Mähr, Maximilian; Oehlen, Jens

Conference Paper

Leveling the Playing Field: Knowledge Production in the Digital Age

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage"

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Mähr, Maximilian; Oehlen, Jens (2023) : Leveling the Playing Field: Knowledge Production in the Digital Age, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2023: Growth and the "sociale Frage", ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/277709

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

Maximilian Mähr† Jens Oehlen‡

January 2022

Abstract

Most scientific papers are not freely available—even though access to existing knowledge is crucial for pushing the research frontier. In this paper, we quantify the effects of lifting access restrictions on scientific research, focusing on Sci-Hub. Sci-Hub is an online platform founded in 2011 that offers access to most scientific articles worldwide and for free. We build a global panel of sub-national units over two decades using 300 million geo-coded download requests and the near-universe of scientific articles. Relying on an instrumented difference-in-differences design, we document that Sci-Hub has significantly and meaningfully changed consumption patterns of existing scientific works. After the website's launch, references to closed-access publications increased markedly. Greater access to frontier knowledge, in turn, led to higher-quality research output but only in middle-income countries.

Keywords: : Science of Science, Open Access

JEL Codes: L86, O30

*We are grateful to Antonio Ciccone, Matthew Gentzkow, Giuseppe Sorrenti and David Strömberg for advice as well as César Hidalgo for a helpful discussion. We thank audiences at the 7th IZA Workshop on the Economics of Education, WICK #10 Workshop at Collegio Carlo Alberto, University of Mannheim, Stanford University, and Stockholm University for comments.

†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 produced by scientific 'giants' (Azoulay, Graff Zivin and Wang, 2010; Iaria, Schwarz and Waldinger, 2018). 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. For example, only 12 percent of peer-reviewed academic journals are published under open access, the practice of providing online access to scientific information free of charge (Archambault et al., 2014).

To what extent do monetary restrictions inhibit further knowledge production? Despite potentially grave impacts, rigorous evidence on this question is surprisingly scant. The key reason is that endogeneity concerns easily plague empirical analyses. Researchers' access conditions are typically tied to their academic institutions. Hence, any comparison will likely capture unobserved differences in their broader academic environment.

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 transformed when vast amounts of existing knowledge become freely available through *Sci-Hub*. Sci-Hub is an online media tool that offers free access to most scientific articles worldwide. Launched in 2011, the website has garnered a global audience currently counting roughly 3 million paper downloads per day.[1] Relying on the website's diffusion through social networks, we implement an instrumented difference-in-differences design using a global panel of sub-national units from 2000 to 2022. We find that Sci-Hub had a significant and quantitatively large impact on the type of research consumed: doubling Sci-Hub intensity in a given region leads researchers to include five percentage points more closed-access papers in their reference lists. In return, we find that this increase in frontier research consumption transitions into the production of higher quality research, but only in middle-income countries.

The primary data underlying our empirical exercises are Sci-Hub server log files. Starting from 2011, the year Sci-Hub went online, until 2017 we observe 300 million access requests worldwide. Each data point contains the time, date, paper, and, most importantly, the IP's associated geolocation. Mapping these downloads, we build a global time-varying measure

[1] Source: sci-hub.se/stats, accessed on 26th of November 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 Mio unique visitors per day in 2017 (Fiorini, Lipman and Lu, 2017).

of Sci-Hub intensity for sub-national regions worldwide. The second primary data source is OpenAlex. OpenAlex is the successor project to Microsoft Academic Graph and includes global information on scientific articles. We use these data to construct measures of citations to closed-access papers and the geography of newly written scientific articles.

Leveraging our large data, we begin by documenting four descriptive facts. First, we show that monetary restrictions are pervasive, yet particularly binding for top-quality journals. On average, only 15% of journals operate under open access regimes. Yet, in the top percentile of all journals, as measured by impact factor, the fraction of open access publications amounts to only 5%. If scientists had bulk access through their libraries and institutions, this would not be a significant issue. However, our second fact speaks against an equal distribution of access. We find that institutions in lesser developed regions are much less likely to have JSTOR subscriptions, a common online library with bulk access to almost 3000 journals.[2] Third, the unequal distribution of access does not simply mimic an unequal distribution of demand for high-quality knowledge. In the Sci-Hub data, we find that most downloads per researcher stem predominantly from developing and emerging countries. We further show that the demand is particularly large for high-quality journals. Finally, we document significant differences in the production of high-quality research across less and highly-developed countries. 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 to what extent access restrictions *cause* the unequal distribution of high-quality knowledge production.

We answer this question using an instrumented difference-in-differences framework. We do so because Sci-Hub traffic across the world is not randomly distributed. We, therefore, isolate quasi-exogenous variation through social networks. Akin to papers in the existing media literature, we argue that social connections often drive technology adoption. Since Sci-Hub was founded in Almaty, Kazakhstan, and does not operate with large marketing budgets, we use friendship links measured by Facebook to construct a global diffusion network. Using connectedness to the Almaty region interacted with a post-2011 indicator variable, we then instrument for Sci-Hub intensity. The identifying assumption is that scientific outcomes in regions with different degrees of connectedness to Almaty would have followed parallel trends without the rise of Sci-Hub.

We conduct several tests on the validity of the identification strategy. First, we show that con-

[2]Sourve: https://about.jstor.org/librarians/journals/, accessed on 25th of January 2023.

nectedness is not associated with differential scientific trends in the ten years prior to the launch of Sci-Hub. Second, we run horse races with connectedness to major cities in neighboring countries of Khazakstan. We consistently find that Almaty is a strong predictor of Sci-Hub take-up, 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, whereas connectedness to other regions does not. Finally, throughout our analyses, we control for a host of covariates. In particular, we include 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 started referencing more paywalled papers. The estimates are quantitatively large. Doubling Sci-Hub traffic leads to a five percentage point increase in references to closed-access publications, where the pre-period mean is 68 percent. We also identify important heterogeneity. First, the effects are largest in fields where paywalls are more pronounced. Second, we show that the largest reference gains accrue to papers published most recently and in higher-ranked journals. Notably, we estimate decreases in references to low-quality journals. Affected researchers do not simply extend reference lists and cite more papers, instead, the quality of reference lists increases. Our interpretation is that Sci-Hub enables scientists to read and reference significantly more frontier research, which would not have been possible without Sci-Hub.

Do these differences in the consumption of frontier research translate into the production of higher-quality research? To answer this question, we focus on citations to regions with high versus low levels of Sci-Hub traffic. In this exercise, the empirical results are more nuanced. On average, we do not find a statistically significant effect of Sci-Hub on citations, a standard measure of scientific quality. However, this clouds substantial and expected heterogeneity. While researchers in highly developed countries appear to be prominent users of Sci-Hub, these scientists, in all likelihood, already had access to a large number of high-quality closed-access journals (e.g., through their libraries). On the other hand, scientists in the least developed regions might be unable to conduct high-quality research even with access due to a lack of other meaningful resources. Consistent with these arguments, we identify positive effects of Sci-Hub traffic on citations only in middle-income countries. Our interpretation is that the quality of papers in these regions has increased. However, we acknowledge that these novel citations may

also result from reciprocal referencing. At a minimum, the results imply greater recognition of work from previously disadvantaged regions.

Our results contribute to three distinct 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). In seminal work, McCabe and Snyder (2005) explore how open access relates to the size and quality of journals. In recent years, the literature has become more empirical and data-heavy, identifying determinants of more and better research activity. Different factors such as the role of peers (Waldinger, 2012), intellectual property rights (Williams, 2013; Murray et al., 2016), international cooperation (Iaria, Schwarz and Waldinger, 2018; Yin et al., 2021; Jia et al., 2022), income inequality (Agarwal and Gaule, 2020) and competition (Hill and Stein, 2021) have been identified as more and less important pillars of scientific advancement. Relative to these existing works, we focus on 'access' as a key pillar of knowledge creation.

While our study is 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. For a systematic review, see Langham-Putrow, Bakker and Riegelman (2021). 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. Surprisingly, while open access papers gain more views and downloads, they are not cited more often. McCabe and Snyder (2014) use a difference-in-differences design with journal-level variation and find increases in citations of approximately 8% when a journal moves from paid to open access. However, this analysis is confounded by plausible time-varying changes such as editorial overhaul. Finally, Bryan and Ozcan (2021) show that open access mandates imposed by the National Institutes of Health (NIH) significantly increased industry use of academic science. Our results advance the quality of existing evidence, both in scope and depth. Contrary to all existing work, we study a global natural experiment with long time horizons without restrictions to a specific scientific field.

The second literature we advance is centered around 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, 2020; Cagé et al., 2022), Facebook (Müller and Schwarz, 2021), VKontake (Enikolopov, Makarin and Petrova, 2020; Bursztyn et al., 2019) or Craigslist (Seamans and Zhu, 2014; Djourelova, Durante and Martin, 2021) 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.

Finally, we add to quickly growing work on the benefits of knowledge and technology transfer. Giorcelli (2019) shows how knowledge transfers in management practices raised firm performance, and Li and Giorcelli (2022) show the benefits of know-how transfer on plant output. Our study analyzes transfers of scientific insights and their effect on subsequent knowledge production.

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 qualitative scientific standards, curating and disseminating academic works—they do not internalize the benefits of offering free access. As a result, knowledge through openly accessible publications is likely an under-provided public good.

Inhibited by access restrictions, in 2011, a young graduate student from Almaty, Kazakhstan, founded 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 constituting 85% of all scholarly output, and in 2017 the platform had 500,000 daily visitors (Bohannon, 2016). In October 2022, the website counted almost 100 Mio downloads worldwide.[3]

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

The inception of Sci-Hub marked a fundamental shift in access to scholarly literature across the globe. 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. Appendix Section A and Figure A.1 document a step-by-step use case.

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 Virgina (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 websites through which Defendant Sci-Hub engages in unlawful practices."*

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 of 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, the founding place of Sci-Hub. 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 became 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.[4] First, the log files 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 network (VPN) is used. While VPN usage probably introduces noise in downloads, it is unlikely to invalidate our identification strategy and bias our results. First, VPNs were not as ominous 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 rely 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 level at which our instrument is observed. 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 5 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 Table 1).

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

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 MAGs coverage (Scheidsteger and Haunschild, 2022), which already, before its discontinuation, outperformed 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 publication-specific information is attached. Importantly, this includes the publication year, the host venue (in most cases, journals), and a list of referenced works.[5] 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, each representing an author and their affiliated institution at the time of publication. Based on the affiliation of authors and the geolocation of institutions[6], 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—this can be interpreted as knowledge consumption. Citations, on the

[5]OpenAlex provides several other pieces of information we do not utilize, e.g., the paper title or the paper abstract. A complete list of available characteristics can be found here.

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

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.[7]

Matching Open-access Status, Quality, and Field We corroborate each work with journal-specific metrics provided by Scopus's[8] yearly ranking of peer-reviewed journals. All journal measures retrieved through Scopus are fixed in 2011[9] to rule out that our results are driven by time trends in either 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.[10] 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 authors recorded in OpenAlex between 2008 and 2012. Moreover, we construct measures for the number of research institutions 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.

[7] Citations are (noisily) 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).

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

[9] 2011 is the earliest year for which Scopus journal metrics are available.

[10] This ignores for now that some journals have a mixed open-access policy 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.

Aggregation The final step aggregates publication, citation, and reference data across years and sub-national units. Panels A and B of Table 3 provide summary statistics on the number of research institutes and researchers in sub-national units. Panels C, D, and E of Table 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 and of which 56% 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,

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

measures the relative probability of a FB friendship between sub-national unit i and sub-national unit j.[11] 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 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.[12] Second, the Social Connectedness Index is unavailable for countries that restrict FB usage. Figure 6 Panels (b) and (c) give a spatial overview of raw and residualized connectedness

[11] 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.

[12] We discuss threats to identification in greater detail in Section 4.1.

between subnational regions and Almaty. Notably, there is no information on Russia, China, and Iran, among others.

Table 1 Panel B provides summary statistics of CON_i^j for Almaty, Nur-Sultan (the Kazakh capital), Kazakhstan[13], 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. 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 with zero observations can be sensitive to the scaling of the outcome if treatment affects the extensive margin (Chen and Roth, 2022; Mullahy and Norton, 2022).[14] 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. We provide two pieces of evidence to substantiate this argument. First, we empirically check if treatment

[13]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: $CON_i^j = \sum_{r_i} \sum_{r_j} \text{PopShare}_{r_i} \times \text{PopShare}_{r_j} \times CON_{r_i}^{r_j}$.

[14]In particular, Chen and Roth (2022) 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.

affects the extensive margin by regressing an indicator for generating any past research output on an indicator for having any Sci-Hub download in the observation period. Appendix Table A.2 shows Sci-Hub usage does not affect the probability of (first-time) entering the academic landscape, implying that treatment does not affect the extensive margin. Second, when disaggregating our results across regions in Figures 10 and 11, we find a null effect in developing countries where most regions have no or only minimal research output, giving credence to the claim that Sci-Hub mainly changes existing research patterns.

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, 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.

Kazakh graduate student Alexandra Elbakyan founded Sci-Hub in Almaty, from where it is run until today, with a small team of developers. As a result, researchers with pre-existing social ties to Almaty were more likely to be early adopters of Sci-Hub. Relying on path dependence in technology adoption (Arthur, 1989), early exposure to Sci-Hub continues to be a strong predictor of sub-national Sci-Hub usage today (akin to Müller and Schwarz, 2020; Enikolopov, Makarin and Petrova, 2020). 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 following first-stage equation:

$$\ln \text{Down}_{it} = \boldsymbol{\alpha} + \beta_1 \ln \text{CON}_i^{\text{Almaty}} \times \mathbb{1}_{t>2010} + \sum_n \delta_2^{(n)} \ln \text{CON}_i^{\text{n}} \times \mathbb{1}_{t>2010} + \boldsymbol{X}_{i2010}\boldsymbol{\gamma}_t + \varepsilon_{it} \quad \text{(IV1)}$$

where $\ln \text{Down}_{it}$ 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[15], each interacted with a post-2010 dummy. By this, 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 extensively controls for possible unobserved factors affecting both Sci-Hub downloads and social ties to Almaty. In particular, α_i captures time-invariant sub-national-specific factors. Furthermore, $\alpha_{c(i)t}$ accounts for time-varying country-specific factors like education reforms affecting all sub-national units $c(i)$ in year t. The function $c(i)$ maps sub-national units to countries. Finally, we control flexibly for several covariates[16] measured in 2010 interacted with year dummies. Unexplained variation enters the error ε_{it}, clustered at the country level.

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

$$\ln Y_{it} = \boldsymbol{\alpha} + \beta_2 \ln \widehat{\text{Down}}_{it} + \sum_n \delta_2^{(n)} \ln \text{CON}_i^{\text{n}} \times \mathbb{1}_{t>2010} + \boldsymbol{X}_{i2010}\boldsymbol{\gamma}_t + \eta_{it} \tag{IV2}$$

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 other variables are the same as in Equation (IV1).

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).

4.1 Threats to Identification

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.

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

[16] The list of control variables includes measure 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).

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? On average, only 15% of all journals make published articles available for free. Beyond this first data moment, Figure 1 shows large heterogeneity in open access regimes across two dimensions: field and journal quality. We find that open access is most prevalent in the life and health sciences and slightly less so in the physical and social sciences. Moreover, fairly consistent across fields, the number of open-access journals dwindles toward the top of the journal quality distribution. The key takeaway is that scientific knowledge is not only highly restricted across fields but these restrictions are particularly severe for knowledge residing in top 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 2 we show that this is not the case. We proxy for library access using institutional JSTOR subscriptions. JSTOR is an online library covering roughly 12 million items and access to about 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 in 2012 are largely unequally distributed across universities. While 30% of all institutions in developed regions have subscriptions, the fraction is reduced to roughly 10% in lesser developed regions.[17]

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, we show the distribution of downloaded papers by varying degrees of journal quality. Unsurprisingly, we find that articles

[17] In Appendix Figure A.5 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.

from top journals are in disproportionately high demand. 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 six times more 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 lesser developed regions of the world.

Finally, we turn to the production of scientific knowledge. In Figure 4, 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 likely shaped by a multitude of different factors. Yet, in the subsequent analyzes, we show that access plays a meaningful role in explaining the geography of scientific knowledge production.

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 take-up. The dynamic event study estimates are shown in Figure 7 Panel (a). According to the point estimates, connectedness is a strong and highly significant predictor. Note that by construction, we cannot estimate pre-trend coefficients because both the platform and downloads did not yet exist before 2011. Complementary, Figure 7 Panel (b) shows a binned scatterplot of the first stage correlation, again focusing on our most demanding specification. It illustrates the range of variation and provides evidence that the linear model provides a good

approximation of the data. The corresponding static estimates are presented in Table 4. 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 20. Once we control for connectedness to neighboring country capitals in column (4) and educational metrics in column (5), the coefficient remains consistent when introducing additional control variables. In Appendix Table A.1 we show that the first stage is not sensitive to applying the inverse hyperbolic sine transformation.

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 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 5. 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. Second, we re-estimate the first-stage equation by independently considering social ties to all other sub-national units. This exercise allows us to compare the estimate for Almaty with all other regions in our data. In Figure 8, 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 9 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. Taking the point estimates at face value, doubling a region's 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.[18] The static equivalents to the dynamic reduced form effects are displayed in Panel A of Table 6.[19] 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 12. 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 connectedness to other regions.

Returning to Figure 9, 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 detriment of references to open-access publications. Panel (d) indicates a slight drop of 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 and corresponds with average academic publication lags.

IV Combining our first stage and reduced form results, Panel B of Table 6 displays the 2SLS estimates on references for our most demanding specification. We find that doubling Sci-Hub 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. In Appendix Section B we discuss and show robustness to weak-IV considerations.

Heterogeneity by Field Fields differ in their prevalence of open versus closed-access journals. We test whether this variation moderates our Sci-Hub estimates. In particular, we disaggregate our data and re-run our 2SLS regression for different scientific fields. In Figure 14, we show that the increase in references to restricted access is particularly large in fields with higher restriction rates. Consequently, these are also the fields that experience the largest drop in open-access references.

[18] In Appendix Section C, we show robustness to potential linear and non-linear violations of the parallel trends assumption following Roth (2022) and Rambachan and Roth (2022).

[19] 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.

Heterogeneity by Quality and Age Next, we ask which exact types of works are being substituted as a result of Sci-Hub. We break down all references along two dimensions, the quality 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 in a connected region). Figure 13 plots the 2SLS estimates on different subsamples of references. The positive effect is highly concentrated among paywalled, high-quality journals (top two deciles) and articles published most recently (two to four years ago). On the other hand, Sci-Hub is associated with significant reductions in low-quality closed-access and low-quality open-access references. Interestingly, references to high-quality open-access publications remain unaffected. Hence, scientists appear not to unconsciously select restricted access publications as references but merely 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.

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), (c) and (e)). We find that increases in references to paywalled papers are driven by developed and emerging regions. Interestingly, the point estimates and dynamics are very similar in both regions, whereas absent in developing countries. In the latter, we see no measurable impact on the share of restricted access references.

5.3 Effects on Knowledge Production

The evidence gathered so far documents that Sci-Hub has profoundly impacted what researchers read and reference. We now ask 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 production of new scientific works.

Number of Publications First, we estimate equation IV2 using the number of newly written articles as the main outcome. The corresponding estimates are shown in Table 7. Interestingly, we don't find any effects of Sci-Hub on the number of new publications. We further test if there are distributional changes. If Sci-Hub led to better papers published in better journals, but not more, we would expect increases in higher-ranked publications at the expense of lower-ranked publications. In columns (3) to (7) we estimate the effect of Sci-Hub intensity on publications

across different journal-rank quintiles. Again, we do not find any evidence for distributional changes. If anything, we estimate slight reductions in newly written articles across the full journal-ranking spectrum.

Citations Next, we turn to a standard quality measure of scientific output: citations. In Table 8 we present results to estimates of equation IV2 using the number of citations accruing to researchers in a given region as the main outcome. If access to frontier research leads to higher quality works, we would expect increases in citations to regions with higher Sci-Hub traffic. While we find positive estimates—doubling Sci-Hub traffic leads to 6.3% more citations from peer-reviewed journals—we lack sufficient power to reject the null hypothesis of no effect.

However, this non-significant increase in citations clouds important heterogeneity. A more nuanced picture emerges when we split 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 results are presented in Table 9 and the right-hand panels of Figure 10. Allowing for heterogeneous effects, we find positive and significant increases in citations concentrated in middle-income countries following 2011. We interpret this as Sci-Hub only having tangible quality effects in regions where access restrictions were previously a binding constraint. Researchers in richer regions typically have access to at least some high-quality journals through their universities, while low-income regions in all likelihood lack other capacities to utilize Sci-Hub availability.

Results to the heterogeneity exercise are presented in Table 9. We estimate quantitatively large effects. Doubling Sci-Hub intensity in emerging regions is associated with approximately 15% more citations from peer-reviewed journals. The size is particularly striking when we compare our estimates to the existing literature. On an individual researcher level, Jacob and Lefgren (2011) find that an NIH grant (worth approximately $1.7 million) is associated with citation increases of 7% per grant recipient. Jia et al. (2022) find 7% citation decreases due to NIH investigations into US-Cina collaborations. Our preferred interpretation is that greater access has led to large quality increases in research by scientists in middle-income countries. However, we acknowledge that our evidence is suggestive. An equally plausible mechanism could be that more references to previously closed-access papers are now met with reciprocal citations. This would imply that Sci-Hub is leading to greater recognition of work from previously disadvantaged regions.

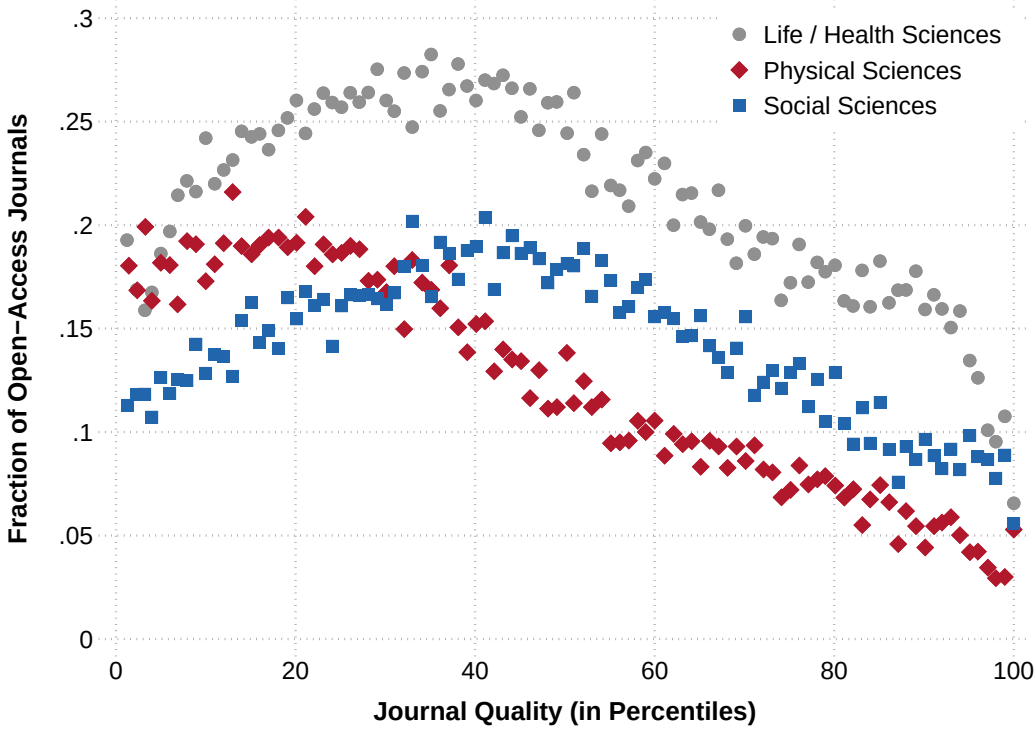
6 Conclusion

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 two 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 closed-access references and low-quality open-access references with closed-access articles at the research frontier. Second, while we do not find that Sci-Hub had a statistically significant effect on higher-quality publications on average, we document significant increases in citations to articles from middle-income countries. Our results suggest that open-accessible research is likely an underprovided public good, specifically in emerging regions. With a slowdown in disruptive science afoot (Park, Leahey and Funk, 2023), the policy takeaway is clear: Governments should actively implement measures to reduce closed-access rates.

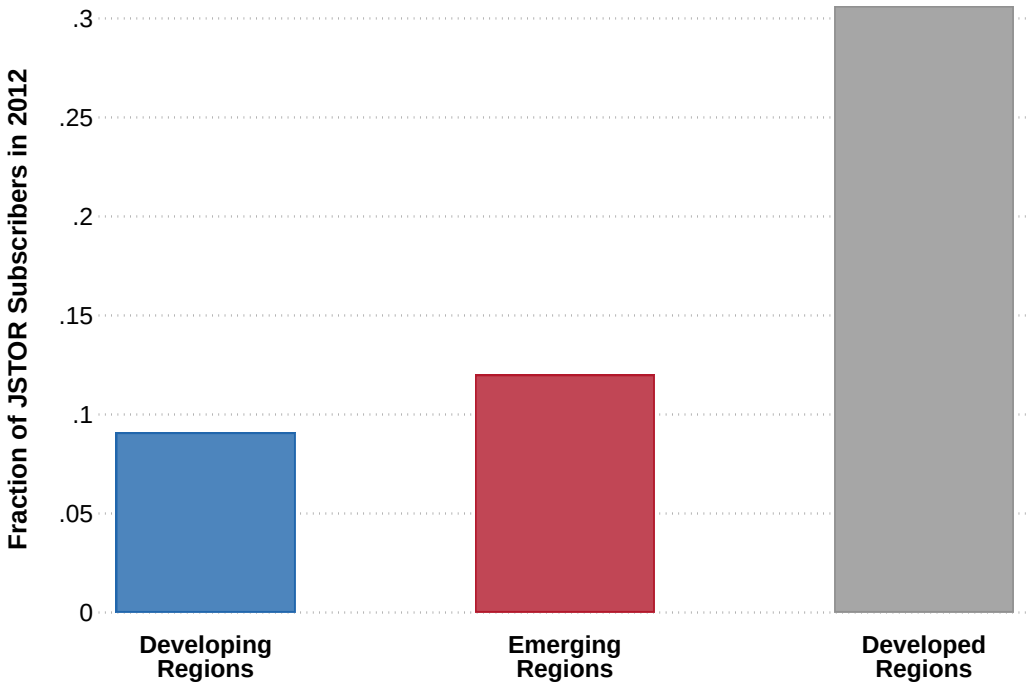
7 Figures

Figure 1: Fraction of Open-Access Journal by Journal Quality across Fields



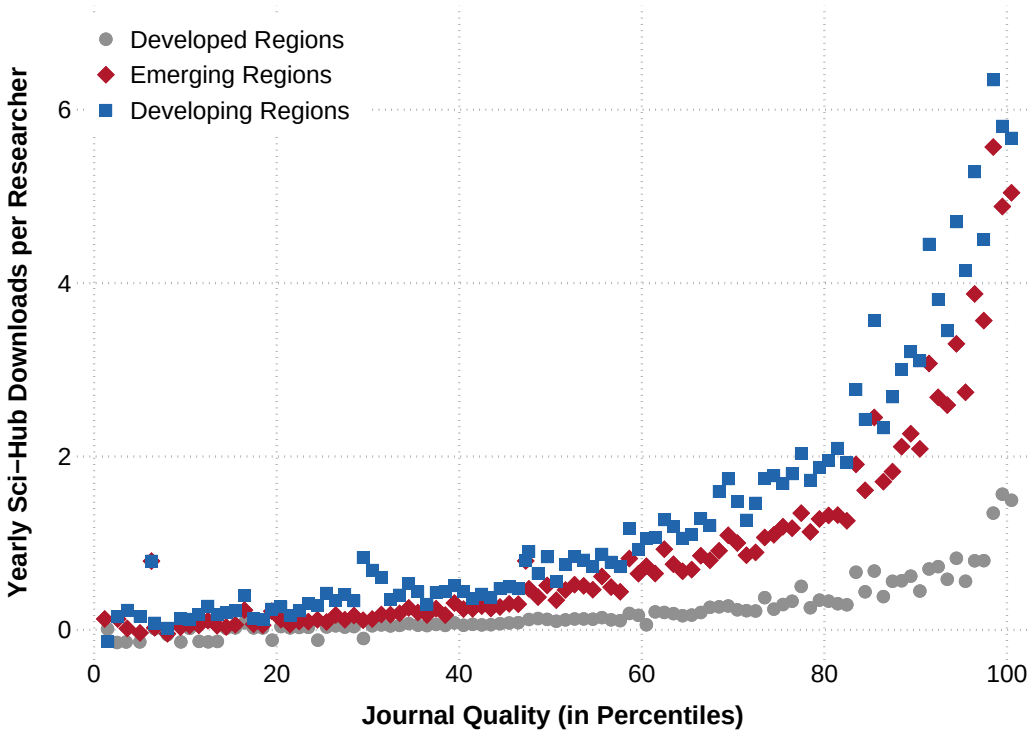
Note: The figure depicts the fraction of open-access journals by quality across fields. For all journals, open-access status and journal quality are based on measures provided by Scopus. In particular, journals are declared as open-access status if the journal is listed in the Directory of Open Access 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.

Figure 2: Fraction of JSTOR Subscribers by Region in 2012



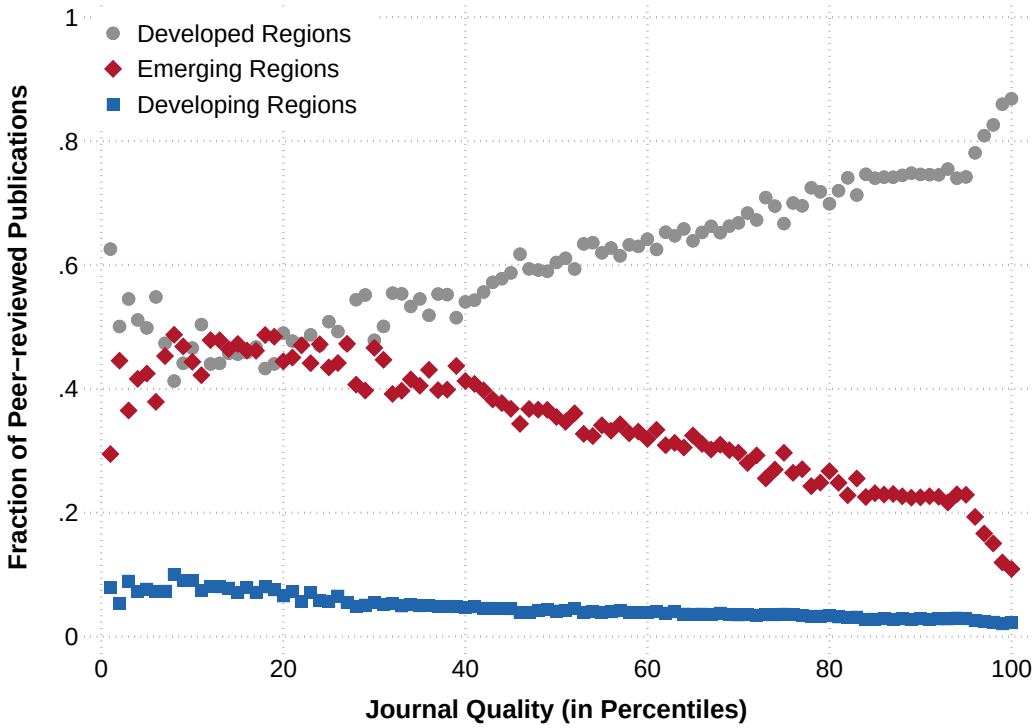
Note: The figure depicts the mean fraction of JSTOR subscribers per research institute across regions in 2012.

Figure 3: Average Yearly Sci-Hub Downloads per Researcher by Journal Quality and Region



Note: The figure shows the average annual Sci-Hub downloads per researcher by journal quality in the different regions. The figure includes all peer-reviewed scientific papers recorded in Sci-Hub log files from 2011 to 2013 and 2015 to 2017. Journal quality percentiles are based on the average number of citations from peer-reviewed articles per publication in Scopus. Classification of sub-national units into developed, emerging, and developing regions follows WorldBank's Atlas method, which classifies economies based on their gross national income per capita (WB, 2022).

Figure 4: Fraction of Peer-reviewed Publications by Journal Quality across Regions

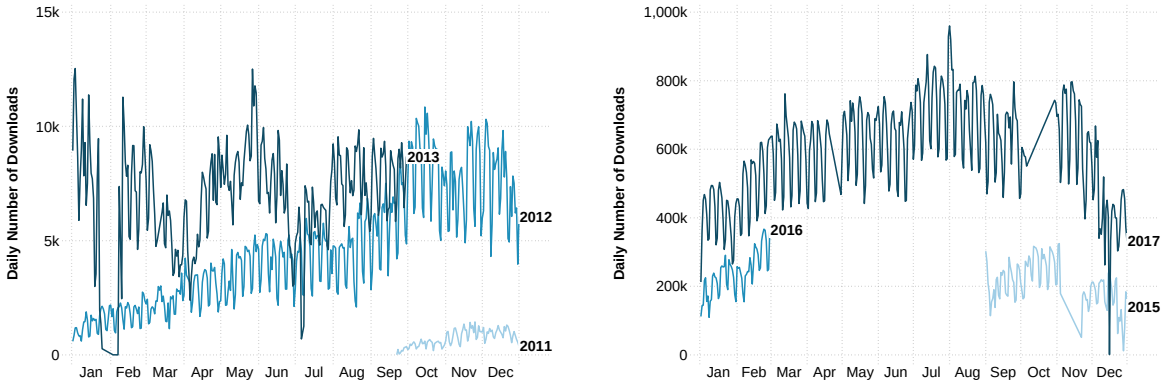


Note: The figure depicts the fraction of peer-reviewed publications by journal quality across regions. The figure includes all publications between 2000 and 2022 that are recorded in OpenAlex and are assigned to a journal. For all journals, peer-review status and journal quality are based on measures provided by Scopus. Journal quality percentiles are based on the average number of citations from peer-reviewed articles per publication. Classification of sub-national units into developed, emerging, and developing regions follows WorldBank's Atlas method, which classifies economies based on their gross national income per capita (WB, 2022).

Figure 5: Sci-Hub Downloads over Time

(a) 2011-2013

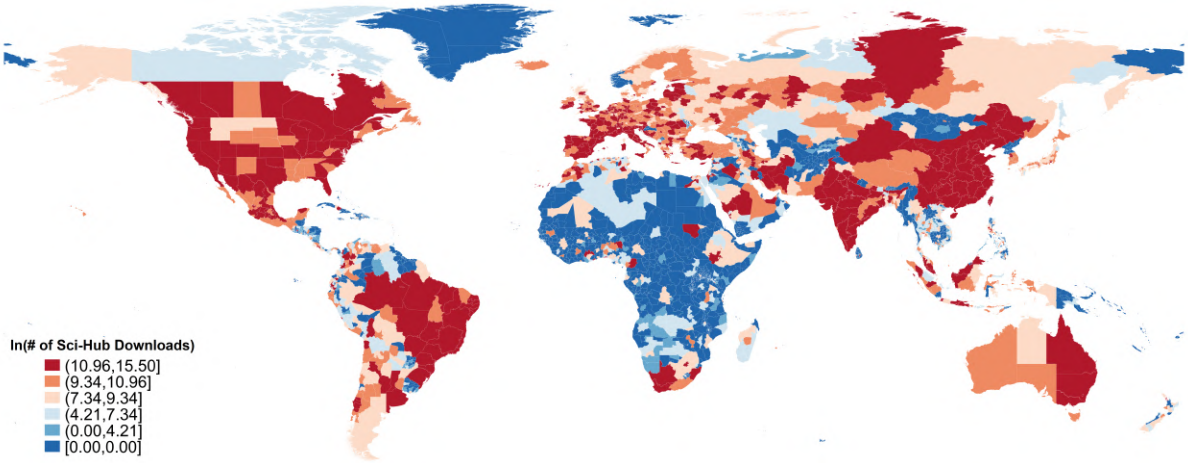
(b) 2015-2017



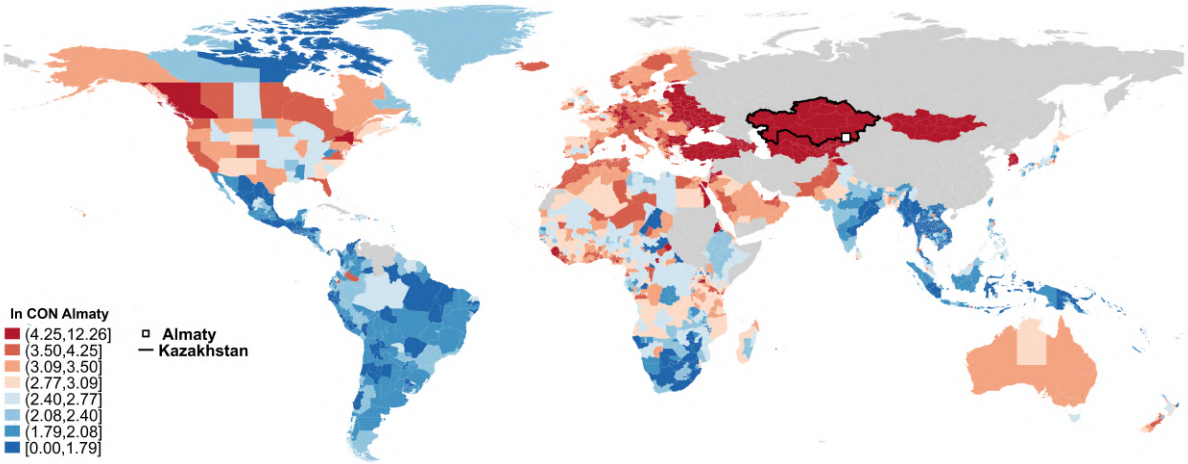
Note: The figure shows the average daily Sci-Hub downloads by year. The figure includes all downloads recorded in Sci-Hub log files from 2011 to 2013 and 2015 to 2017.

Figure 6: Descriptive by Sub-national Units

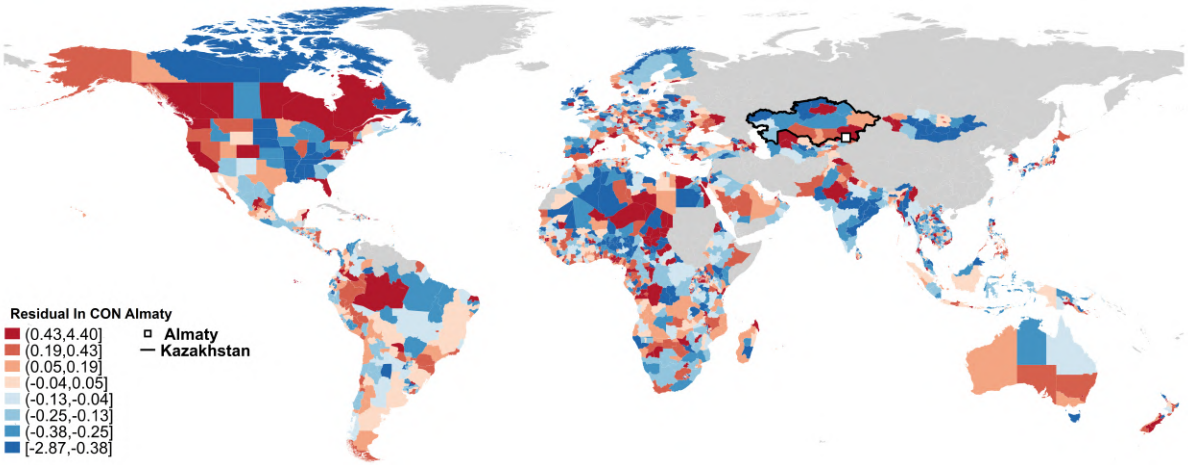
(a) Sci-Hub Downloads



(b) Social Ties to Almaty



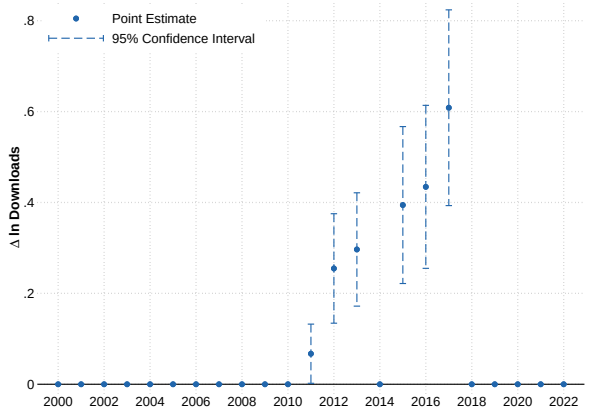
(c) Residualized Social Ties to Almaty



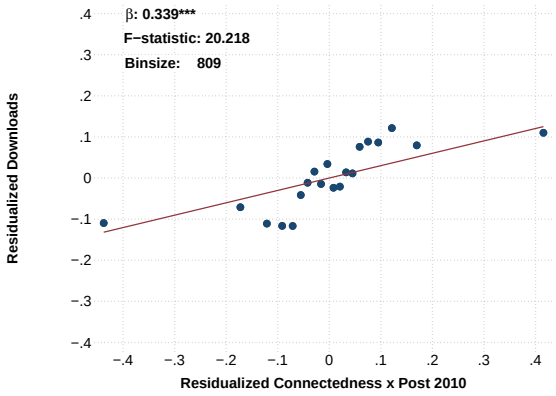
Note: Figure 6a shows the spatial distribution of Sci-Hub downloads across sub-national units. Figure 6b depicts how social ties to Almaty vary across sub-national units. The borders of Kazakhstan are marked by a black line. The location of Almaty is marked by the white square outlined in black.

Figure 7: First Stage – Visual Evidence

(a) Event Study

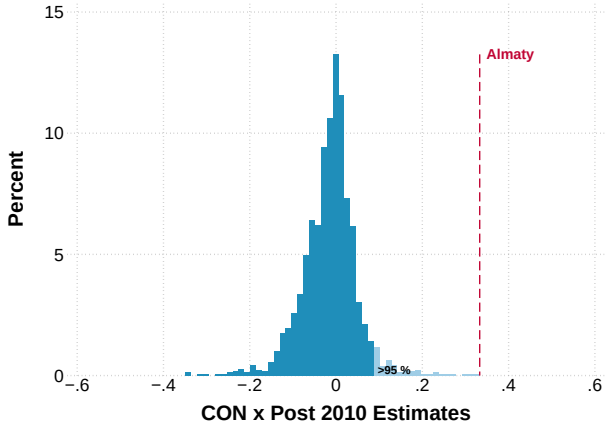


(b) Binned Scatterplot

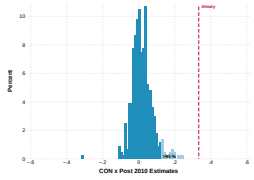


Note: Panel (a) shows point estimates and confidence intervals of the dynamic effects corresponding to the specification in Table 4 Panel A column (8). Panel (b) plots the residuals and coefficient estimate of the corresponding static difference-in-differences model.

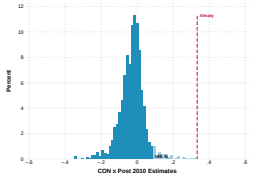
Figure 8: First Stage – Placebo Estimates



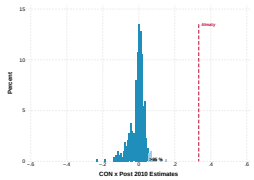
(a) All Regions



(b) Developed



(c) Emerging

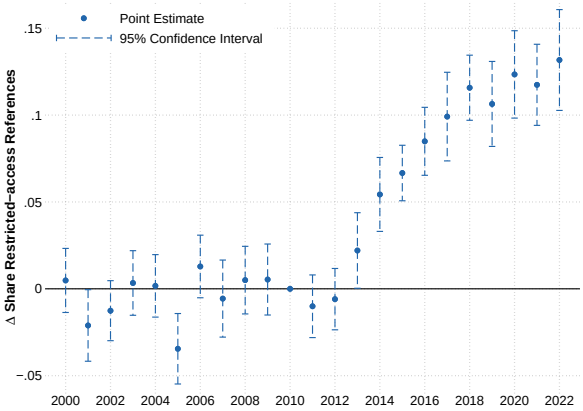


(d) Developing

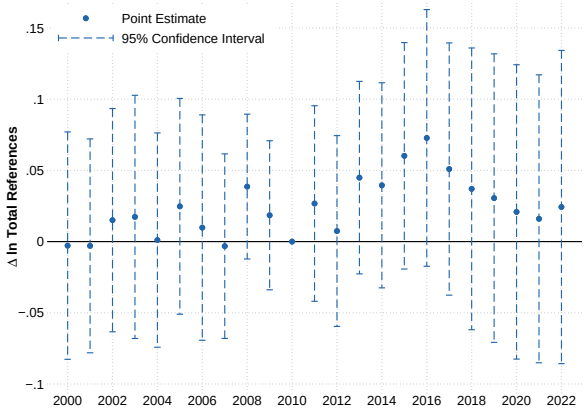
Note: Panel (a) shows the distribution of point-estimates when re-estimating Equation IV1 by iteratively replacing social connectedness to Almaty with social ties to all other sub-national units. Panels (b) to (d) show the distribution of point-estimates within specific regions. Classification of sub-national units into developed, emerging, and developing regions follows WorldBank's Atlas method, which classifies economies based on their gross national income per capita (WB, 2022). In all figures, the dotted red line corresponds to the point estimate in Panel A column 8 of Table 4.

Figure 9: Reduced Form Event Studies

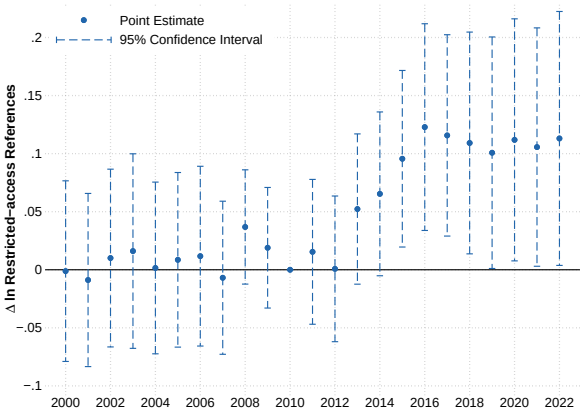
(a) Share Restricted-access



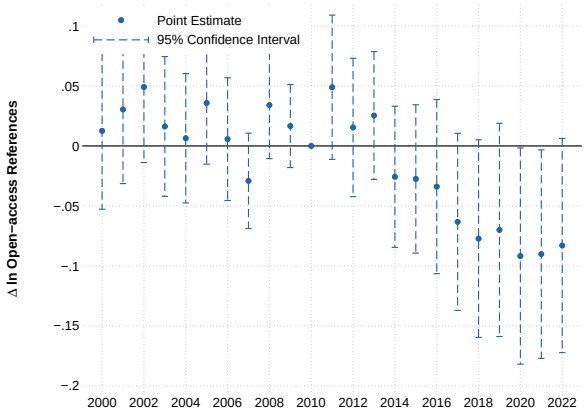
(b) Total



(c) Restricted-access



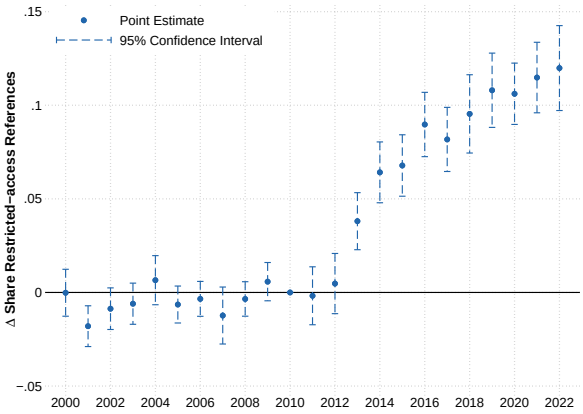
(d) Open-access



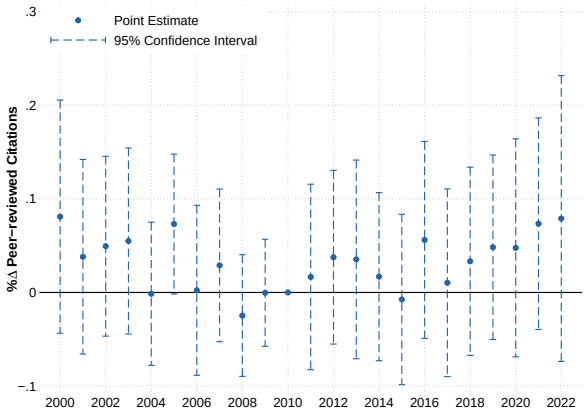
Note: The figure shows reduced form event study estimates with corresponding 95% confidence intervals for the outcomes and specification displayed in Table 6 Panel A. The post-2010 indicator in equation IV1 is replaced with a full set of annual indicators, omitting 2010, the year before Sci-Hub was established.

Figure 10: Reduced Form Event Studies by Region (1)

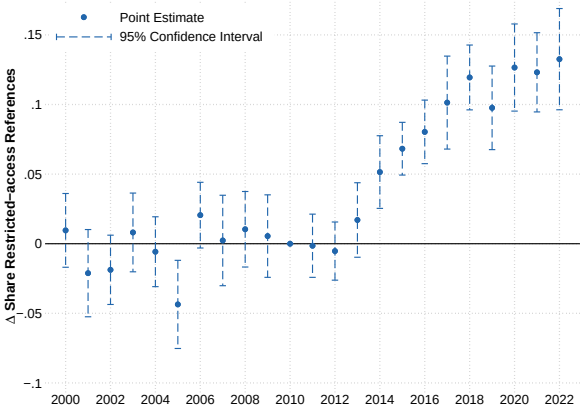
(a) Share Restricted-access (Developed)



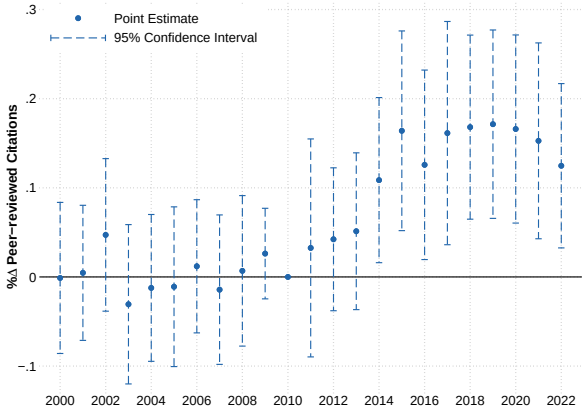
(b) Citations (Developed)



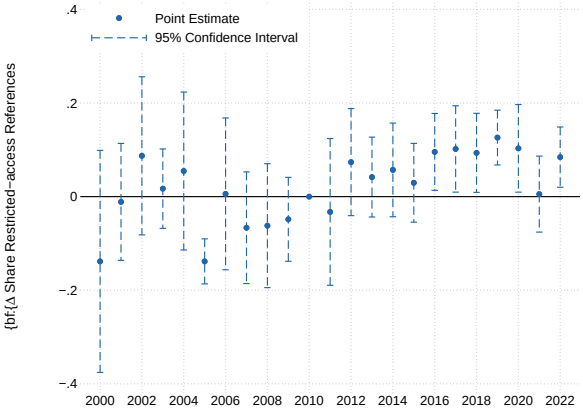
(c) Share Restricted-access (Emerging)



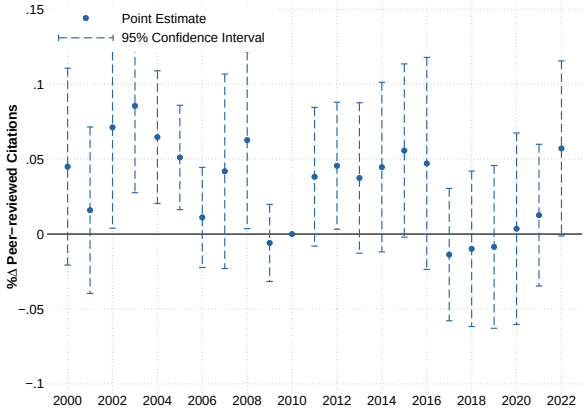
(d) Citations (Emerging)



(e) Share Restricted-access (Developing)

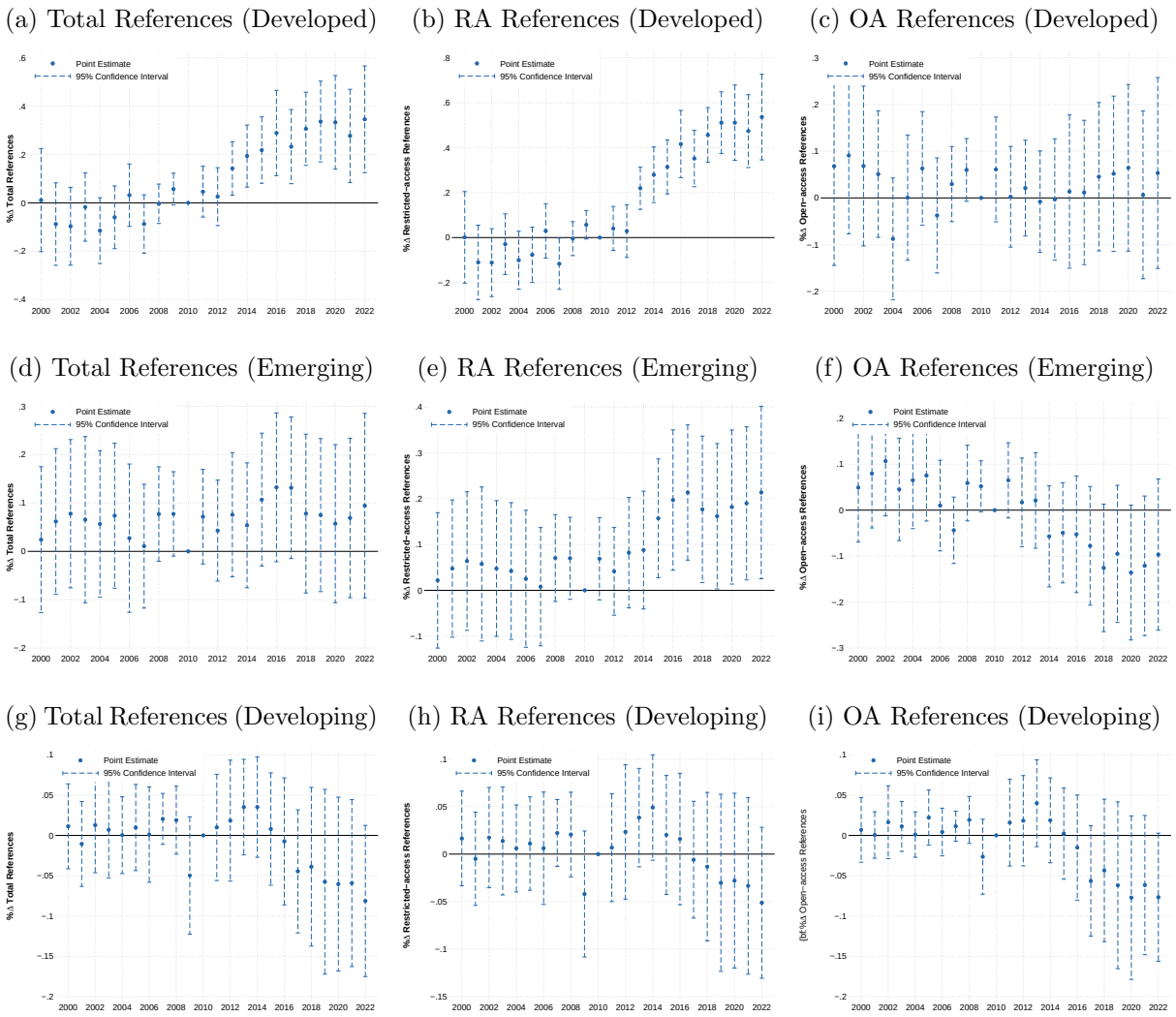


(f) Citations (Developing)



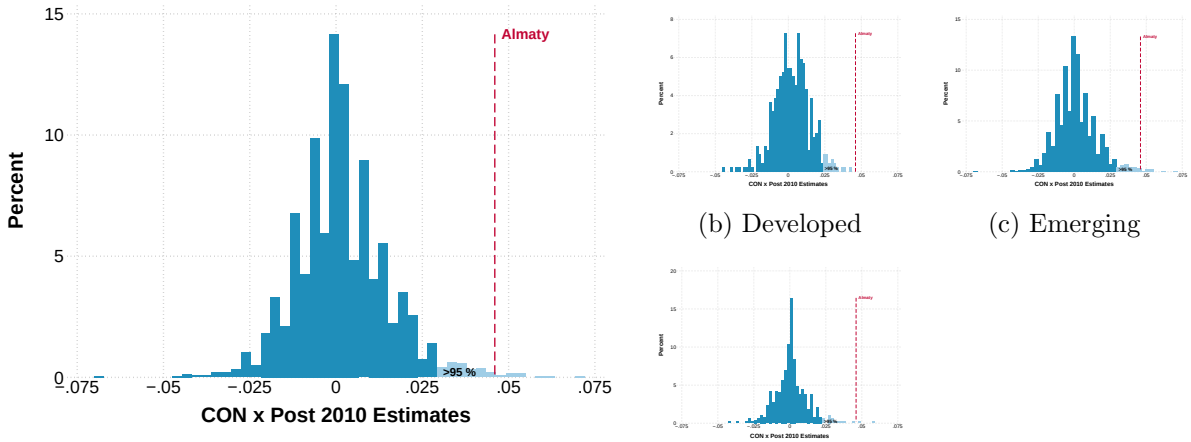
Note: The figure shows reduced form event study estimates with corresponding 95% confidence intervals for the outcomes and specification displayed in Table 6 Panel A. The post-2010 indicator in equation IV1 is replaced with a full set of annual indicators, omitting 2010, the year before Sci-Hub was established.

Figure 11: Reduced Form Event Studies by Region (2)



Note: The figure shows reduced form event study estimates with corresponding 95% confidence intervals for the outcomes and specification displayed in Table 6 Panel A. The post-2010 indicator in equation IV1 is replaced with a full set of annual indicators, omitting 2010, the year before Sci-Hub was established.

Figure 12: Reduced Form – Placebo Estimates

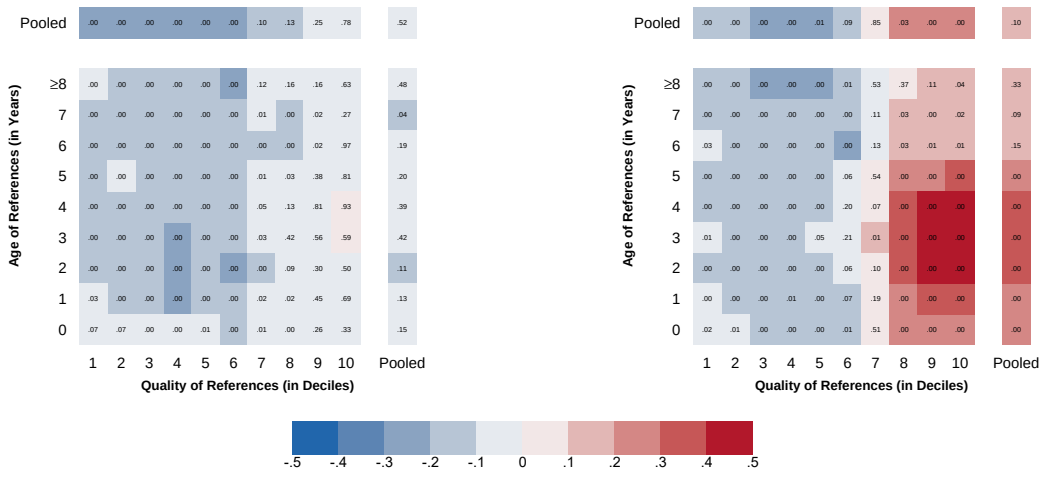


(a) All Regions (b) Developed (c) Emerging (d) Developing

Note: Figure 12a shows the distribution of point-estimates when re-estimating Equation IV1 by iteratively replacing social connectedness to Almaty with social ties to all sub-national units. Figures 12b to 12d show the distribution of point-estimate within specific regions. Classification of sub-national units into developed, emerging, and developing regions follows WorldBank's Atlas method, which classifies economies based on their gross national income per capita (WB, 2022). In all figures, the dotted red line corresponds to the point estimate in Panel B column 8 of Table 4.

Figure 13: Change in Reference Dynamics by Age and Quality

(a) Open-access (b) Restricted-access

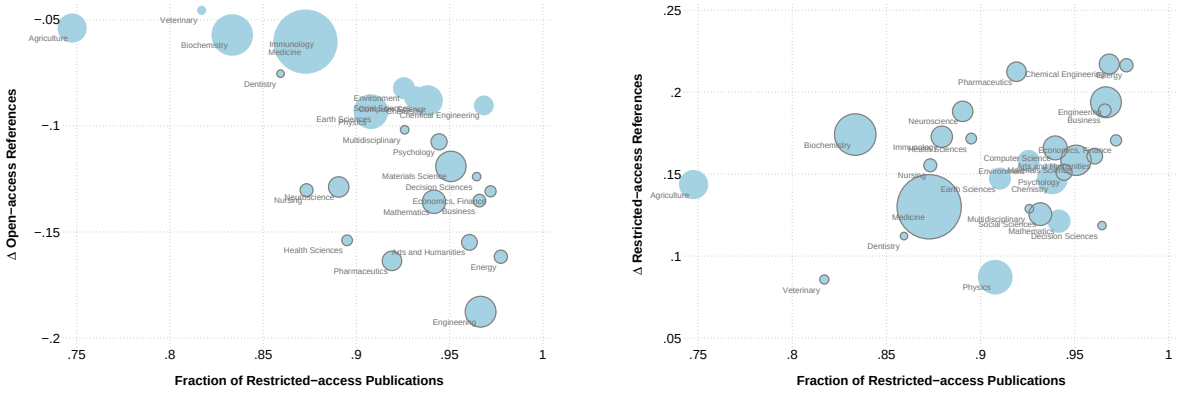


Note: The figure shows disaggregated 2SLS estimates for the number of open-access and restricted-access references according to the specification in Panel B of table 6. Specifically, the number of references is disaggregated by age and quality of the referenced papers. The age corresponds to the year 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.

Figure 14: Change in Reference Dynamics by 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 6. 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. A grey outline indicates that the estimate is significant at 5%.

8 Tables

Table 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 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 shows 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 are sub-national from 2000 to 2022.

Table 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 4: First Stage Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: ln Downloads								
ln CON Almaty × Post 2010	0.617***	0.647***	0.747***	0.454***	0.290***	0.299***	0.340***	0.339***
	(0.072)	(0.068)	(0.124)	(0.096)	(0.069)	(0.071)	(0.076)	(0.075)
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	74.401	90.766	36.214	22.417	17.534	17.763	19.819	20.218
Panel B: ln Cumulative Downloads								
ln CON Almaty × Post 2010	0.883***	0.980***	0.937***	0.579***	0.396***	0.403***	0.455***	0.452***
	(0.105)	(0.095)	(0.149)	(0.118)	(0.087)	(0.089)	(0.097)	(0.094)
Observations	56,051	56,051	54,832	54,832	54,832	54,832	54,832	54,832
Number of Clusters	195	195	142	142	142	142	142	142
F-statistic	70.221	105.564	39.768	24.127	20.534	20.597	22.157	23.161
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. Standard errors are clustered at the country level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 5: First Stage Estimates – Horse Race

	Dependent Variable: ln Downloads					
	(1)	(2)	(3)	(4)	(5)	(6)
ln CON Almaty × Post 2010	0.273***	0.290***	0.320***	0.308***	0.339***	0.318***
	(0.064)	(0.078)	(0.072)	(0.072)	(0.075)	(0.080)
ln CON KAZ excl. Almaty × Post 2010	–	-0.035	–	–	–	–
	(–)	(0.073)	(–)	(–)	(–)	(–)
ln CON Nur-Sultan × Post 2010	–	–	-0.072*	–	-0.065	-0.067
	(–)	(–)	(0.043)	(–)	(0.044)	(0.043)
ln CON Bishkek × Post 2010	–	–	–	-0.069*	-0.062*	-0.067*
	(–)	(–)	(–)	(0.037)	(0.037)	(0.038)
ln CON Ashgabat × Post 2010	–	–	–	-0.020	-0.018	-0.019
	(–)	(–)	(–)	(0.015)	(0.015)	(0.015)
ln CON Tashkent × Post 2010	–	–	–	0.029	0.044	0.028
	(–)	(–)	(–)	(0.046)	(0.049)	(0.054)
ln CON Dushanbe × Post 2010	–	–	–	–	–	0.003
	(–)	(–)	(–)	(–)	(–)	(0.036)
ln CON Ulaanbaatar × Post 2010	–	–	–	–	–	0.054
	(–)	(–)	(–)	(–)	(–)	(0.042)
ln CON Kyiv × Post 2010	–	–	–	–	–	0.017
	(–)	(–)	(–)	(–)	(–)	(0.057)
Observations	40,444	40,444	40,444	40,444	40,444	40,444
F-statistic	18.342	13.698	19.900	18.522	20.218	15.607
Fixed Effects						
Sub-national	✓	✓	✓	✓	✓	✓
Year × Country	✓	✓	✓	✓	✓	✓
CON Neighboring Capitals	✓	✓	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓	✓	✓

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

Table 6: 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	-0.028	0.066	0.046***
	(0.042)	(0.031)	(0.042)	(0.009)
Observations	40,528	40,528	40,528	19,481
Panel B: 2SLS				
ln Downloads	0.108	-0.085	0.191	0.046***
	(0.123)	(0.093)	(0.121)	(0.009)
Observations	40,444	40,444	40,444	19,410
F-statistic	20.436	20.436	20.436	20.436
Panel C: OLS				
ln Downloads	-0.018*	-0.016*	-0.014	0.001**
	(0.010)	(0.009)	(0.010)	(0.001)
Observations	40,444	40,444	40,444	19,410
Fixed Effects				
Sub-national	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) across various specifications. Standard errors are clustered at the country level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 7: 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	-0.008	-0.003	-0.011	-0.012	-0.008	-0.005	-0.009
	(0.027)	(0.025)	(0.013)	(0.015)	(0.019)	(0.020)	(0.021)
Observations	40,528	40,528	40,528	40,528	40,528	40,528	40,528
Panel B: 2SLS							
ln Downloads	-0.026	-0.009	-0.032	-0.032	-0.026	-0.012	-0.027
	(0.081)	(0.074)	(0.038)	(0.045)	(0.056)	(0.059)	(0.064)
Observations	40,444	40,444	40,444	40,444	40,444	40,444	40,444
F-statistic	20.436	20.436	20.436	20.436	20.436	20.436	20.436
Panel C: OLS							
ln Downloads	-0.008	-0.006	0.017***	0.010	0.003	-0.004	0.000
	(0.007)	(0.007)	(0.005)	(0.007)	(0.006)	(0.006)	(0.006)
Observations	40,444	40,444	40,444	40,444	40,444	40,444	40,444
Fixed Effects							
Sub-national	✓	✓	✓	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) across various specifications. Standard errors are clustered at the country level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 8: Change in Citation Patterns

	Number of Citations			
	Total	Non-peer-reviewed	Peer-reviewed	Cross-field
	(1)	(2)	(3)	(4)
Panel A: Reduced Form				
ln CON Almaty × Post 2010	0.036	0.005	0.040	-0.033
	(0.028)	(0.025)	(0.028)	(0.040)
Observations	40,528	40,528	40,528	40,528
Panel B: 2SLS				
ln Downloads	0.115	0.019	0.129	-0.085
	(0.086)	(0.074)	(0.087)	(0.116)
Observations	40,444	40,444	40,444	40,444
F-statistic	20.436	20.436	20.436	20.436
Panel C: OLS				
ln Downloads	-0.007	0.001	-0.006	-0.011
	(0.007)	(0.007)	(0.007)	(0.012)
Observations	40,444	40,444	40,444	40,444
Fixed Effects				
Sub-national	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) across various specifications. Standard errors are clustered at the country level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table 9: 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	-0.007	-0.040	-0.007	0.143
	(0.058)	(0.052)	(0.058)	(0.090)
ln CON Almaty × Post 2010 × Emerging	0.096	0.103	0.149**	-0.036
	(0.078)	(0.098)	(0.074)	(0.123)
ln CON Almaty × Post 2010 × Developing	0.027	-0.012	0.032	-0.052
	(0.029)	(0.022)	(0.030)	(0.043)
Observations	40,528	40,528	40,528	40,528
Panel B: 2SLS				
ln Downloads × Developed	0.053	-0.022	0.060	0.056
	(0.077)	(0.063)	(0.078)	(0.093)
ln Downloads × Emerging	0.145*	0.105	0.157*	-0.053
	(0.086)	(0.096)	(0.085)	(0.140)
ln Downloads × Developing	0.127	-0.029	0.147	-0.199
	(0.125)	(0.093)	(0.127)	(0.158)
Observations	40,444	40,444	40,444	40,444
F-statistic	18.797	18.797	18.797	18.797
Panel C: OLS				
ln Downloads × Developed	-0.014	-0.018*	-0.015	-0.024
	(0.012)	(0.010)	(0.012)	(0.021)
ln Downloads × Emerging	-0.013	-0.007	-0.012	-0.017
	(0.016)	(0.016)	(0.016)	(0.026)
ln Downloads × Developing	0.000	0.015	0.002	-0.001
	(0.011)	(0.011)	(0.011)	(0.018)
Observations	40,444	40,444	40,444	40,444
Fixed Effects				
Sub-national	✓	✓	✓	✓
Country × Year	✓	✓	✓	✓
CON Neighb. Capitals	✓	✓	✓	✓
Controls in 2010 × Year FE	✓	✓	✓	✓

Note: The table displays regression results from Equation (IV2) across various specifications. Standard errors are clustered at the country level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

References

Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. 2015. "Radio and the Rise of the Nazis in Prewar Germany." *The 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.

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 IVe." *NBER Working Paper*, 29417.

Archambault, Éric, Didier Amyot, Philippe Deschamps, Aurore Nicol, Françoise Provencher, Lise Rebout, and Guillaume Roberge. 2014. "Proportion of Open Access Papers Published in Peer-reviewed Journals at the European and World Levels - 1996–2013." In *Study to Develop a Set of Indicators to Measure Open Access*. Vol. RTD-B6-PP-2011-2. European Commission.

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.

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." *The 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. 2022. "Log-like? Identified ATEs Defined with Zero-valued Outcomes are (arbitrarily) Scale-dependent." *Working Paper*.

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." *The Quarterly Journal of Economics*, 122(3): 1187–1234.

Djourelova, Milena, Ruben Durante, and Gregory Martin. 2021. "The impact of online competition on local newspapers: Evidence from the introduction of Craigslist."

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

Dustmann, Christian, Attila Lindner, Uta Schönberg, Matthias Umkehrer, and

Philipp Vom Berge. 2022. "Reallocation effects of the minimum wage." *The Quarterly Journal of Economics*, 137(1): 267–328.

Eastern District Court of Virgina, 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.

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." *The 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.

Giorcelli, Michela. 2019. "The long-term effects of management and technology transfers." *American Economic Review*, 109(1): 121–52.

Guriev, Sergei, Nikita Melnikov, and Ekaterina Zhuravskaya. 2021. "3g internet and confidence in government." *The Quarterly Journal of Economics*, 136(4): 2533–2613.

Hill, Ryan, and Carolyn Stein. 2021. "Race to the bottom: Competition and quality in science." Working Paper.

Iaria, Alessandro, Carlo Schwarz, and Fabian Waldinger. 2018. "Frontier knowledge and scientific production: evidence from the collapse of international science." *The Quarterly*

Journal of Economics, 133(2): 927–991.

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.

Jacob, Brian A, and Lars Lefgren. 2011. "The impact of research grant funding on scientific productivity." *Journal of public economics*, 95(9-10): 1168–1177.

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. 2022. "The Impact of US-China Tensions on US Science." National Bureau of Economic Research.

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. "JUE Insight: 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, Bo, and Michela Giorcelli. 2022. "Technology Transfer and Early Industrial Development: Evidence from the Sino-Soviet Alliance."

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,Äô COCI: a multidisciplinary comparison of coverage via citations." *Scientometrics*, 126(1): 871–906.

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.

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. 2020. "From Hashtag to Hate Crime: Twitter and Anti-minority Sentiment." *SSRN Working Paper*, 3149103.

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.

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 ‚Äì 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.

Rambachan, Ashesh, and Jonathan Roth. 2022. "A More Credible Approach to Parallel Trends." Working Paper.

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

Roth, Jonathan. 2022. "Pretest with caution: Event-study estimates after testing for parallel trends." *American Economic Review: Insights*, 4(3): 305–22.

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 preprint arXiv: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.

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

Stoy, Lennart, Morais 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." *The 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." *The review of economic studies*, 79(2): 838–861.

WB, The World Bank Group. 2022. "The World Bank Atlas Method – Detailed Methodology." URL: *https://datahelpdesk.worldbank.org/knowledgebase/articles/378832-the-world-bank-atlas-method-detailed-methodology*, Accessed: 2022-11-29.

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." *The Quarterly Journal of Economics*, 129(4): 1947–1994.

Yin, Yian, Yuxiao Dong, Kuansan Wang, Dashun Wang, and Benjamin Jones. 2021. "Science as a Public Good: Public Use and Funding of Science." *NBER Working Paper*, 28748.

Appendices

A Using Sci-Hub

In Appendix Figure A.1 we present an exemplary use case of Sci-Hub. Panel (a) shows the Sci-Hub home page as of November 2022, which users can find easily through a web search. To request a paper, one can either provide the paper title or the unique digital object identifier (DOI)[20] of a paper, see Panel (b). After confirming the request by clicking 'open', Sci-Hub checks if the paper is stored in its database. If the requested paper is available, it is displayed for download as shown in Panel (c). Next, the user can download the paper. Sci-Hub records all resolved requests, i.e., successful downloads, in so-called log files. Unresolved requests are not saved, i.e., papers that are opened but not downloaded.

B 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, given by the correlation between structural and first-stage residuals. Within this framework, a high degree of endogeneity calls for a strong instrument, i.e., a high first-stage F-statistics. 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.3 shows that our specification exhibits moderate to high levels of endogeneity, exceeding the threshold of $|\rho| > .565$ when considering our main specification. Complementing the bounding exercise on ρ, Table A.3 next reports p-values of the Anderson-Rubin F-test (Anderson and Rubin, 1949) as well as tF-adjusted standard errors (Lee et al., 2022).[21] Under both procedures, our results stay significant at the

[20] Since the turn of the millennium, most scholarly work can be identified through a unique DOI. In academia "Crossref" functions as the official DOI registration agency. The project started in the early 2000s as a cooperative effort among publishers to enable persistent cross-publisher citation linking of academic work. Notably, DOIs can be registered retroactively.

[21] The procedure by Anderson-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

1-percent level even when considering a worst-case endogeneity scenario of $|\rho| = 1$ as assumed when computing tF-adjusted standard errors.

C Parallel Trends

Potential Linear Pre-Trends In our main event study with the share of restricted access citations as outcome, visual inspection of pre-period coefficients suggests that regions with high and low connectedness are generally on similar outcome trajectories prior to the launch of Sci-Hub. However, the pre-period coefficient for 2005 pinpoints a temporary significant difference. To investigate potential linear violations of the parallel trends assumption more systematically we follow the procedure outlined in Roth (2022).

We ask, for what linear trend would we identify at least one statistically significant coefficient in the pre-period 80% of the time. We find that a slope of 0.2 percentage points satisfies these criteria. Is such a linear violation of parallel trends quantitatively meaningful? We depict the potential bias in Figure A.6. For 2022, our last year of data, we measure that 100% more friendship links to Almaty are associated with a 13.4 percentage point increase in references to closed-access articles. Yet, the bias from the linear trend may only account for 3.4 percentage points (conditional on passing the pre-test). Hence, we conclude that bias from reasonable linear violations of parallel trends is not sufficient to explain our difference-in-differences estimates.

Potential Non-Linear Parallel Trend Violations Next, we assess potential non-linear violations of parallel trends. We follow the approach in Rambachan and Roth (2022). Intuitively, their procedure assumes that parallel trend violations in the post-period cannot be much larger than violations in the pre-period. More formally, they propose to bound possible changes in the slope across two subsequent periods by some parameter M following:

$$\Delta^{SD}(M) := \{\delta : |(\delta_{t+1} - \delta_t) - (\delta_t - \delta_{t-1})| \leq M, \forall t\} \tag{A.1}$$

This allows the construction of confidence intervals valid under partial identification. Setting a reasonable value for M, we adopt the procedure in Dustmann et al. (2022). First, we take the estimated pre-trend coefficients and the associated variance-covariance matrix. Under the assumption of jointly normally distributed errors, we simulate the average absolute trend devi-

the first-stage F-statistic and the considered significance level. Both procedures yield correct coverage under arbitrarily weak instruments; however, the expected length of the Anderson-Rubin confidence interval is infinite, while the corresponding tF interval is finite (Lee et al., 2022).

ation in the pre-period. We then set M at the median of the average trend deviations. This procedure yields a value of $M^* = 0.014$.

In Figure A.7 we show associated confidence sets for the average treatment effect in the post-period at different values of M. Our result is robust to different values of M, even allowing for a trend violation that is more than ten times the size of M^*.

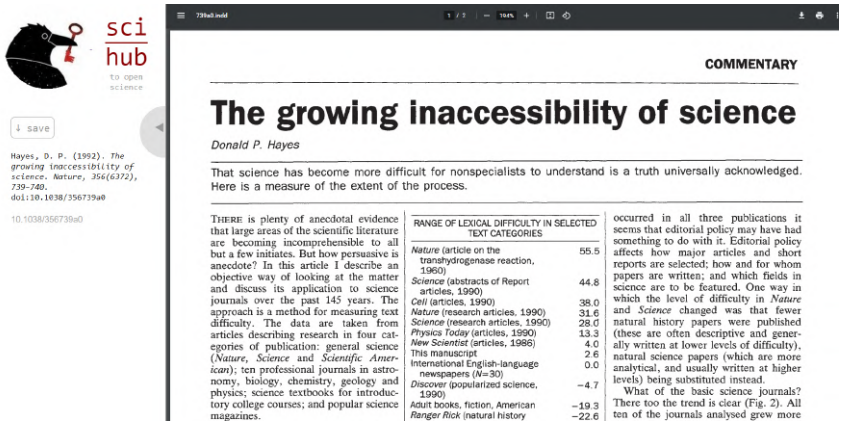
D Additional Figures

Figure A.1: Sci-Hub Usage Example

(a) Sci-Hub Homepage

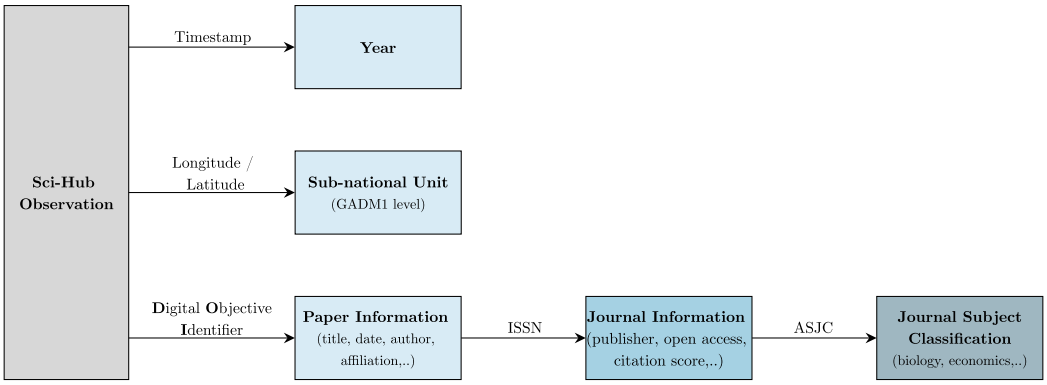
(b) Search Term

(c) Accessed Paper



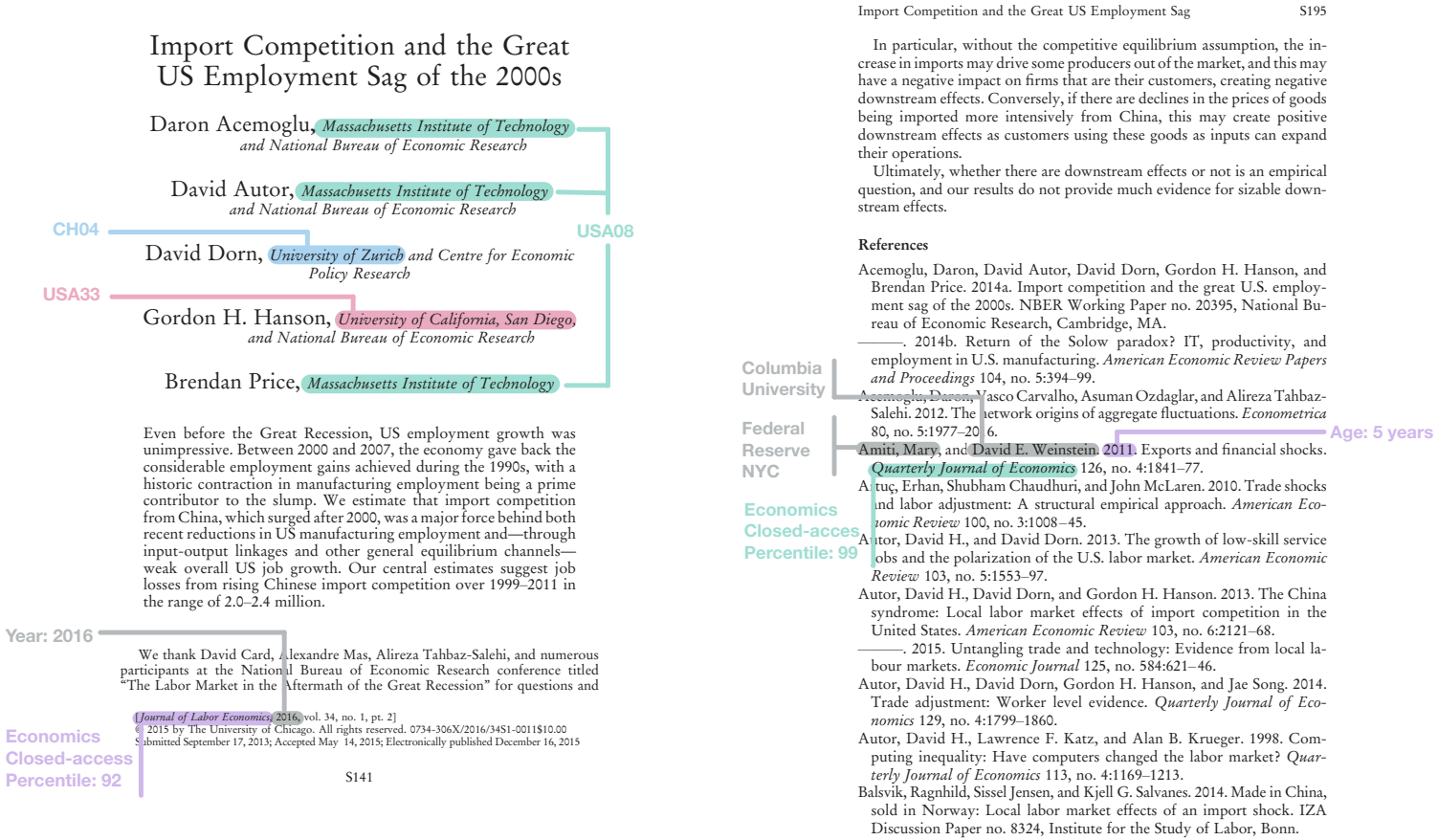
Note: The figure shows an example use case of Sci-Hub. Figure A.1a shows a snapshot of the Sci-Hub home page as of November 3, 2022. To access a paper, the user can specify either the unique digital object identifier or the paper title. If the requested document is available through Sci-Hub, the user can download the document directly after confirming the request by pressing the "open" button as shown in Figure A.1b.

Figure A.2: Sci-Hub Data Structure



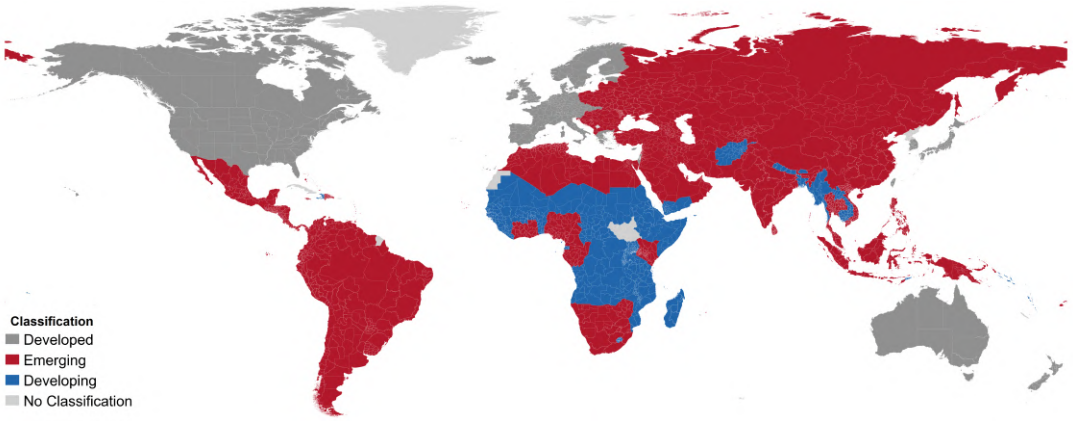
Note: The figure shows the structure of an entry in the Sci-Hub log-files download and describes how it is subsequently processed.

Figure A.3: Research Output Classification Example



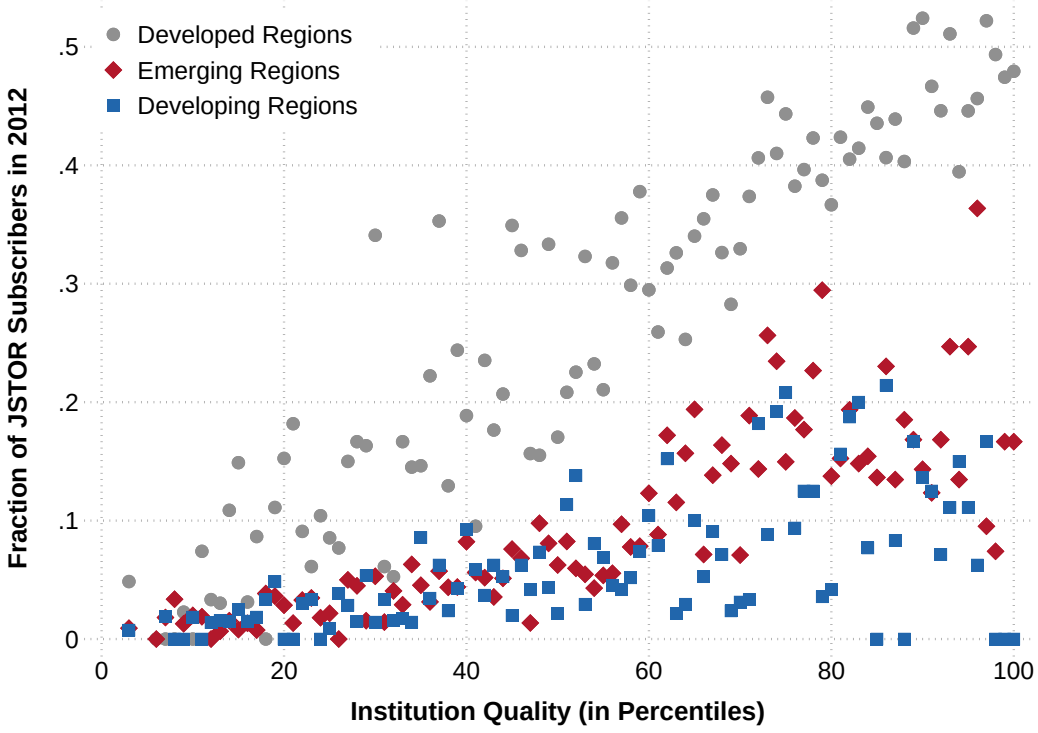
Note: The figure describes the type of characteristics extracted from a publication recorded in OpenAlex.

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.

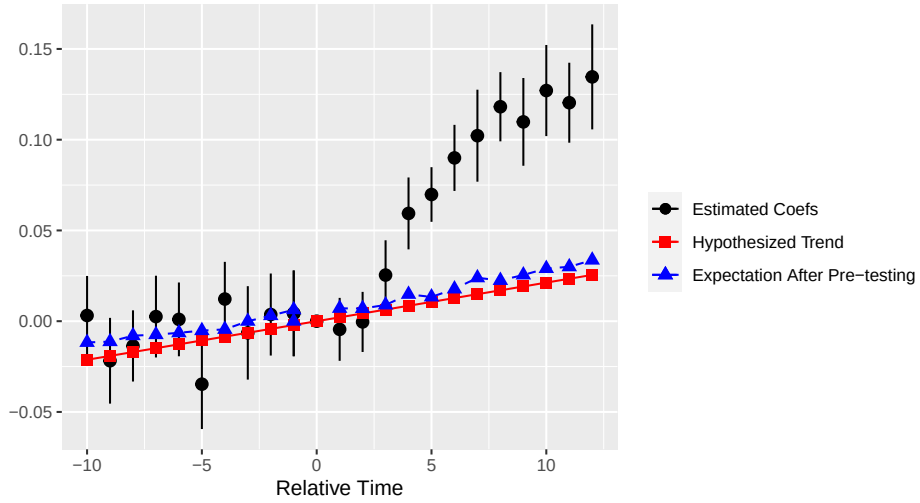
Figure A.5: Fraction of JSTOR Subscribers by Institution Quality and Region



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

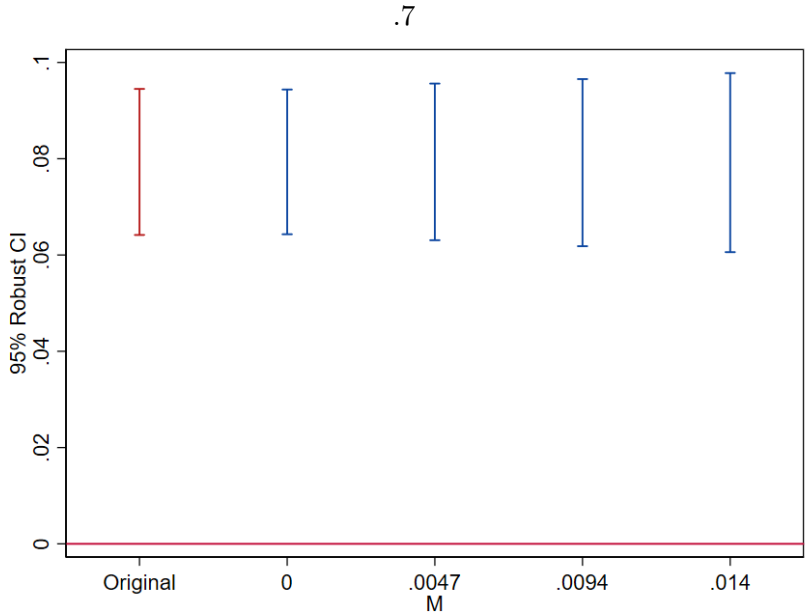
9

Figure A.6: Pre-trends



Note: This figure implements Roth (2022) and depicts linear violations of parallel trends that we would be detected with 80% power in pre-trend tests. The outcome is the Share of Restricted-access References. The corresponding exhibition in the main text is Figure 9 Panel (a).

Figure A.7: Sensitivity to Non-Linear Parallel Trend Violations



Note: This figure implements Rambachan and Roth (2022) and depicts confidence sets for the main effects at different values of M. The outcome is the Share of Restricted-access References. The corresponding estimate is in Table 6 Column (4).

E Additional Tables

Table A.1: First Stage Estimates – Inverse Hyperbolic Sine Transformation

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: ihs Downloads								
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
Panel B: ihs Cumulative Downloads								
ihs CON Almaty × Post 2010	0.859***	0.920***	1.014***	0.631***	0.428***	0.435***	0.494***	0.491***
	(0.089)	(0.081)	(0.156)	(0.125)	(0.092)	(0.093)	(0.102)	(0.098)
Observations	56,051	56,051	54,832	54,832	54,832	54,832	54,832	54,832
Number of Clusters	195	195	142	142	142	142	142	142
F-statistic	92.131	128.275	42.292	25.563	21.625	21.749	23.598	24.840
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 country level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.2: 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 country level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table A.3: Weak IV Robustness – Effect on 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***	-0.024***	0.025***	0.049***	0.049***	0.047***	0.045***	0.046***
	(0.002)	(0.002)	(0.003)	(0.015)	(0.011)	(0.010)	(0.009)	(0.009)
Observations	20,453	20,401	19,410	19,410	19,410	19,410	19,410	19,410
F-statistic	74.401	90.766	36.214	22.417	17.774	17.997	19.998	20.436
Panel B: Weak IV Considerations								
Endogeneity Parameter ρ								
$\max\{\lvert\hat{\rho}_L\rvert, \lvert\hat{\rho}_U\rvert\}$	0.380	0.460	0.480	0.820	0.730	0.720	0.690	0.690
Anderson-Rubin Inference								
p-value	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$	$< .001$
tF-adjusted Standard Errors								
5-percent Significance	(0.002)	(0.002)	(0.020)	(0.020)	(0.015)	(0.014)	(0.012)	(0.012)
1-percent Significance	(0.003)	(0.002)	(0.028)	(0.028)	(0.022)	(0.021)	(0.018)	(0.018)
Fixed Effects								
Sub-national	-	✓	✓	✓	✓	✓	✓	✓
Year × Country	-	-	✓	✓	✓	✓	✓	✓
CON Neighb. Capitals	-	-	-	✓	✓	✓	✓	✓
Controls 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 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 intervals 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 country level. Significance levels are indicated as follows: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.