The Great Firewall and Knowledge Diffusion[*]

Andrew B. Bernard
Tuck@Dartmouth
CEP, CEPR & NBER

Esther Bøler
Imperial B School
CEP & CEPR

Davin Chor
Tuck@Dartmouth
NBER

Sirig Gurung
SIEPR

Wei Lu
World Bank

March 2024
VERY PRELIMINARY. DO NOT CITE OR QUOTE

Abstract

This paper examines the role of internet restrictions in China on the flow of knowledge across borders. China imposes substantial restrictions on the flow of data in and out of the country including blocking Google.com and its subdomains. Webpages of academic researchers that are hosted on sites.google.com are blocked in China. We find that research articles written by authors who host their personal website on sites.google.com have significantly fewer citations and that the reduced citations begin at the same time as the Chinese restrictions on Google. The reductions in citations is even larger for papers that reference China in the title or abstract even if those papers are not hosted on a google site.

JEL Classifications: F13, F14, L86

Keywords: internet, citations, China

[*] We thank Treb Allen, Sam Asher, Yang Jiao, Paul Novosad, and Nina Pavcnik and seminar audiences at Dartmouth and Imperial for helpful comments. Paige Xu and Bashudha Dhamala provided excellent research assistance. All errors are our own.

1 Introduction

National restrictions on internet access are growing. Countries from China to Venezuela to Russia substantially limit which websites their residents can view online. While designed primarily to restrict access to news and non-sanctioned information, these restrictions can have unintended consequences on the flow of other information across borders. This paper examines the unintended effects of China's prohibition on Google on citations of non-Chinese academic research in economics.

China implemented internet restrictions in the 1990s under the Golden Shield project to gather data on internet usage of its citizens/residents and to limit what could be posted online.[1] Our interest is in a subset of the Golden Shield known as the Great Firewall which has the ability to block content by blocking IP addresses and other means. The Great Firewall is the colloquial term for the nation-wide internet control that filters and censors content in mainland China (Ensafi et al. (2015)). The absolute control over internet content enables the government to manage idea, knowledge, and information exchange without international interference, like a "closed economy". Since the Chinese government owns all internet service providers (ISPs), it can monitor and control what information people in China access online.

In 2010, China used the Great Firewall to block Google news and subsequently all Google-related sites began to be unreliable in China including webpage hosting at sites.google.com (GS). As a result, scholars in China, a large and rising share of the global total, were, and continue to be, unable to easily access research posted on those pages. It is well known that it is possible to circumvent the restrictions of The Great Firewall using VPNs and other techniques including private file-sharing; even so, the Chinese authorities have been known to implement counter-measures periodically that obstruct VPNs. The inability to **easily** and reliably search, find, and download content from restricted sites thus raises the cost of accessing academic research hosted on those sites.

This paper examines the (potentially) unintended effect of these restrictions on citations of academic economics research. We call this effect unintended because China does not block access to personal webpages of researchers hosted on other platforms. In addition, the same research working papers that are blocked when hosted on Google Sites (GS) are available if the researcher has an affiliation with a research organization such as the National Bureau of Economic Research (NBER) in Cambridge, MA or the Centre for Economic Policy Research (CEPR) in Paris. Similarly some repositories of working papers are not blocked by the Great Firewall.

We find that, as of 2020, total lifetime citations of foreign papers hosted on GS are significantly lower than those for foreign papers hosted on other services. Using a difference-in-difference

[1] The Great Firewall of China, Bloomberg News October 12, 2017

approach we find that the change in relative citations begins in 2010, precisely when the Great Firewall began blocking Google. After 2010, annual citations fell significantly and persistently for GS-hosted papers relative to non GS-hosted papers. This effect remains even after controlling for paper, year, and paper age fixed effects and differential citation lifecycles across groups of papers.

The Great Firewall does not just block domains of large organizations such as Google, but also attempts to restrict access to information on specific topics. We hypothesize that the Chinese government did not want unfettered access to economic research on China. Looking at papers about China that were hosted on otherwise unblocked websites, we find an even larger decline in relative citations post-2010.

While our baseline specification necessarily focuses on an older sample of papers to look at citations before and after 2010, there is evidence to suggest that the knowledge-restricting effects of the Great Firewall may be much greater now than in the past. China-based authors are a much larger fraction of academic economics researchers and the share of foreign papers hosted on GS has been rising rapidly over time.

This paper contributes to the literature on the diffusion of knowledge and ideas across space. Citations have been found to decrease with distance, and across borders (Peri, 2005).[2] Recent contributions find that this distance effect has decreased over time (Head et al., 2019), and is less relevant for academic papers than it is for patents (Belenzon and Schankerman, 2013).[3] Knowledge flows are found to reach much farther than trade flows (Peri, 2005).

There is a growing body of work examining the increasing importance of China in global knowledge production (Xie and Freeman (2019), Qiu et al. (2022), Aghion et al. (2023)). Some of this work focuses more specifically on the effects of the Great Firewall on knowledge and information flows. Zheng and Wang (2020) find that the Firewall altered the behavior of inventors in China in that they became less able to seek distant knowledge. This led to the economic value of their inventions decreasing, compared to inventors in neighboring regions, not impacted by the Great Firewall (Zheng and Wang (2020)). Similarly, Kong et al. (2022) also find evidence that suggests that the Great Firewall has had a negative effect on foreign knowledge spillovers, as it adversely impacted the intensity and quality of innovation of firms that rely on foreign knowledge, where the affected firms cite fewer foreign patents, and their innovation efficiency declines after Google's exit from China.

The paper is also related to a small literature on other consequences of the Great Firewall. Li et al. (2023) find that limited information accessibility after Google was blocked led to deteriorating export quality. Firms and products facing greater information frictions in China experienced

[2] Ganguli et al. (2020) find that the effect of distance is even stronger when comparing patent interferences, i.e. when two or more independent parties submit close to identical claims of invention nearly simultaneously, than it is for patent citations.

[3] Sin (2018) finds that the distance effect has decreased over time also for book translations.

a decline in their export quality. Wang et al. (2023) show that the Great Firewall is associated with the strategic disclosure of information by Chinese firms. As investors had more difficulty assessing foreign markets and projects, Chinese firms disclosed more optimistic expectations of foreign projects relative to domestic ones.

Unlike these papers, we are able to compare the effects on blocked and unblocked knowledge. We look at the citations on foreign papers rather than the outcomes of Chinese scholars themselves.

The paper proceeds as follows. Section 2 discusses the Great Firewall, aspects of its goals and operation and how it relates to economics research. Section 3 describes the construction of the data on authors, co-authors, and articles. Section 4 discusses potential selection issues and presents our baseline empirical specification. Section 5 then reports the regression findings and extensions and discusses implications for more recent years. Appendices report regression coefficients and document in more detail the process of data acquisition and cleaning.

2 The Great Firewall

The Golden Shield project is the Chinese nationwide network-security fundamental constructional project by the e-government of the People's Republic of China. This project includes a security management information system, a criminal information system, an exit and entry administration information system, a supervisor information system, a traffic management information system, among others.[4]

The Great Firewall is a component of the Golden Shield project and is designed to block access to selected foreign websites and information deemed unacceptable by the Chinese authorities and to control (and slow) cross-border internet traffic. Sites may be selected to be blocked because of their overall purpose, i.e. the New York Times (news) and DuckDuckGo (search) or because they contain keywords or phrases deemed to be inadmissible. If a link is closed then other links from the same machine will likely also be blocked, a form of "guilt by association".

Some commonly used technical methods for the Great Firewall include IP blocking, DNS spoofing filtering and redirection, URL filtering, and VPN blocking among others. Though a user could technically bypass the Great Firewall using VPNs or Proxies, the Great Firewall uses deep packet inspection and machine learning to shut down suspected VPN or proxy tunnels, and as of today, many fewer commercial VPN services are viable in China compared to a few years ago (Tang, 2016). While there has been evidence that suggests disparity between the number of servers found in each city, Wright argues that there is no discernible overall geographic pattern to the nature or extent of filtering despite significant variation (Wright, 2012). Therefore, the Great Firewall may be better understood as a decentralized and semi-privatized operation in which

[4] https://en.wikipedia.org/wiki/Great_Firewall.

low-level filtering decisions are left to local authorities and organizations and high-level control is loose regarding implementation (Wright, 2012).

The first part of the Great Firewall known as the Golden Shield Project was enacted at the end of 2006, making it difficult for citizens to reach the foreign sites that the government deemed as illegal, including Google. By 2010 Google decided to redirect its mainland China customers to its Hong Kong-based site (BBC News, 2010). While Google was technically "banned" in mainland China in 2010, it was still accessible by mainland Chinese only with occasional breakdowns. In mid-2014, the CCP intensified its crack-down on the Google search engine and Google-related products due to the 25th year anniversary of Tiananmen; Google, the primary search engine in the rest of the world, has since been erased entirely from the Chinese market.

The blocking of related websites is precisely the source of restrictions that are the subject of this paper. Google.com was blocked in 2010 because the main search feature of Google was returning results deemed unacceptable to the Chinese government. In addition, the news feature of Google provided access to restricted foreign news sites and content and Google refused to self-censor its search and news results. The web-hosting feature of Google at sites.google.com was likely not considered problematic by the Chinese government per se but instead was "guilty" merely by being a subdomain of the larger Google site. Other examples of the (potentially) unintended consequences of blocking Google are the failure of websites with Google fonts, Google Analytics, Google CDNs and reCaptcha.[5] Google.com and sites.google.com remain blocked in China to this day.

2.1 VPNs

Just because webpages hosted on Google are not easily visible in China does not mean that they are completely unavailable. Residents of countries with substantial online government censorship are not the only internet users who may want to hide their country of origin or establish a secure connection to a website. Access to country-restricted content, such as Netflix US and BBC iPlayer, is marketed as a feature by providers of Virtual Private Network (VPN) software.

VPNs are advertised as a means to protect one's personal data and to shield online activity from outside observers, either government or internet service providers or other online agents, by rerouting the user's internet traffic through a remote server before sending it on to the final destination.

However, in China, only VPNs authorized by the Chinese government can be legally used. These authorized VPNs have backdoors that allow the government to monitor activity. Other VPNs, while not legal, are available in China. However they do not provide uninterrupted easy

[5] The problems are not limited to Google, Facebook, X/Twitter and even AWS might lead to a site not loading in China.

access to restricted content. One feature of the Great Firewall is that access today may not guarantee access one day later (or even one hour later). The ultimate effect is to dramatically raise the search cost of acquiring information on restricted internet sites. In practice this means that a working VPN is best used to go to specific sites, and not for more informal internet searches. If a researcher knows which particular site they want to visit, a VPN may solve that problem. If they are not sure what sites they should visit to learn about current research, the cost may be substantially higher or prohibitive.

2.2 The role of working papers in economics

This study focuses on the effect of the Great Firewall through sites.google.com (GS) on the citations of economics research outside China. The field of economics is particularly useful in this regard because there is a widespread norm of sharing and presenting working papers early in the research process, usually many years before formal publication in a refereed journal.[6] Unlike many other academic fields, researchers in economics learn about these early stage papers, cite them and are influenced by them before formal publication. In fact, it is generally not considered appropriate for papers presented at conferences or in departmental seminars to have been accepted for publication.

This feature of the economics literature means that online posting of working papers, often through personal websites, is a crucial means of disseminating current research. Economics researchers are likely to look at personal web pages to see what topics are being researched. Authors who host their personal webpages on GS will be harder to find and thus less likely to be cited by China-based researchers.

In academic economics, as in the broader global economy, China has become increasingly important in the last two decades. China has roughly the same number of colleges and universities as the US (Statista.com) and the fraction of papers in economics journals with an author based at a Chinese university rose from 5.2% (816) in 2004 to 9.5% (5077) in 2020.[7]

The rise of China-based economic researchers means that an increasing share of citations in economics are coming from China and these cited references are less likely to be papers hosted on GS websites because of the Great Firewall.

2.3 Possible Solutions and Mitigation

In spite of the growth in Chinese economic research and the concurrent difficulty in seeing websites hosted on GS, there are several reasons that the Great Firewall might have a small impact on the

[6] Publication in an academic journal does not guarantee articles will be available in China. Spring Nature self-censored 1000s of articles at China's request in 2017 (Reuters, 2017). Websites of academic publishers have also been blocked in China.

[7] Results from searches on WebofScience.com using CU=(china) and SO=(economics) and YR=(2004 or 2020).

flow of knowledge across the border in the economics profession.

First, it is possible that VPNs and other workarounds might mitigate the lack of information for Chinese scholars.[8] Second, if non-Chinese researchers are aware of the problem, they have an incentive to shift their web hosting away from Google, resulting in fewer personal research webpages being hosted on GS. However, the fraction of papers with at least one author on GS has steadily increased over our sample period.

Finally authors may submit their papers to other sites that aggregate and disseminate economic research. However, not all these sites are available in China. The largest repository for disseminating economics working papers, SSRN.org, is blocked in China, while RePEc.org, another large internet repository of academic material, is not blocked.

The fact that webpages are blocked in China if they are hosted on GS would be less problematic if non-Chinese researchers were aware both that their webpages are being blocked in China and that the restriction has a noticeable effect on the dissemination of their research.

3 Data

Our goal is to determine whether non-Chinese academic economists are cited less often if they host their personal webpage on Google Sites (GS). To do this, we take a paper-based approach-rather than an author-based approach in order to control for as many unobserved characteristics of the authors as possible. Running regressions of total author citations on a dummy for the GS characteristic of the personal webpage fails to control for substantial heterogeneity in author ability, institutional visibility and other strong determinants of citations. Following a paper-based approach with paper fixed effects helps control for a large number of time-invariant unobservables.

The ideal dataset would have comprehensive information on cited as well as citing papers including date of first dissemination online, date of any eventual journal publication, existence in working paper series, the names of all authors, and citations by year. Webhosting sites would be randomly assigned across authors.

Data for authors on cited papers would include the existence of a personal webpage, the hosting site, home institution, age, gender, and additional institutional affiliations (including thinktanks). In addition, to confirm that any observed effects on citations are due to the Great Firewall, we would like to have citing paper information with the location of all the authors. While we are able to assemble a version of the data for cited papers we have not been able to collect information on citing papers and the location of their authors.

[8] One such workaround involves a network of Chinese researchers sharing non-Chinese working papers on sites in China behind the firewall.

3.1 Sample Creation

We start by assembling a list of academic staff at economics departments/institutions ranked as in the top 50 by Tilburg University in 2019. From that list of 2,991 individuals, we retain those with Google Scholar pages in 2023, resulting in 1,804 unique authors. All papers on the profile pages of these authors are scraped from Google Scholar in 2023 yielding an initial set of 145,590 papers.

Deduplication of papers, both within- and across-authors, trims that initial list to 102,842 main English language papers in the social sciences.[9] 96,987 of these papers have a cumulative citation count of at least one through 2020. Using the fuzzy matching merging score set to 80, the broad sample drops to 94,299 papers.

All 70,995 author/co-author names are extracted from these papers and grouped where necessary using fuzzy matching algorithms (co-author names were less clean due to differing spellings and name formatting) yielding 61,439 authors in total and 59,635 authors not on the original list of authors from the top 50 schools. We did not go back to scrape all the additional papers by the coauthors, so each paper in the dataset contains one or more authors at the top 50 institutions.[10]

For all the authors (original top50 and co-authors), we scrape Google search results to look for personal websites and to determine whether the personal websites are hosted on GS. 7,024 out of 61,439 had personal websites hosted on GS.[11]

In order to determine the start of life for each paper in the sample, we create a *startdate* variable using two distinct pieces of information available in the scraped data from Google Scholar. One, called "publication date" by Google Scholar, is most likely associated with either publication of the paper in an academic journal or the publication of the paper in a formal working paper series. The second variable is the year that the paper is first recorded as being cited by Google Scholar, "first citation". Both of these measures are noisy, with some first citations occurring decades before the paper was circulating and with publication dates not systematically assigned across papers. We manually clean these variables before creating our preferred measure, *Startdate*, as the min of the two variables.[12]

From this list of 94,299 papers, we create our baseline sample for the empirical work. The baseline sample includes papers with a *startdate* between 2000 and 2008 inclusive. In order to establish whether citations are associated with the onset of the Great Firewall prohibition on

[9] Deduplication involved removing multiple copies of same paper names using a fuzzy matching algorigthm. WE also drop any papers in fields outside social sciences.

[10] Given the enhanced visibility of authors at top 50 institutions we may be under-weighting the negative effects on citations of hosting on GS.

[11] We do not have precise information about when any particular website was created or when it started being hosted on GS. The Wayback Machine potentially contains such data but gives relatively imprecise information and is contingent on when then sites and papers were indexed.

[12] The data appendix describes our effort to clean these variables.

Google, we include papers that are first cited/published before 2010 so we can determine if they have differential trends before and after the GF restrictions.[13] We choose 2000 as the oldest *startdate* to make sure each paper could have been searchable on the public internet. We include annual citations for each paper from 2004 until 2020. We start in 2004 as Google Scholar citations seem less comprehensive in earlier years. These restrictions trim the larger sample down to 27,918 unique papers in the baseline sample.

For each paper, we have the following variables: *startdate*, citations by year, the number of authors, the number of authors with a personal page hosted on GS, the number of authors at top 50 institutions, the rank (1-50) of the institution for top50 authors, and the title and abstract.

4 Empirics

To begin, by way of motivation, we check whether the cumulative citations are lower for papers with at least one author with a personal webpage on GS. We run a cross-sectional regression for the year 2020 on the large sample of 94,299 papers of the form:

$$CumCite_p = \beta \mathbf{I}(AtLeastOne)_p + \sum_a \beta_a \mathbf{I}(PaperAge_p = a) + \varepsilon_p \tag{1}$$

where $CumCite_p$ is the cumulative number of citations for the paper since $startdate_p$, $(AtLeastOne)_p$ is an indicator variable if the paper has at least one author with a personal page on GS, and $PaperAge_p$ is the number of years since $startdate_p$.

Table 1 reports the results for all papers and papers with a *startdate* on or after 2000, without and with paper age fixed effects. The simple cross-section shows a significant negative coefficient on the GS variable. Controlling for paper age, we find 8.62 fewer citations for papers who had at least one author hosting a webpage on sites.google.com. Given that the mean level of citations for this sample of papers is 97.7 and the median is 18, this represents a substantial reduction in citations over the life of the paper.

4.1 Selection and GS

In this section we discuss possible issues with paper and author characteristics that could be correlated with GS webhosting and might affect the path of citations over the life of a paper. Authors who choose to host on GS might have different characteristics and the nature of the *AtLeastOne* variable might induce a correlation between GS and other paper characteristics.

[13] As mentioned above, the precise timing of the prohibition on Google.com is hard to date but during and after 2010 the site was almost continuously blocked by the Great Firewall.

Table 1: Cumulative Citations and Google Sites, 2020

	All papers	
At Least One GS Author	-65.18	-16.64
	(4.86)	(3.31)
Paper Age Dummies	N	Y
Obs.	94,299	94,299
	Startdate 2000-2020	
At Least One GS Author	-33.40	-8.62
	(2.87)	(2.89)
Paper Age Dummies	N	Y
Obs.	68,296	68,296

We use two variables to help control for omitted author and paper characteristics. First, we consider the rank of the institution of the highest ranked author on the paper.[14] Authors at higher ranked institutions may have less need for personal web pages to advertise their papers and their papers may have a different citation lifecycle due to the high visibility of the authors and their institutions. We create three sets of papers, one if *MaxRank* is in the top 10 departments, one if *MaxRank* is in the rank 11-20 group, and one if the highest ranked author is in at an institution ranked 21-50.[15] These three sets of papers have significantly different shares of papers with *AtLeastOne* = 1, ranging from a high of 27.8% for the 21-50 group to 20.3% in the 11 to 20 group to a low of 11.4% for papers with an author in the top ten departments. Lower ranked authors are more likely to have a GS personal webpage.

Figure 1a shows that the citation path over the age of the paper for the three groups of papers also varies substantially. That the level of citations is positively correlated with the *MaxRank* of the authors is probably unsurprising. However the top-ranked group shows a more rapid increase in citations early in the paper lifecycle, peaking around year 10, and declining less afterwards. The lowest rank group peaks lower and earlier and declines rapidly. These results and the large variation in GS hosting argue for the inclusion of separate *paperage* effects for the different sets of papers.

Our measure of GS hosting, *AtLeastOne*, equals one if any of the authors hosts a personal webpage on GS. This means that papers with more authors will have a higher probability of *AtLEastOne*=1 even if hosting is randomly assigned across authors. In the baseline sample the fraction of GS hosted papers rises with the number of authors, 11.8%, 21.6%, 25.0%, and 29.4%

[14] Since every paper has at least one author at a top 50 institution this is equivalent to the highest rank of the top50 authors.

[15] The three groups have 8,904, 7,523 and 11,126 papers respectively.

for 1, 2, 3, and 4+ authors. As multi-authored papers tend to receive more citations (171) than single-authored papers (127), we separately control for *paperage* and author numbers in our regression.[16] Figure 1b shows the citation path over the age of the paper for papers with 1,2, 3, and 4+ authors. Except for single authored papers, the lifecycle of citations is similar across the groups.

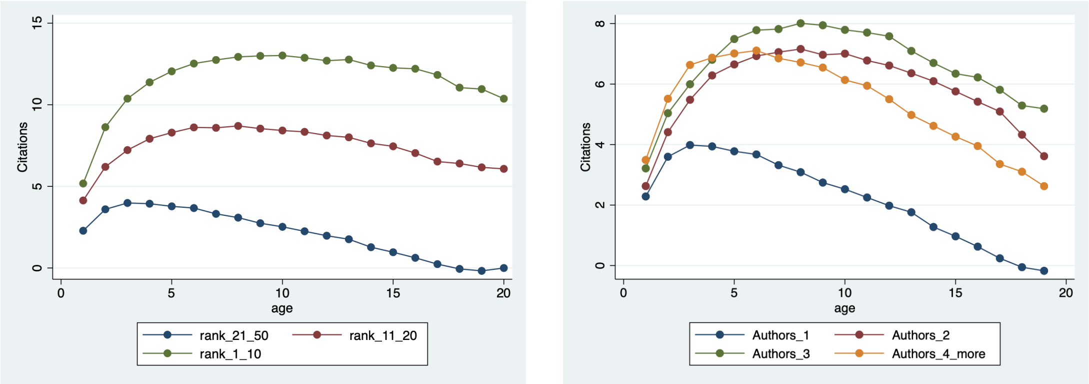
4.2 Baseline

The substantially lower lifetime citations for papers with *AtLeastOne* provides suggestive evidence that the Great Firewall has reduced citations. We now turn to our preferred difference-in-difference specification where we examine the paths of citations for individual papers before and after the blocking of Google by the Great Firewall. Splitting the sample into papers with no authors on GS versus papers with at least one author on GS allows us to check for pretrends as well as different relative citation rates after 2010.

Figure 1: Citations by Paper Age

(a) Paper and MaxRank Group

(b) Paper Age and Author Number Group



Sample includes papers with startdate between 2000-2008 inclusive
Basic specification: paper, year, and paperage FEs

We regress annual citations for each paper on year dummies interacted with an indicator for at least one author having a personal webpage on sites.google.com (*AtLeastOne*), a complete set of paper age (*PaperAge*) dummies interacted with both the four groups of number of authors (*NumAuths*) and the three groups of maximum author institution rank (*MaxRank*), as well as year and paper fixed effects.

[16] We include four dummies and interactons with *paperage*for papers with 1, 2, 3, and 4+ authors.

The baseline specification is

$$
\begin{aligned}
C_{pt} \;=\; & \sum_t \beta_{Gt}(D_t \times \mathbf{I}(AtLeastOne_{ip})) + \\
& + \sum_n \sum_a \beta_{na}\mathbf{I}(NumAuths_p = n)\mathbf{I}(PaperAge_{pt} = a) \\
& + \sum_m \sum_a \beta_{ma}\mathbf{I}(MaxRank_p = n)\mathbf{I}(PaperAge_{pt} = a) \\
& D_t + D_p + \varepsilon_{pt} \qquad (2)
\end{aligned}
$$

with $\beta_{G2010} = 0$ and standard errors clustered at the paper level.

5 Results

5.1 Baseline

Figure 2 shows the coefficients of *AtLeastOne* by year in the baseline specification.[17] Following the blocking of Google in 2010 the relative citation rates for papers hosted on GS is lower, usually significant and increasing in magnitude. On average in the post-treatment period, these papers received 0.44 fewer citations per year, or 3.8% fewer citations overall. There is no evidence of pretrends, none of the coefficients before 2010 are significantly different from zero.

5.2 China Papers

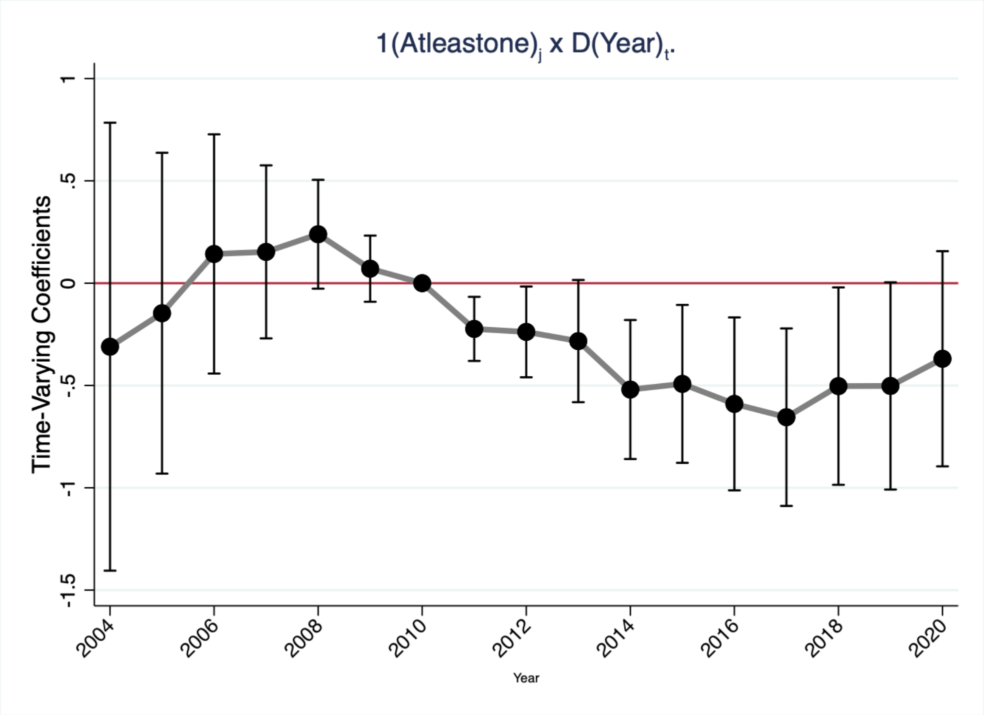
As mentioned earlier, the Great Firewall both blocks specific domains such as the Wall Street Journal as well as sites containing individual words or content that are deemed inadmissable. While the blocking of Google.com is well known and the timing is fairly well documented, it is less clear what other content is not allowed to pass through.

Here, we hypothesize that the Great Firewall is also blocking pages that contain economics papers about China itself. Controlling news and information about China from outside sources is one of the stated purposes of the Great Firewall so, it stands to reason, that the Chinese government might not want academic papers about China to be readily available. We construct a variable, *ChinaReference*, that equals one if the title or abstract of the paper contains "China" and if the paper does not have an author with a personal page on GS.

Since websites on GS are already blocked, any such paper would already be largely unavailable in China. In fact, in the baseline specification, papers with *ChinaReference*=1 would be in the control sample and, if blocked separately, would reduce the estimated coefficients on *AtLeastOne* after 2010.

[17] Appendix A reports the coefficients and standard errors.

Figure 2: Effects of At Least One Author Hosting on GoogleSites



(a) Sample includes papers with *startdate* between 2000-2008

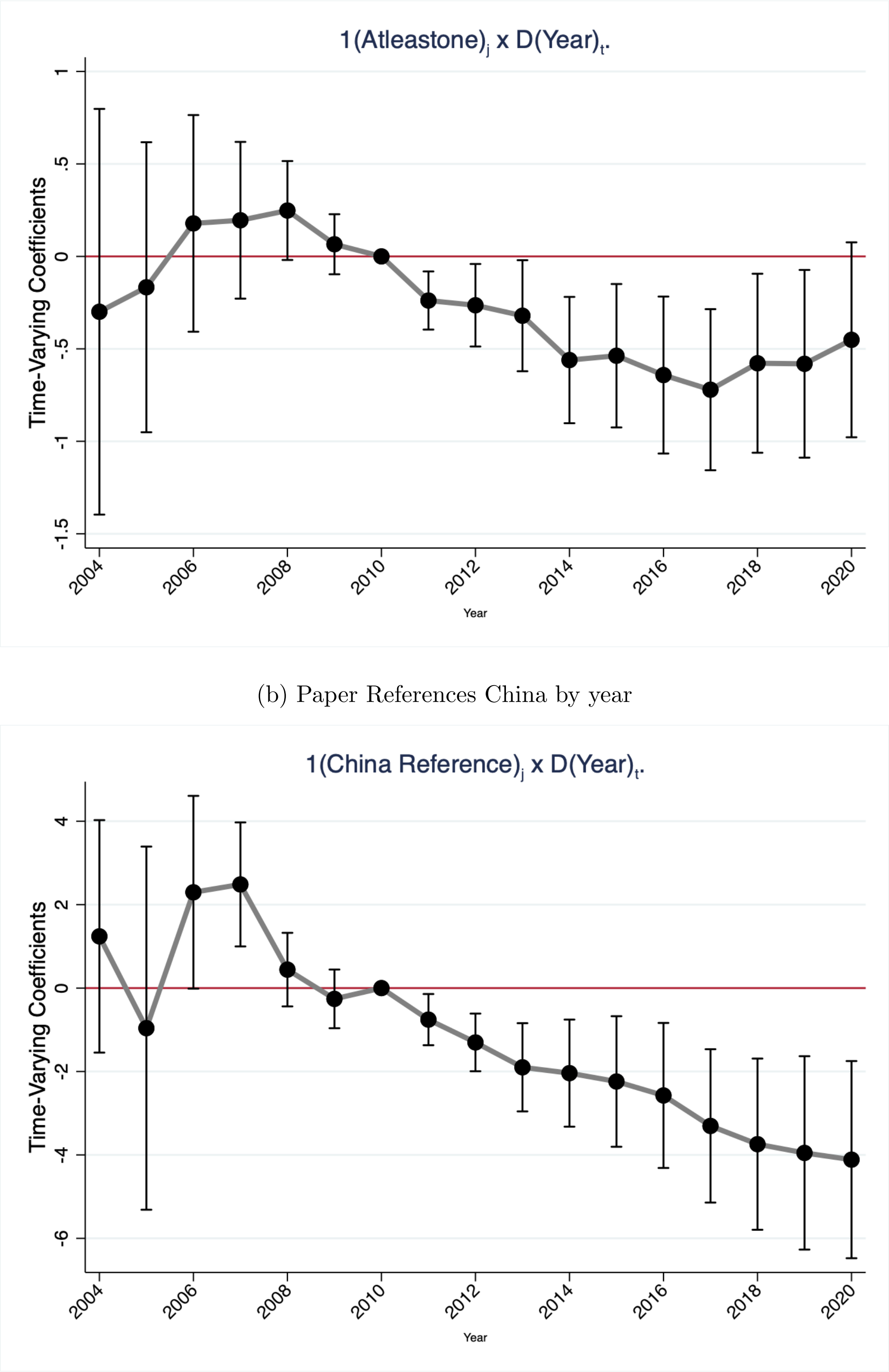
(b) Basic specification: paper, year, and paperage FEs, Paperage interactions w Author Number groups and MaxRank groups

We are uncertain when China began blocking content more broadly, so we assume it also occurred in 2010 and augment the baseline specification with a full set of year dummies interacted with *ChinaReference* as given in equation 3

$$\begin{aligned} C_{pt} \quad = \quad & \sum_t \beta_{Gt}(D_t \times \mathbf{I}(AtLeastOne_{ip})) + \\ & \sum_t \beta_{Ct}(D_t \times \mathbf{I}(ChinaRef_{ip})) + \\ & \quad + \sum_n \sum_a \beta_{na} \mathbf{I}(NumAuths_p = n) \mathbf{I}(PaperAge_{pt} = a) \\ & \quad + \sum_m \sum_a \beta_{ma} \mathbf{I}(MaxRank_p = n) \mathbf{I}(PaperAge_{pt} = a) \\ & \quad D_t + D_p + \varepsilon_{pt}. \end{aligned} \tag{3}$$

Figure 3a shows the year-by-year coefficients for *AtLeastOne*. The pattern from the baseline specification is preserved although the coefficents after 2010 are more negative, averaging -0.459, and the standard errors are smaller. This suggests that blocked papers about China had been

Figure 3: Effects on Papers that Reference China



(a) *AtLeastOne* googlesite by year

(b) Paper References China by year

(c) Sample includes papers with *startdate* between 2000-2008 inclusive

(d) Basic specification: paper, year, and paperage FEs, Paperage interactions w Author Number groups and MaxRank groups grouped with unblocked papers in the baseline results.

Figure 3b shows the year-by-year coefficients on the *ChinaReference* variable. Again the trend is downward, papers not on GS that have China in the title or abstract are significantly less likely to be cited after 2010 relative to other non-GS papers.[18] The magnitude of the coefficients are dramatically larger, with the cumulative effect between 2011 and 2020 amounting to 24.3 fewer citations. The average paper in this group had 149 cites by 2020.

5.3 Extensions and Robustness

Generally, the Great Firewall does not block individual papers but rather blocks the IP address hosting undesirable content. This means that researchers who do not use GS but write a paper about China may also have their webpage blocked. The problem with examining this additional "guilt by association" effect is the number of such papers and thus authors is small in the baseline sample.

The findings above suggest a large negative impact of hosting personal websites on GS paper citations as well as large impacts of the Great Firewall on citations for papers about China. However, there are methods to make such papers visible. Most university webpages outside China are not blocked nor are websites of prominent economic thinktanks such as the National Bureau of Economic Research in the US or the Centre for Economic Policy Research in France.[19] Assembling membership lists of those organizations may let us see if they offset the effects of GS or China papers.

Robustness

We have done some robustness checks on the results. Changing the fuzzy merging score in the bigram algorithm does not change the results although the effects are smaller on *AtLeastOne* if the matching parameter is less stringent. This may be the result of duplicate non-GS papers in the sample.

Shifting the *startdate* range earlier to 1996-2004, we find similar patterns although the sample is much smaller and the standard errors increase and coefficient magnitudes are slightly smaller. Shifting the citation years earlier to 2004-2016 also does not change the patterns. Again the sample is smaller but the effects remain.

5.4 Implications

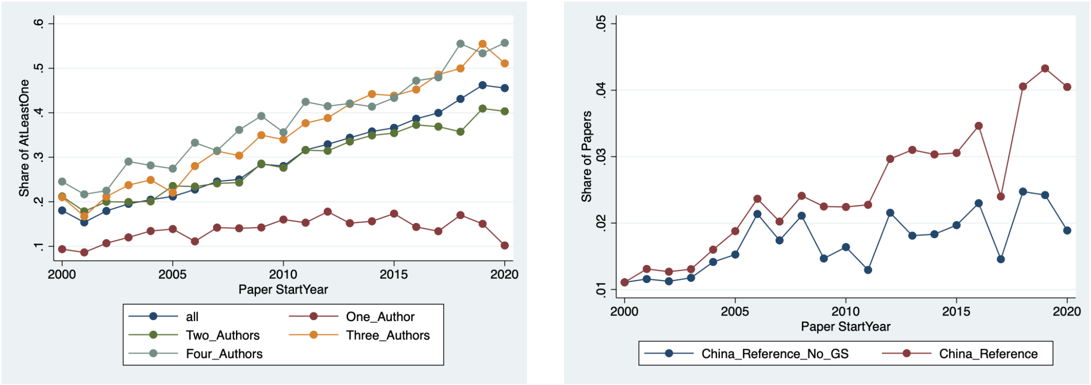
By construction the baseline sample of papers for our empirical exercise had to predate the imposition of the restrictions on Google-related sites in 2010. We used papers whose start date

[18] There is a possibility that the blocking of papers about China started earlier, i.e. in 2007

[19] University webpages in Taiwan are blocked.

was between 2000 and 2008 inclusive. This sample therefore consists of relatively old papers. However the growth in the use of personal websites, and Google Sites in particular, means that the aggregate impact of these restrictions is potentially much greater for the full sample of papers from recent years.

Figure 4: Papers Over Time



(a) Papers with AtLeastOne GS Author

(b) Papers about China

Figure 4a shows the fraction of papers with at least one GS by paper start year and the number of authors on the paper. Overall the fraction of new papers with at least one GS has risen from 18.1 % in 2000 to 45.5% in 2020. Most of that increase occurs after 2008, the last year of new papers in our baseline sample. Considering papers grouped by the number of authors, we find that GS papers have risen in all categories, but the fastest increase is for papers with three and four+ authors. At the same time those two categories have increased their share of papers from 28.% in 2000 to 63.0% in 2020.

The 10 year cumulative loss of citations ranges between 4.3 and 4.8 citations and the 10 year cumulative citations in the sample for papers not on GS has mean 96.8 and median 23.

In line with the rise of China as an economic power, the fraction of new papers referencing China has also increased in the past several decades from 1.1% in 2000 to 4.1% in 2020 as shown in Figure 3b. Surprisingly, new papers that reference China in the most recent decade (2011-2020) are slightly more likely to have at least one author with a page on Google Sites. It does not appear that writing about China increases awareness of the fact that personal pages on Google Sites are blocked in China.[20]

[20] The fact that one of the authors of this paper continues to host a personal academic webpage on GS suggests perhaps that researchers do not care about the loss of citations, even after becoming aware of the issue.

6 Conclusion

This paper examines the effects of internet restrictions on the flow of knowledge across borders. When China used the Great Firewall to block Google search and news in 2010, it had the (potentially) unintended consequence of also blocking personal webpages hosted on sites.google.com. The work of academic researchers outside China hosting on GS became much harder to find by China-based scholars.

Using data on academic economics papers, we show that the blocking of Google significantly lowered annual citations by authors with a GS webpapge. The results are large in magnitude and persistent over time. We find that papers about China, not hosted on GS, also showed a very large and persistent drop in citations starting around the same year.

Our data does not allow us to determine definitively that the drop in citations was due to papers written by China-based researchers. We have no data on citing papers and their links to cited papers. In addition, access to a counterpart to Google Scholar for Chinese authors and papers has been blocked outside of China by the Great Firewall.

While this paper focuses on a single academic discipline, the results strongly suggest that non-Chinese research has become harder to access inside China. The long term effects of such reduced access are an important topic for further research.

References

Aghion, P., C. Antonin, L. Paluskiewicz, D. Stromberg, X. Sun, R. Wargon, and K. Westin (2023). Does Chinese Research Hinge on US Coauthors? Evidence from the China Initiative. Technical report, College de France.

Belenzon, S. and M. Schankerman (2013, 07). Spreading the Word: Geography, Policy, and Knowledge Spillovers. *The Review of Economics and Statistics 95*(3), 884–903.

Ensafi, R., P. Winter, A. Museen, and R. J. Crandall (2015). Analyzing the Great Firewall of China Over Space and Time. *Proceedings on Privacy Enhancing Technologies 2015*(1), 61–76.

Ganguli, I., J. Lin, and N. Reynolds (2020, April). The paper trail of knowledge spillovers: Evidence from patent interferences. *American Economic Journal: Applied Economics 12*(2), 278–302.

Head, K., Y. A. Li, and A. Minondo (2019, 10). Geography, Ties, and Knowledge Flows: Evidence from Citations in Mathematics. *The Review of Economics and Statistics 101*(4), 713–727.

Kong, D., C. Lin, L. Wei, and J. Zhang (2022). Information accessibility and corporate innovation. *Management Science 68*(11), 7837–7860.

Li, G., H. Ding, and F. Jia (2023). Information Accessibility and Export Quality: Evidence from China. *Review of International Economics*, 1–30.

Peri, G. (2005). Determinants of knowledge flows and their effect on innovation. *The Review of Economics and Statistics 87*(2), 308–322.

Qiu, S., C. Steinwender, and P. Azoulay (2022). Who Stands on the Shoulders of Chinese (Scientific) Giants? Evidence from Chemistry. Technical Report 30772, NBER.

Reuters (2017). Springer Nature blocks access to certain articles in China. November 1, 2017.

Sin, I. (2018). The gravity of ideas: How distance affects translations. *The Economic Journal 128*(615), 2895–2932.

Wang, K., X. Yu, and B. Zhang (2023). Panda games: Corporate disclosure in the eclipse of search. *Management Science 69*(6), 3263–3284.

Xie, Q. and R. B. Freeman (2019). Bigger than you thought: China's contribution to scientific publications and its impact on the global economy. *China & World Economy 27*(1), 1–27.

Zheng, Y. and Q. R. Wang (2020). Shadow of the Great Firewall: The Impact of Google Blockade on Innovation in China. *Strategic Management Journal 41*(12), 2234–2260.

A Appendix

Regression results

Results for Figure 2

	(1)
VARIABLES	Citation Count
2004 x *AtLeastOne*	-0.311
	[0.665]
2005 x *AtLeastOne*	-0.147
	[0.477]
2006 x *AtLeastOne*	0.143
	[0.355]
2007 x *AtLeastOne*	0.153
	[0.257]
2008 x *AtLeastOne*	0.239
	[0.162]
2009 x *AtLeastOne*	0.071
	[0.098]
2010 x *AtLeastOne*	0.000
	[0.000]
2011 x *AtLeastOne*	-0.223**
	[0.095]
2012 x *AtLeastOne*	-0.238*
	[0.135]
2013 x *AtLeastOne*	-0.283
	[0.182]
2014 x *AtLeastOne*	-0.520**
	[0.207]
2015 x *AtLeastOne*	-0.492**
	[0.235]
2016 x *AtLeastOne*	-0.590**
	[0.257]
2017 x *AtLeastOne*	-0.655**
	[0.264]
2018 x *AtLeastOne*	-0.503*
	[0.293]
2019 x *AtLeastOne*	-0.502
	[0.308]
2020 x *AtLeastOne*	-0.369
	[0.320]
Observations	394,862

Paper, year, and paper age fixed effects included.
Paper age interactions with MaxRank groups and Author Number groups included.
Robust standard errors in brackets.
*** p<0.01, ** p<0.05, * p<0.1

Results for Figure 3

VARIABLES	Citation Count		
2004 x *AtLeastOne*	-0.299	2004 x China Reference	1.240
	[0.667]		[1.694]
2005 x *AtLeastOne*	-0.167	2005 x China Reference	-0.961
	[0.477]		[2.646]
2006 x *AtLeastOne*	0.178	2006 x China Reference	2.297
	[0.356]		[1.405]
2007 x *AtLeastOne*	0.195	2007 x China Reference	2.485***
	[0.258]		[0.903]
2008 x *AtLEastOne*	0.248	2008 x China Reference	0.442
	[0.163]		[0.536]
2009 x *AtLEastOne*	0.066	2009 x China Reference	-0.259
	[0.099]		[0.428]
2010 x *AtLEastOne*	0.000	2010 x China Reference	0.000
	[0.000]		[0.000]
2011 x *AtLEastOne*	-0.239**	2011 x China Reference	-0.756**
	[0.096]		[0.373]
2012 x *AtLEastOne*	-0.264*	2012 x China Reference	-1.303***
	[0.136]		[0.420]
2013 x *AtLEastOne*	-0.321*	2013 x China Reference	-1.899***
	[0.182]		[0.643]
2014 x *AtLeastOne*	-0.561***	2014 x China Reference	-2.039***
	[0.208]		[0.780]
2015 x *AtLeastOne*	-0.537**	2015 x China Reference	-2.240**
	[0.236]		[0.952]
2016 x *AtLeastOne*	-0.642**	2016 x China Reference	-2.574**
	[0.258]		[1.057]
2017 x *AtLeastOne*	-0.721***	2017 x China Reference	-3.305***
	[0.265]		[1.118]
2018 x *AtLeastOne*	-0.578**	2018 x China Reference	-3.743***
	[0.294]		[1.247]
2019 x *AtLeastOne*	-0.581*	2019 x China Reference	-3.952***
	[0.309]		[1.409]
2020 x *AtLeastOne*	-0.451	2020 x China Reference	-4.113***
	[0.320]		[1.437]
Observations	394,862		

B Appendix

Data Details

B.1 Website Scraping/Classification

- Used https://serpapi.com/SerpAPI to retrieve search results for authors affiliated with top 50 Economics graduate program institutions.
  - Search criterion: "{Name} Economics Personal website".
  - Search results from SerpAPI are in JSON format which is easy to handle in python as a dictionary. Saved all search results (the searcher plan allows up to 2,000 downloads per hour. Can be faster if you upgrade).
- Screened for google, github, wix, weebly, wordpress, personal, academic, and other sites based on website urls that appeared in the google search[21].

B.2 Authors from Top 50 institutions

- After initial screen of sites returned from google search, downloaded the source codes for the "personal" and "other" websites to find google, wix, wordpress and weebly websites that are not obvious to screen using the url. Searched for company watermarks and copyright notices. Below are the updated updated number of authors with a site on a given domain (github site source code does not have anything unique about it other than the url, so it wasn't worth screening for those.):
  - # Authors with no sites: 8
  - # Authors with Google Sites: 368 (19%)
  - # Authors with GitHub Sites: 25
  - # Authors with Weebly Sites: 12
  - # Authors with Wix Sites: 14
  - # Authors with Wordpress Sites: 22
  - # Authors with Personal Sites: 357
  - # Authors with Academic Sites: 1940
  - # Authors with Other Sites: 513
  - # websites reclassified after source code check: 45
- Further screened for whether there was a perfect match for author name in the website source code. This was done for google sites only. After this, there are 357 authors with a google site match (18%).
- Each observation is by author-website in the final dataset, with indicators for each of the eight possible domain classifications.

[21] Domain classification criteria are in the appendix

B.3 Coauthors

- Initial screen and source code check for domain verification are identical to the Author website classification.
- After initial screen of sites returned from google search and further checking website source code to verify website domain::
  - # Authors with no sites: 4978
  - # Authors with Google Sites: 5993
  - # Authors with GitHub Sites: 359
  - # Authors with Weebly Sites: 195
  - # Authors with Wix Sites: 199
  - # Authors with Wordpress Sites: 15
  - # Authors with Personal Sites: 9158
  - # Authors with Academic Sites: 45928
  - # Authors with Other Sites: 18672
  - # websites reclassified after source code check: 27
- Coauthor names are not in a standardized format, so source code check for presence of name in the website source code is particularly helpful for these names. Again, this was done for google sites only. After this, there are 5096 authors with a google site match (8.5%)[22].

B.4 Domain Classification Criteria

- Google Sites:
  - a url in the format “https://sites.google.com/. . . ” is flagged as google site. Further, the site must have some permutation of the first name, last name (or middle name if present) in order to screen for related author google sites that come up in the search results.
  - Some google sites have custom urls (eg. “https://www.yeowhweechua.com/” is a google website). These urls fall in the “personal” category (defined below), and a source code check is done to determine whether the site is hosted on the google domain.
- Github Sites:
  - Github sites tend to consistently have a url of the above format. Unfortunately, there is nothing in the source code that uniquely identifies a github site other than the url.
- Weebly Sites:
  - Of the format “https:/____.weebly.com/. . . ”. The url must also have some permutation of the name where underscored.

[22] Alternatively, matched each coauthor with the author affiliated with a top 50 institution with whom they coauthor and the name of the paper that they worked on and checked for matches with Top50 author name and a fuzzy match for paper name in the website in the research page instead. While this screen is more stringent, it performed similarly to searching for a match with coauthor name in the website source code.

  - Weebly site urls can be customized. In the source code, I have searched for “Powered by Weebly” or “weebly.com” in the “personal” and “other” sites to categorize them as a weebly site.

- Wix Sites:
  - Of the format “https://____.wixsite.com/...” or “https://____.wix.com/...”. The url must also have some permutation of the name where underscored.
  - Wix site urls can be customized. In the source code, I have searched for “Wix.com Website builder” in the “personal” and “other” sites to categorize them as a wix site.

- Wordpress Sites:
  - Of the format “https://____.wordpress.com/...”. The url must also have some permutation of the name where underscored.
  - Wordpress site urls can be customized. In the source code, I have searched for “wordpress.com” in the “personal” and “other” sites to categorize them as a wix site. This is a very weak check, but doesn’t seem like there is anything else in the source code that is clearly unique to a wordpress site.

- Personal Sites:
  - Of the format “https://____.(com—net—info)/...”. Where the underscored. Part is some variation of the author name. Includes common ones like “https://{firstname}{lastname}.(com—n obviously, but also sometimes relies on first 3 letters of first name or first 5 letters of last name for matching (since names like Alexander abbreviates to Alex, long Russian last names have shorter versions, etc).
  - The source codes of these websites are checked to categorize them as one of the above categories.

- Academic Sites:
  - Simply, a site with “.edu”. Also, foreign domains tend to be academic sites (“.fr”, “.uk”, “.de”, etc) usually with multiple matches per author. Sites for graduate students, alumni, student page profiles and websites that lead to pdfs are excluded.

- Other Sites:
  - Essentially all other sites, with a small caveat described below.
  - All websites that obviously do not host research have been removed. These include social media sites, media sites, blogs, websites of popularly affiliated companies, organizations with about pages but no research page etc. Also, any page that leads to a pdf or document.

B.5 Paper Title Deduplication

In the process of paper title deduplication, a deduplication run was executed using the matchit command to eliminate multiple copies of the same paper within a given author. This was particularly relevant for instances like “The Missing Profits of Nations” by Zucman, where variations in paper names (“The Missing Profits of Nations (Working Paper No. 24701)”, “i Zucman, G.(2018):“The Missing Profits of Nations””) resulted in separate line entries on the author’s Google Scholar page that the prior fuzzy merging algorithm failed to identify as the same paper. Additionally, to account for different cutoffs in the bigram-based fuzzy merging score, five new citation count variables were generated: citation_count_75, citation_count_80 (used as the baseline dependent variable), citation_count_85, citation_count_90, and citation_count_95. These variables correspond to varying thresholds in the bigram-based fuzzy merging score, where a higher score indicates a greater share of identical bigrams across any two paper titles being compared.

Paper Start Year

B.5.1 pub_year Cleaning

- Cleaned the publication year for papers that show pub_year $\leq$ 1975. This includes mostly inaccurately scraped pub_year or old scriptures wrongly classified to an author due to a common name (e.g., John Taylor).
- Cleaned up pub_year $\geq$ 2024 due to inaccurate scraping.
- Cleaned the irrelevant entries for those with length(paper_title) $\leq$ 10, which was mostly inaccurate scraping; not found in author’s Google Scholar’s page and no info on other details (title, publisher, coauthors, etc.), or non-economic papers falsely classified under the author’s name.
- Cleaning the irrelevant entries for those with length(paper_title) $>$ 11 & length(paper_title) $\leq$ 15 (similar reasons as above).

The pub_year variable underwent a series of cleaning procedures to address some concerns in the output scraped by the API. Firstly, entries with a pub_year $\leq$ 1975 were addressed. This subset primarily contained inaccurately scraped pub_year values or cases where old papers were erroneously linked to authors due to common names (e.g., “John Taylor”). The next step involved cleaning entries with pub_year $>$ 2023, targeting inaccuracies arising from flawed scraping processes. Further cleaning was carried out for entries with a paper title length of less than or equal to 10 characters, and a similar process was applied to those with titles ranging from 11 to 15 characters. In both cases, the decision to clean was motivated by the identification of wrongly classified papers, such as incoherent paper titles that were misclassified on the author’s Google Scholar page and have since been removed, as well as instances of non-economic papers falsely associated with the author’s name.

B.5.2 Min Year Cleaning

- Sieved out papers that either have: (i) a very early year of first citation; or (ii) have a year of first citation that precedes by many years the pub_year.
- Cleaned observations with min_year $\leq$ 1975.
- Dropped the non_econ papers.
- Fixed the citation_count for those that are falsely shown as having past citations, but don’t actually have them.

Fixed the publication year for those with incorrect publication year(papers/books that have multiple volumes or have been republished, or have had updated publications years later) With diff_year being defined as min_year-pub_year

Cleaned up the pub_year for papers with diff_year < -10. There are two main categories of noise that were cleaned out: (a) non-econ papers that are once again attributed to namesakes; or (b) instances where the citation count bar graph when scraped had an erroneous early year entry followed by a long series of zero citatons before the actual citation counts start. In the latter case, going back to Google Scholar now oftentimes shows that the early year noise is cleaned out subsequently by Google Scholar itself. For (a), we drop the paper. For (b), we drop the erroneous first year and regenerate the cumulative citations variable.

Additionally, papers with either an unusually early year of first citation or a first citation year significantly preceding the pub_year were filtered out. For entries with min_year less than or equal to 1975, a multi-step cleaning process was implemented. This included dropping non-economic papers, rectifying citation_count for cases falsely indicating past citations no longer reflected on the Google Scholar page, and correcting publication years for entries with inaccuracies due to multiple volumes, republishing, or updates. Entries with a difference between min_year and pub_year less than -10 underwent this further cleaning. Two primary categories of noise were addressed during the cleaning process: (a) non-economic papers incorrectly attributed to namesakes, leading to their exclusion; and (b) instances where the citation count bar graph, when initially scraped, contained an erroneous early year entry followed by a prolonged zero time series before the actual citation counts commenced. In the latter case,revisiting Google Scholar often revealed that the early year noise was subsequently cleaned out by Google Scholar itself. For category (a), the decision was made to drop the respective paper. In the case of category (b), the erroneous first year was excluded, and the cumulative citations variable was regenerated.