. (2023). Work from home and productivity: Evidence from personnel and analytics data on information technology professionals. *Journal of Political Economy Microeconomics*, 1(1):7–41.

Gibson, J., Olivia, S., Boe-Gibson, G., and Li, C. (2021). Which night lights data should we use in economics, and where? *Journal of Development Economics*, 149:102602.

Giroud, X., Lenzu, S., Maingi, Q., and Mueller, H. (2022). Propagation and amplification of local productivity spillovers. *NBER Working Paper*.

GitHub (2021). The 2021 state of the octoverse. *Company Report*.

Gitta, S. and Ikoja-Odongo, J. R. (2003). The impact of cybercafés on information services in uganda. *First Monday*.

Glaeser, C. K., Glaeser, S., and Labro, E. (2023). Proximity and the management of innovation. *Management Science*, 69(5):3080–3099.

Glaeser, E. L., Kallal, H. D., Scheinkman, J. A., and Shleifer, A. (1992). Growth in cities. *Journal of Political Economy*, 100(6):1126–1152.

Glaeser, E. L. and Mare, D. C. (2001). Cities and skills. *Journal of Labor Economics*, 19(2):316–342.

Glaeser, E. L. and Ponzetto, G. A. M. (2010). Did the Death of Distance Hurt Detroit and Help New York? In *Agglomeration Economics*. University of Chicago Press.

Glennon, B. (2024). Skilled immigrants, firms, and the global geography of innovation. *Journal of Economic Perspectives*, 38(1):3–26.

Gokmen, G. (2017). Clash of civilizations and the impact of cultural differences on trade. *Journal of Development Economics*, 127:449–458.

Goldbeck, M. (2023). Bit by bit: Colocation and the death of distance in software developer networks. *ifo Working Paper*.

Goldfarb, A. and Tucker, C. (2019). Digital economics. *Journal of Economic Literature*, 57(1):3–43.

Goldin, I., Koutroumpis, P., Lafond, F., and Winkler, J. (2024). Why is productivity slowing down? *Journal of Economic Literature*, 62(1):196–268.

Gomez, R. (2014). When you do not have a computer: Public-access computing in developing countries. *Information Technology for Development*, 20(3):274–291.

Gomez-Herrera, E., Martens, B., and Turlea, G. (2014). The drivers and impediments for cross-border e-commerce in the eu. *Information Economics and Policy*, 28:83–96.

Gordon, R. (2017). *The Rise and Fall of American Growth: The US Standard of Living since the Civil War*. Princeton University Press.

Gorodnichenko, Y. and Roland, G. (2017). Culture, institutions, and the wealth of nations. *Review of Economics and Statistics*, 99(3):402–416.

Gousios, G. (2013). The ghtorent dataset and tool suite. *IEEE Conference on Mining Software Repositories (MSR)*, pages 233–236.

Graafland, J. and de Jong, E. (2022). The moderating role of culture on the benefits of economic freedom: Cross-country analysis. *Journal of Comparative Economics*, 50(1):280–292.

Graham, M., Andersen, C., and Mann, L. (2015). Geographical imagination and technological connectivity in east africa. *Transactions of the Institute of British Geographers*, 40(3):334–349.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.

Gray, J. V., Siemsen, E., and Vasudeva, G. (2015). Colocation still matters: Conformance quality and the interdependence of r&d and manufacturing in the pharmaceutical industry. *Management Science*, 61(11):2760–2781.

Greenstein, S. (2015). *How the Internet Became Commercial: Innovation, Privatization, and the Birth of a New Network*. Princeton University Press.

Greenstone, M., Hornbeck, R., and Moretti, E. (2010). Identifying agglomeration spillovers: Evidence from winners and losers of large plant openings. *Journal of Political Economy*, 118(3):536–598.

Greenwood, M. J. (1973). The geographic mobility of college graduates. *The Journal of Human Resources*, 8(4):506–515.

Greenwood, M. J. (1975). Research on internal migration in the united states: A survey. *Journal of Economic Literature*, pages 397–433.

Grewal, R., Lilien, G. L., and Mallapragada, G. (2006). Location, location, location: How network embeddedness affects project success in open source systems. *Management Science*, 52(7):1043–1056.

Griffith, R., Lee, S., and van Reenen, J. (2011). Is distance dying at last? falling home bias in fixed-effects models of patent citations. *Quantitative Economics*, 2(2):211–249.

Griffith, T. L., Sawyer, J. E., and Neale, M. A. (2003). Virtualness and knowledge in teams: Managing the love triangle of organizations, individuals, and information technology. *MIS Quarterly*, pages 265–287.

Grogan, L. and Sadanand, A. (2013). Rural Electrification and Employment in Poor Countries: Evidence from Nicaragua. *World Development*, 43:252–265.

Grossman, G. M. and Rossi-Hansberg, E. (2008). Trading tasks: A simple theory of offshoring. *American Economic Review*, 98(5):1978–1997.

Guiso, L., Sapienza, P., and Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives*, 20(2):23–48.

Guiso, L., Sapienza, P., and Zingales, L. (2009). Cultural biases in economic exchange? *Quarterly Journal of Economics*, 124(3):1095–1131.

Haapanen, M. and Tervo, H. (2012). Migration of the highly educated: Evidence from residence spells of university graduates. *Journal of Regional Science*, 52(4):587–605.

Haftu, G. G. (2019). Information communications technology and economic growth in sub-saharan africa: A panel data approach. *Telecommunications Policy*, 43(1):88–99.

Hahn, J., Moon, J. Y., and Zhang, C. (2008). Emergence of new project teams from open source software developer networks: Impact of prior collaboration ties. *Information Systems Research*, 19(3):369–391.

Hakim Orman, W. (2008). Giving it away for free? the nature of job-market signaling by open-source software developers. *The BE Journal of Economic Analysis & Policy*, 8(1).

Hall, R. E. and Jones, C. I. (1999). Why do some countries produce so much more output per worker than others? *Quarterly Journal of Economics*, 114(1):83–116.

Hamilton, B. H., Nickerson, J. A., and Owan, H. (2003). Team incentives and worker heterogeneity: An empirical analysis of the impact of teams on productivity and participation. *Journal of Political Economy*, 111(3):465–497.

Hamilton Research (2020). Africa bandwidth maps.

Hann, I.-H., Roberts, J. A., and Slaughter, S. (2004). Why developers participate in open source software projects: An empirical investigation. *International Conference on Information Systems (ICIS)*.

Hann, I.-H., Roberts, J. A., and Slaughter, S. A. (2013). All are not equal: An examination of the economic returns to different forms of participation in open source software communities. *Information Systems Research*, 24(3):520–538.

Hanson, G. and Xiang, C. (2011). Trade barriers and trade flows with product heterogeneity: An application to us motion picture exports. *Journal of International Economics*, 83(1):14–26.

Harrigan, J., Reshef, A., and Toubal, F. (2021). The march of the techies: Job polarization within and between firms. *Research Policy*, 50(7):104008.

Harrigan, J., Reshef, A., and Toubal, F. (2023). Techies and firm level productivity. *NBER Working Paper*.

Hars, A. and Ou, S. (2002). Working for free? motivations for participating in open-source projects. *International Journal of Electronic Commerce*, 6(3):25–39.

Haussen, T. and Uebelmesser, S. (2018). Job changes and interregional migration of graduates. *Regional Studies*, 52(10):1346–1359.

Havranek, T. and Irsova, Z. (2017). Do borders really slash trade? a meta-analysis. *IMF Economic Review*, 65:365–396.

Head, K., Li, Y. A., and Minondo, A. (2019). Geography, ties, and knowledge flows: Evidence from citations in mathematics. *Review of Economics and Statistics*, 101(4):713–727.

Head, K. and Mayer, T. (2010). Illusory border effects: Distance mismeasurement inflates estimates of home bias in trade. *The Gravity Model in International Trade*.

Head, K. and Mayer, T. (2021). The united states of europe: A gravity model evaluation of the four freedoms. *Journal of Economic Perspectives*, 35(2):23–48.

Helliwell, J. F. and Verdier, G. (2001). Measuring internal trade distances: A new method applied to estimate provincial border effects in canada. *Canadian Journal of Economics*, pages 1024–1041.

Hellmanzik, C. and Schmitz, M. (2015). Virtual proximity and audiovisual services trade. *European Economic Review*, 77:82–101.

Hellmanzik, C. and Schmitz, M. (2016). Gravity and international services trade: The impact of virtual proximity. *European Economic Review*, 77:82–101.

Hellmanzik, C. and Schmitz, M. (2017). Taking gravity online: The role of virtual proximity in international finance. *Journal of International Money and Finance*, 77:164–179.

Helpman, E. (2009). *The Mystery of Economic Growth*. Harvard University Press.

Henderson, J. V. and Kriticos, S. (2018). The Development of the African System of Cities. *Annual Review of Economics*, 10(1):287–314.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2011). A Bright Idea for Measuring Economic Growth. *American Economic Review*, 101(3):194–199.

Henderson, J. V., Storeygard, A., and Weil, D. N. (2012). Measuring Economic Growth from Outer Space. *American Economic Review*, 102(2):994–1028.

Hendricks, K., Weiss, A., and Wilson, C. (1988). The war of attrition in continuous time with complete information. *International Economic Review*, pages 663–680.

Hersche, M. and Moor, E. (2020). Identification and estimation of intensive margin effects by difference-in-difference methods. *Journal of Causal Inference*, 8(1):272–285.

Hertel, G., Niedner, S., and Herrmann, S. (2003). Motivation of software developers in open source projects: An internet-based survey of contributors to the linux kernel. *Research Policy*, 32(7):1159–1177.

Hinds, P. J. and Bailey, D. E. (2003). Out of sight, out of sync: Understanding conflict in distributed teams. *Organization Science*, 14(6):615–632.

Hinds, P. J. and Mortensen, M. (2005). Understanding conflict in geographically distributed teams: The moderating effects of shared identity, shared context, and spontaneous communication. *Organization Science*, 16(3):290–307.

Hjort, J. and Poulsen, J. (2019). The Arrival of Fast Internet and Employment in Africa. *American Economic Review*, 109(3):1032–1079.

Hjort, J. and Tian, L. (2021). The Economic Impact of Internet Connectivity in Developing Countries. *Working Paper*.

Hoekman, J., Frenken, K., and Tijssen, R. J. W. (2010). Research collaboration at a distance: Changing spatial patterns of scientific collaboration within europe. *Research Policy*, 39(5):662–673.

Hoffmann, M., Nagle, F., and Zhou, Y. (2024). The value of open source software. *Harvard Business School Strategy Unit Working Paper*.

Hofstede, G. (2011). Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, 2(1):8.

Holmström, B. (1999). Managerial incentive problems: A dynamic perspective. *The Review of Economic Studies*, 66(1):169–182.

Huang, P. and Zhang, Z. (2016). Participation in open knowledge communities and job-hopping. *MIS Quarterly*, 40(3):785–806.

Huang, R. (2007). Distance and trade: Disentangling unfamiliarity effects and transport cost effects. *European Economic Review*, 51(1):161–181.

Invesco (2019). China embarks on digital silk road. *Invesco Insight*.

ITU (2019). *Economic Impact of Broadband in LDCs, LLDCs and SIDS*. International Telecommunications Union.

Jaffe, A. B., Trajtenberg, M., and Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations. *Quarterly Journal of Economics*, 108(3):577–598.

Jang, S. (2017). Cultural brokerage and creative performance in multicultural teams. *Organization Science*, 28(6):993–1009.

Jedrusik, A. and Wadsworth, P. (2017). Patent protection for software-implemented inventions. *WIPO Magazine*, pages 7–11.

Jedwab, R., Kerby, E., and Moradi, A. (2017). History, path dependence and development: Evidence from colonial railways, settlers and cities in kenya. *Economic Journal*, 127(603):1467–1494.

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

Johnson, J. P. (2002). Open source software: Private provision of a public good. *Journal of Economics & Management Strategy*, 11(4):637–662.

Johnson, K. P. and Kort, J. R. (2004). 2004 redefinition of the bea economic areas. *Survey of Current Business*, 75(2):75–81.

Jones, B. F. (2009). The burden of knowledge and the death of the renaissance man: Is innovation getting harder? *The Review of Economic Studies*, 76(1):283–317.

Jones, B. F. (2021). The rise of research teams: Benefits and costs in economics. *Journal of Economic Perspectives*, 35(2):191–216.

Jorgenson, D. W. and Stiroh, K. J. (1999). Information technology and growth. *American Economic Review*, 89(2):109–115.

Karl, K. A., Peluchette, J. V., and Aghakhani, N. (2022). Virtual work meetings during the covid-19 pandemic: The good, bad, and ugly. *Small Group Research*, 53(3):343–365.

Keller, W. and Yeaple, S. R. (2013). The gravity of knowledge. *American Economic Review*, 103(4):1414–1444.

Kende, M. and Rose, K. (2015). Promoting Local Content Hosting to Develop the Internet Ecosystem. Technical report, Internet Society.

Kerr, W. R. (2020). *The Gift of Global Talent: How Migration Shapes Business, Economy & Society*. Stanford University Press.

Kešeljević, A. and Spruk, R. (2023). Long-term effects of the yugoslav war. *Defence and Peace Economics*, 34(4):410–436.

Kitimbo, T. (2023). Why ugandan internet cafes did not die. *Nile Post*.

Koçak, Ö. and Puranam, P. (2022). Separated by a common language: How the nature of code differences shapes communication success and code convergence. *Management Science*, 68(7):5287–5310.

Kodrzycki, Y. K. (2001). Migration of recent college graduates: Evidence from the national longitudinal survey of youth. *New England Economic Review*, pages 13–34.

Kolko, J. (2012). Broadband and Local Growth. *Journal of Urban Economics*, 71(1):100–113.

Kondo, J., Li, D., and Papanikolaou, D. (2021). Trust, collaboration, and economic growth. *Management Science*, 67(3):1825–1850.

Korkmaz, G., Calderón, J. B. S., Kramer, B. L., Guci, L., and Robbins, C. A. (2024). From github to gdp: A framework for measuring open source software innovation. *Research Policy*, 53(3):104954.

Krishnamurthy, S. (2006). On the intrinsic and extrinsic motivation of free/libre/open source (floss) developers. *Knowledge, Technology & Policy*, 18(4):17–39.

Krishnamurthy, S., Ou, S., and Tripathi, A. K. (2014). Acceptance of monetary rewards in open source software development. *Research Policy*, 43(4):632–644.

La Porta, R., Lopez-de Silanes, F., Shleifer, A., and Vishny, R. (1999). The quality of government. *Journal of Law, Economics, and Organization*, 15(1):222–279.

Lagakos, D. (2020). Urban-rural gaps in the developing world: Does internal migration offer opportunities? *Journal of Economic Perspectives*, 34(3):174–192.

Lahiri, N. (2010). Geographic distribution of r&d activity: How does it affect innovation quality? *Academy of Management Journal*, 53(5):1194–1209.

Lakhani, K. R. and Wolf, R. G. (2003). Why hackers do what they do: Understanding motivation and effort in free/open source software projects. *Open Source Software Projects*.

Larch, M., Wanner, J., Yotov, Y. V., and Zylkin, T. (2019). Currency unions and trade: A ppml re-assessment with high-dimensional fixed effects. *Oxford Bulletin of Economics and Statistics*, 81(3):487–510.

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., and Margetts, H. (2020). Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

LeBlanc, M. and Shrum, W. (2017). The evolution of ghanaian internet cafés, 2003–2014. *Information Technology for Development*, 23(1):86–106.

Lee, G. K. and Cole, R. E. (2003). From a firm-based to a community-based model of knowledge creation: The case of the linux kernel development. *Organization Science*, 14(6):633–649.

Lee, K., Miguel, E., and Wolfram, C. (2020). Experimental evidence on the economics of rural electrification. *Journal of Political Economy*, 128(4):1523–1565.

Lee, S. (2019). Learning-by-moving: Can reconfiguring spatial proximity between organizational members promote individual-level exploration? *Organization Science*, 30(3):467–488.

Leiserson, C. E., Thompson, N. C., Emer, J. S., Kuszmaul, B. C., Lampson, B. W., Sanchez, D., and Schardl, T. B. (2020). There's plenty of room at the top: What will drive computer performance after moore's law? *Science*, 368(6495):eaam9744.

Lendle, A., Olarreaga, M., Schropp, S., and Vézina, P.-L. (2016). There goes gravity: ebay and the death of distance. *Economic Journal*, 126(591):406–441.

Leppämäki, M. and Mustonen, M. (2009). Skill signalling with product market externality. *Economic Journal*, 119(539):1130–1142.

Lerner, J. and Tirole, J. (2002). Some simple economics of open source. *The Journal of Industrial Economics*, 50(2):197–234.

Lerner, J. and Tirole, J. (2005a). The economics of technology sharing: Open source and beyond. *Journal of Economic Perspectives*, 19(2):99–120.

Lerner, J. and Tirole, J. (2005b). The scope of open source licensing. *Journal of Law, Economics, and Organization*, 21(1):20–56.

Leuven, E., Akerman, A., and Mogstad, M. (2021). Information frictions, internet and the relationship between distance and trade. *American Economic Journal: Applied Economics*.

Levin, N. and Duke, Y. (2012). High spatial resolution night-time light images for demographic and socio-economic studies. *Remote Sensing of Environment*, 119:1–10.

Levin, N., Kyba, C. C., Zhang, Q., Sánchez de Miguel, A., Román, M. O., Li, X., Portnov, B. A., Molthan, A. L., Jechow, A., Miller, S. D., Wang, Z., Shrestha, R. M., and Elvidge, C. D. (2020). Remote sensing of night lights: A review and an outlook for the future. *Remote Sensing of Environment*, 237.

Lewer, J. J. and van den Berg, H. (2008). A gravity model of immigration. *Economics Letters*, 99(1):164–167.

Li, X., Li, D., Xu, H., and Wu, C. (2017). Intercalibration between DMSP/OLS and VIIRS Night-time Light Images to Evaluate City Light Dynamics of Syria's Major Human Settlement During Syrian Civil War. *International Journal of Remote Sensing*, 38(21):5934–5951.

Li, X., Zhou, Y., Zhao, M., and Zhao, X. (2020). A Harmonized Global Nighttime Light Dataset 1992–2018. *Scientific Data*, 7(1):1–9.

Li, Y. A. (2014). Borders and distance in knowledge spillovers: Dying over time or dying with age? evidence from patent citations. *European Economic Review*, 71:152–172.

Lifshitz-Assaf, H. and Nagle, F. (2021). The digital economy runs on open source. here's how to protect it. *Harvard Business Review*.

Lin, Y., Frey, C. B., and Wu, L. (2023). Remote collaboration fuses fewer breakthrough ideas. *Nature*, 623:987–991.

Lipsey, R. G., Carlaw, K. I., and Bekar, C. T. (2005). *Economic Transformations: General Purpose Technologies and Long-term Economic Growth*. Oxford University Press.

Long, J. (2009). Open source software development experiences on the students' resumes: Do they count? *Journal of Information Technology Education: Research*, 8(1):229–242.

Lubwama, J. (2023). Internet cafes introduced uganda to the internet. *Rest of World*.

Luca, M. (2015). User-generated content and social media. In *Handbook of Media Economics*, volume 1, pages 563–592. Elsevier.

Ma, L., Wu, J., Li, W., Peng, J., and Liu, H. (2014). Evaluating saturation correction methods for DMSP/OLS nighttime light data: A case study from china's cities. *Remote Sensing*, 6(10):9853–9872.

Määttä, I., Ferreira, T., and Lessmann, C. (2022). Nighttime lights and wealth in very small areas: Namibian complete census versus dhs data. *Review of Regional Research*, 42(2):161–190.

Määttä, I. and Lessmann, C. (2019). Human lights. *Remote Sensing*, 11(19):2194.

Machin, S., Salvanes, K. G., and Pelkonen, P. (2012). Education and mobility. *Journal of the European Economic Association*, 10(2):417–450.

Mack, C. A. (2011). Fifty years of moore's law. *IEEE Transactions on Semiconductor Manufacturing*, 24(2):202–207.

MacKinnon, J. G. and Magee, L. (1990). Transforming the dependent variable in regression models. *International Economic Review*, pages 315–339.

Majchrzak, A., Rice, R. E., Malhotra, A., King, N., and Ba, S. (2000). Technology adaptation: The case of a computer-supported inter-organizational virtual team. *MIS Quarterly*, pages 569–600.

Manning, A. and Petrongolo, B. (2017). How local are labor markets? evidence from a spatial job search model. *American Economic Review*, 107(10):2877–2907.

Marlow, J. and Dabbish, L. (2013). Activity traces and signals in software developer recruitment and hiring. *ACM SIGCHI Conference on Computer Supported Cooperative Work (CSCW)*, pages 145–156.

Marshall, A. (1920). *Principles of Economics*. MacMillan.

Masaki, T., Ochoa, R., and Rodríguez-Castelán, C. (2020). Broadband internet and household welfare in senegal. *IZA Discussion Paper*.

Maurseth, P. B. and Verspagen, B. (2002). Knowledge spillovers in europe: A patent citations analysis. *Scandinavian Journal of Economics*, 104(4):531–545.

Mayer, T. and Zignago, S. (2011). Notes on cepii's distances measures: The geodist database. *CEPII Working Paper*.

Maznevski, M. L. and Chudoba, K. M. (2000). Bridging space over time: Global virtual team dynamics and effectiveness. *Organization Science*, 11(5):473–492.

Mbarika, V., Kah, M. M. O., and Keita, M. (2004). The diffusion of cyber cafés in sub-saharan africa: Country case studies. *IRMA International Conference*, pages 964–967.

McCallum, J. (1995). National borders matter: Canada-us regional trade patterns. *American Economic Review*, 85(3):615–623.

McKague, D., Zurbuchen, T. H., Donajkowski, T., Ervin, J., Heckathorn, D., and Moran, K. (2009). Imagine africa: Providing internet to the developing world. *IEEE Aerospace Conference (AeroConf)*, pages 1–9.

Megasis Network (2023). Ai and language translation: Breaking down language barriers. *Medium*.

Melitz, J. (2008). Language and foreign trade. *European Economic Review*, 52(4):667–699.

Melitz, J. and Toubal, F. (2014). Native language, spoken language, translation and trade. *Journal of International Economics*, 93(2):351–363.

Mellander, C., Lobo, J., Stolarick, K., and Matheson, Z. (2015). Night-time light data: A good proxy measure for economic activity? *PLoS ONE*, 10(10):1–18.

Michaels, G., Natraj, A., and van Reenen, J. (2014). Has ict polarized skill demand? evidence from eleven countries over twenty-five years. *Review of Economics and Statistics*, 96(1):60–77.

Miklós-Thal, J. and Ullrich, H. (2015). Belief precision and effort incentives in promotion contests. *Economic Journal*, 125(589):1952–1963.

Miklós-Thal, J. and Ullrich, H. (2016). Career prospects and effort incentives: Evidence from professional soccer. *Management Science*, 62(6):1645–1667.

Millimet, D. L. and Osang, T. (2007). Do state borders matter for us intranational trade? the role of history and internal migration. *Canadian Journal of Economics*, 40(1):93–126.

Mokyr, J. (2002). *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton University Press.

Montobbio, F. and Sterzi, V. (2013). The globalization of technology in emerging markets: A gravity model on the determinants of international patent collaborations. *World Development*, 44:281–299.

Moretti, E. (2021). The effect of high-tech clusters on the productivity of top inventors. *American Economic Review*, 111(10):3328–75.

Moretti, E. and Yi, M. (2023). Size matters: The benefits of large labor markets for job seekers. *Working Paper*.

Morgan, T. (2022). supercompress: Stata module to run compress on all datasets in a folder and its subfolders. *Statistical Software Components*.

Muenchen, R. A. (2012). The popularity of data analysis software. *R4Stats*.

Mullahy, J. and Norton, E. C. (2022). Why transform y? a critical assessment of dependent-variable transformations in regression models for skewed and sometimes-zero outcomes. *NBER Working Paper*.

Murdock, G. P. (1959). *Africa: Its Peoples and their Culture History*. McGraw Hill Text.

Mwesige, P. G. (2004). Cyber elites: A survey of internet café users in uganda. *Telematics and Informatics*, 21(1):83–101.

Myatt, D. P. and Wallace, C. (2002). Equilibrium selection and public-good provision: The development of open-source software. *Oxford Review of Economic Policy*, 18(4):446–461.

Nagle, F. (2018). Learning by contributing: Gaining competitive advantage through contribution to crowdsourced public goods. *Organization Science*, 29(4):569–587.

Nagle, F. (2019). Open-source software and firm productivity. *Management Science*, 65(3):1191–1215.

Nagle, F. (2022). Strengthening digital infrastructure: A policy agenda for free and open source software. *Brookings Institution Policy Brief*.

Nagle, F., Wheeler, D. A., Lifshitz-Assaf, H., Ham, H., and Hoffman, J. (2020). Report on the 2020 foss contributor survey. *The Linux Foundation Core Infrastructure Initiative*.

Ndiomewese, I. (2015). A tribute to cybercafes and their undeniable role in nigeria's internet revolution. *Techpoint Africa*.

Ngai, L. R. and Pissarides, C. A. (2007). Structural change in a multisector model of growth. *American Economic Review*, 97(1):429–443.

Ngari, L. and Petrack, S. A. (2019). Internet infrastructure in africa. *Empower Africa*.

Nordhaus, W. and Chen, X. (2015). A sharper image? estimates of the precision of nighttime lights as a proxy for economic statistics. *Journal of Economic Geography*, 15(1):217–246.

Nunn, N. and Puga, D. (2012). Ruggedness: The Blessing of Bad Geography in Africa. *Review of Economics and Statistics*, 94(1):20–36.

Nyezi, C. (2012). Case study: Connecting rural africa to the internet. *How We Made It in Africa*.

Obradovich, N., Özak, Ö., Martín, I., Ortuño-Ortín, I., Awad, E., Cebrián, M., Cuevas, R., Desmet, K., Rahwan, I., and Cuevas, Á. (2022). Expanding the measurement of culture with a sample of two billion humans. *Journal of the Royal Society Interface*, 19(190):20220085.

OECD (2020). *Africa's Urbanisation Dynamics 2020*. OECD Publishing.

OECD (2021). *The Digital Transformation of SMEs*. OECD Publishing.

Olofinlua, O. (2015). How mobile internet killed off cyber cafés in nigeria. *Quartz*.

Olson, G. M. and Olson, J. S. (2000). Distance matters. *Human-Computer Interaction*, 15(2-3):139–178.

Osho, O. and Adepoju, S. A. (2016). Cybercafés in nigeria: Curse to the internet. *International Conference on Information and Communication Technology and Applications (ICTA)*.

Osterloh, M. and Rota, S. (2007). Open source software development: Just another case of collective invention? *Research Policy*, 36(2):157–171.

O'Mahony, S. (2003). Guarding the commons: How community managed software projects protect their work. *Research Policy*, 32(7):1179–1198.

O'Neil, M., Muselli, L., Cai, X., and Zacchiroli, S. (2022). Co-producing industrial public goods on github: Selective firm cooperation, volunteer-employee labour and participation inequality. *New Media & Society*.

Page, S. E. (2010). *Diversity and Complexity*. Princeton University Press.

Palfrey, T. R. and Rosenthal, H. (1984). Participation and the provision of discrete public goods: A strategic analysis. *Journal of Public Economics*, 24(2):171–193.

Pallais, A. (2014). Inefficient hiring in entry-level labor markets. *American Economic Review*, 104(11):3565–3599.

Park, M., Leahey, E., and Funk, R. J. (2023). Papers and patents are becoming less disruptive over time. *Nature*, 613(7942):138–144.

Parrotta, P., Pozzoli, D., and Pytlikova, M. (2014). Labor diversity and firm productivity. *European Economic Review*, 66:144–179.

Pentland, A. S. (2012). The new science of building great teams. *Harvard Business Review*, 90(4):60–69.

Peri, G. (2005). Determinants of knowledge flows and their effect on innovation. *Review of Economics and Statistics*, 87(2):308–322.

Picci, L. (2010). The internationalization of inventive activity: A gravity model using patent data. *Research Policy*, 39(8):1070–1081.

Piopiunik, M., Schwerdt, G., Simon, L., and Woessmann, L. (2020). Skills, signals, and employability: An experimental investigation. *European Economic Review*, 123:103374.

Polzer, J. T., Crisp, C. B., Jarvenpaa, S. L., and Kim, J. W. (2006). Extending the faultline model to geographically dispersed teams: How colocated subgroups can impair group functioning. *Academy of Management Journal*, 49(4):679–692.

Powell, W. W. and Snellman, K. (2004). The knowledge economy. *Annual Review of Sociology*, 30:199–220.

Puranam, P., Alexy, O., and Reitzig, M. (2014). What's 'new' about new forms of organizing? *Academy of Management Review*, 39(2):162–180.

Quadri, S. (2023). How lagos' bubbly internet cafe culture went flat. *Rest of World*.

Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E., and Aral, S. (2022). A causal test of the strength of weak ties. *Science*, 377(6612):1304–1310.

Rascouet, A., Prinsloo, L., and Seal, T. (2020). Faster internet coming to africa with facebook's $1 billion cable. *Bloomberg News*.

Raveendran, M., Puranam, P., and Warglien, M. (2022). Division of labor through self-selection. *Organization Science*, 33(2):810–830.

Reagans, R., Argote, L., and Brooks, D. (2005). Individual experience and experience working together: Predicting learning rates from knowing who knows what and knowing how to work together. *Management Science*, 51(6):869–881.

Ren, Y., Chen, J., and Riedl, J. (2016). The impact and evolution of group diversity in online open collaboration. *Management Science*, 62(6):1668–1686.

Roberts, J. A., Hann, I.-H., and Slaughter, S. A. (2006). Understanding the motivations, participation, and performance of open source software developers: A longitudinal study of the apache projects. *Management Science*, 52(7):984–999.

Rodríguez-Castelán, C., Ochoa, R. G., Lach, S., and Masaki, T. (2021). Mobile internet adoption in west africa. *IZA Discussion Paper*.

Roessler, P., Carroll, P., Myamba, F., Jahari, C., Kilama, B., and Nielson, D. (2021). The economic impact of mobile phone ownership: Results from a randomized controlled trial in tanzania. *Working Paper*.

Romer, P. M. (1986). Increasing returns and long-run growth. *Journal of Political Economy*, 94(5):1002–1037.

Romer, P. M. (1990). Endogenous technological change. *Journal of Political Economy*, 98(5):71–102.

Roodman, D. (2024). The arrival of fast internet and employment in africa: Comment. *arXiv*.

Rotondi, V., Kashyap, R., Pesando, L. M., Spinelli, S., and Billari, F. C. (2020). Leveraging mobile phones to attain sustainable development. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, 117(24):13413–13420.

Royston, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 43(3):429–453.

Royston, P. and Sauerbrei, W. (2008). *Multivariable Model-Building: A Pragmatic Approach to Regression Anaylsis Based on Fractional Polynomials for Modelling Continuous Variables*. John Wiley & Sons.

Rud, J. P. (2012). Electricity Provision and Industrial Development: Evidence from India. *Journal of Development Economics*, 97(2):352–367.

Sakketa, T. G. (2023). Urbanisation and rural development in sub-saharan africa: A review of pathways and impacts. *Research in Globalization*, page 100133.

Samuel, A. (2015). Collaborating online is sometimes better than face-to-face. *Harvard Business Review*, 17.

Santamaría, M., Ventura, J., and Yeşilbayraktar, U. (2023a). Borders within europe. *Working Paper*.

Santamaría, M., Ventura, J., and Yeşilbayraktar, U. (2023b). Exploring european regional trade. *Journal of International Economics*, page 103747.

Seliger, F., Kozak, J., and de Rassenfosse, G. (2019). Geocoding of Worldwide Patent Data. *Harvard Dataverse*.

Shah, S. K. (2006). Motivation, governance, and the viability of hybrid forms in open source software development. *Management Science*, 52(7):1000–1014.

Simon, H. A. (1979). Rational decision making in business organizations. *American Economic Review*, 69(4):493–513.

Singh, J. and Marx, M. (2013). Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity. *Management Science*, 59(9):2056–2078.

Smirnova, I., Reitzig, M., and Alexy, O. (2022). What makes the right oss contributor tick? treatments to motivate high-skilled developers. *Research Policy*, 51(1):104368.

Solimano, A. (2006). *The International Mobility of Talent and Its Impact on Global Development: An Overview*. ECLAC.

Southwood, R. (2022). Bandwidth as the digital economy's fuel: Getting sub-saharan africa connected (1991–2015). In *Africa 2.0: Inside a Continent's Communications Revolution*, pages 48–82. Manchester University Press.

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, pages 355–374.

Spolaore, E. and Wacziarg, R. (2009). The diffusion of development. *Quarterly Journal of Economics*, 124(2):469–529.

Startlin (2016). History of github. *Infographic*.

Steegmans, J. and de Bruin, J. (2021). Online housing search: A gravity model approach. *PLoS ONE*, 16(3):e0247712.

Steinwender, C. (2018). Real effects of information frictions: When the states and the kingdom became united. *American Economic Review*, 108(3):657–96.

Stewart, K. J. and Gosain, S. (2006). The impact of ideology on effectiveness in open source software development teams. *MIS Quarterly*, pages 291–314.

Storeygard, A. (2016). Farther on down the road: Transport costs, trade and urban growth in sub-Saharan Africa. *Review of Economic Studies*, 83(3):1263–1295.

Stork, C., Calandro, E., and Gamage, R. (2014). The future of broadband in africa. *info*, 16(1):76–93.

Strong, N. and Xu, X. (2003). Understanding the equity home bias: Evidence from survey data. *Review of Economics and Statistics*, 85(2):307–312.

Surakka, S. (2007). What subjects and skills are important for software developers? *Communications of the ACM*, 50(1):73–78.

Suri, T. and Bhattacharya, K. (2022). Impacts of the internet on the poor. *Working Paper*.

Synopsys (2023). Open source security and risk analysis report 2023. *Report*.

Tadesse, B. and White, R. (2010). Cultural distance as a determinant of bilateral trade flows: Do immigrants counter the effect of cultural differences? *Applied Economics Letters*, 17(2):147–152.

Thompson, H. G. and Garbacz, C. (2011). Economic impacts of mobile versus fixed broadband. *Telecommunications Policy*, 35(11):999–1009.

Thompson, P. and Fox-Kean, M. (2005). Patent citations and the geography of knowledge spillovers: A reassessment. *American Economic Review*, 95(1):450–460.

Thung, F., Bissyande, T. F., Lo, D., and Jiang, L. (2013). Network structure of social coding in github. *IEEE Conference on Software Maintenance and Reengineering (CSMR)*, pages 323–326.

Tinbergen, J. (1962). An analysis of world trade flows. *Shaping the World Economy*, 3:1–117.

Tubino, L., Cain, A., Schneider, J.-G., Thiruvady, D., and Fernando, N. (2020). Authentic individual assessment for team-based software engineering projects. *IEEE International Conference on Software Engineering, Education and Training (CSEE&T)*, pages 71–81.

Urry, J. (2002). Mobility and proximity. *Sociology*, 36(2):255–274.

Uzzi, B., Mukherjee, S., Stringer, M., and Jones, B. (2013). Atypical combinations and scientific impact. *Science*, 342(6157):468–472.

van der Kamp, J. (1977). The gravity model and migration behaviour: An economic interpretation. *Journal of Economic Studies*, 4(2):89–102.

van der Wouden, F. and Youn, H. (2023). The impact of geographical distance on learning through collaboration. *Research Policy*, 52(2):104698.

van Knippenberg, D. and Schippers, M. C. (2007). Work group diversity. *Annual Review of Psychology*, 58:515–541.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives*, 28(2):3–28.

Venhorst, V., van Dijk, J., and van Wissen, L. (2011). An analysis of trends in spatial mobility of dutch graduates. *Spatial Economic Analysis*, 6(1):57–82.

Vidoni, M. (2022). A systematic process for mining software repositories: Results from a systematic literature review. *Information and Software Technology*, 144:106791.

Visser, R. (2019). The effect of the internet on the margins of trade. *Information Economics and Policy*, 46:41–54.

von Engelhardt, S. and Freytag, A. (2013). Institutions, culture, and open source. *Journal of Economic Behavior & Organization*, 95:90–110.

von Krogh, G., Haefliger, S., Spaeth, S., and Wallin, M. W. (2012). Carrots and rainbows: Motivation and social practice in open source software development. *MIS Quarterly*, pages 649–676.

von Krogh, G., Spaeth, S., and Lakhani, K. R. (2003). Community, joining, and specialization in open source software innovation: A case study. *Research Policy*, 32(7):1217–1241.

von Proff, S., Duschl, M., and Brenner, T. (2017). Motives behind the mobility of university graduates: A study of three german universities. *Review of Regional Research*, 37:39–58.

Wachs, J., Nitecki, M., Schueller, W., and Polleres, A. (2022). The geography of open-source software: Evidence from github. *Technological Forecasting and Social Change*, 176:121478.

Wagner, S. and Ruhe, M. (2018). A systematic review of productivity factors in software development. *arXiv Preprint*.

Waldinger, F. (2012). Peer effects in science: Evidence from the dismissal of scientists in nazi germany. *The Review of Economic Studies*, 79(2):838–861.

Waltman, L., Tijssen, R. J. W., and van Eck, N. J. (2011). Globalisation of science in kilometres. *Journal of Informetrics*, 5(4):574–582.

Wang, Y., Wang, L., Hu, H., Jiang, J., Kuang, H., and Tao, X. (2022). The influence of sponsorship on open-source software developers' activities on github. *IEEE Computers, Software, and Applications Conference (COMPSAC)*, pages 924–933.

Wei, Y., Liu, H., Song, W., Yu, B., and Xiu, C. (2014). Normalization of time series DMSP-OLS nighttime light images for urban growth analysis with pseudo invariant features. *Landscape and Urban Planning*, 128:1–13.

Weidmann, N. B., Rød, J. K., and Cederman, L.-E. (2010). Representing ethnic groups in space: A new dataset. *Journal of Peace Research*, 47(4):491–499.

Weitzman, M. L. (1998). Recombinant growth. *Quarterly Journal of Economics*, 113(2):331–360.

Wen, W., Ceccagnoli, M., and Forman, C. (2016). Opening up intellectual property strategy: Implications for open source software entry by start-up firms. *Management Science*, 62(9):2668–2691.

Wickham, H., Hester, J., Francois, R., Bryan, J., Bearrows, S., Jylänki, J., and Jørgensen, M. (2024). readr. *CRAN*.

Williams, I., Gyaase, P. O., and Falch, M. (2012). Enhancing rural connectivity through an extended internet cafés business model. *International Telecommunications Society Conference (ITS)*.

Williams, M. D. J. (2010). *Broadband for Africa: Developing Backbone Communications Networks*. World Bank.

Williams, M. D. J., Mayer, R., and Minges, M. (2011). Africa's ICT infrastructure: Building on the mobile revolution. *World Bank Report*.

Wolf, H. C. (2000). Intranational home bias in trade. *Review of Economics and Statistics*, 82(4):555–563.

World Bank (2016). World Development Report 2016: Digital Dividends. *World Bank Report*.

Wright, N. L., Nagle, F., and Greenstein, S. (2023). Open source software and global entrepreneurship. *Research Policy*, 52(9):104846.

Wu, L., Wang, D., and Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, 566(7744):378–382.

Wuchty, S., Jones, B. F., and Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039.

Xu, L., Nian, T., and Cabral, L. (2020). What makes geeks tick? a study of stack overflow careers. *Management Science*, 66(2):587–604.

Yang, L., Holtz, D., Jaffe, S., Suri, S., Sinha, S., Weston, J., Joyce, C., Shah, N., Sherman, K., Hecht, B., and Teevan, J. (2022). The effects of remote work on collaboration among information workers. *Nature Human Behaviour*, 6(1):43–54.

Yotov, Y. V. (2012). A simple solution to the distance puzzle in international trade. *Economics Letters*, 117(3):794–798.

Yotov, Y. V. (2022). On the role of domestic trade flows for estimating the gravity model of trade. *Contemporary Economic Policy*, 40(3):526–540.

Zammuto, R. F., Griffith, T. L., Majchrzak, A., Dougherty, D. J., and Faraj, S. (2007). Information technology and the changing fabric of organization. *Organization Science*, 18(5):749–762.

Zeitlyn, D. (2003). Gift economies in the development of open source software: Anthropological reflections. *Research Policy*, 32(7):1287–1291.