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

Univeristy of Chicago, November 1, 2024

Motivation

National restrictions on internet access are growing.

- designed primarily to restrict access to news, but their effects are more widespread.

China

- Internet restrictions were authorized in in 1997, begun in 1998, and continue to expand their reach
- The Great Firewall, started in 2006, relates primarily to cross-border traffic.

Is the flow of knowledge to China is being (perhaps unintentionally) affected by these internet restrictions?

Motivation

Great Firewall

- blocks Google news and search, google.com (2010)
  - "Guilt by Association"
    - all Google-related sites were blocked
    - including webpage hosting at **sites.google.com**.
- also used by China to restrict access to foreign information/research on China.
    - economics papers about China?

As a result, scholars in China, a large and rising share of the global total, are unable to easily access research posted on those pages.

Motivation

Great Firewall

- blocks Google news and search, google.com (2010)
  - "Guilt by Association"
    - all Google-related sites were blocked
    - including webpage hosting at **sites.google.com**.
- also used by China to restrict access to foreign information/research on China.
    - economics papers about China?

As a result, scholars in China, a large and rising share of the global total, are unable to easily access research posted on those pages.

Motivation

Great Firewall

- blocks Google news and search, google.com (2010)
  - "Guilt by Association"
    - all Google-related sites were blocked
    - including webpage hosting at **sites.google.com**.
- also used by China to restrict access to foreign information/research on China.
    - economics papers about China?

As a result, scholars in China, a large and rising share of the global total, are unable to easily access research posted on those pages.

This Project

This paper examines the effects of the Great Firewall on citations of academic economic research.

1. Does the implementation of the Great Firewall have an effect on the diffusion of knowledge in economics?
   - Are papers hosted on sites.google.com, Google Sites (GS), less likely to be cited?
2. Are papers about China less likely to be cited?
3. Does writing a China paper reduce cites on other papers by the same author?
   - The blocking technology usually works at the IP address rather than the paper.
4. Are Chinese authors less likely to cite GS papers, even after they have been published?

This Project

This paper examines the effects of the Great Firewall on citations of academic economic research.

1. Does the implementation of the Great Firewall have an effect on the diffusion of knowledge in economics?
   - Are papers hosted on sites.google.com, Google Sites (GS), less likely to be cited?
2. Are papers about China less likely to be cited?
3. Does writing a China paper reduce cites on other papers by the same author?
   - The blocking technology usually works at the IP address rather than the paper.
4. Are Chinese authors less likely to cite GS papers, even after they have been published?

The Problem Hits Close to Home

- 45 percent of the other presenters in the Chicago Trade and Spatial Afternoon Seminar Series from last spring and this fall have personal webpages on Google Sites.
- Regular Dartmouth international economics seminar attendees whose personal sites are blocked from view in China
  - Treb Allen
  - Matt Grant
  - Meredith Startz
  - Nathan Zorzi
- *Davin Chor* (huh? still? after working on this paper for years?)

The Problem Hits Close to Home

- 45 percent of the other presenters in the Chicago Trade and Spatial Afternoon Seminar Series from last spring and this fall have personal webpages on Google Sites.
- Regular Dartmouth international economics seminar attendees whose personal sites are blocked from view in China
  - Treb Allen
  - Matt Grant
  - Meredith Startz
  - Nathan Zorzi
- *Davin Chor* (huh? still? after working on this paper for years?)

Literature

- Great Firewall and information flows
  - Li et al. (2023) [export quality]; Kong et al. (2022) [innovation]; Zheng and Wang (2020) [patents]; Wang et al. (2022) [disclosure]
- Knowledge flows across space
  - Peri (2005); Belenzon and Schankerman (2013); Head et al. (2018); Sin (2018); Wuestman et al (2019); Ganguli et al. (2020); Bernard et al. (2023)
- China and global knowledge production
  - Xie and Freeman (2019); Qiu et al. (2022); Aghion et al. (2023); Economist (2024)

Timeline

- 1998
  - Golden Shield project initiated
    - China begins restricting the internet within the country and across borders.
- 2008
  - sites.google.com launched.
- 2009
  - Great Firewall activity increases. YouTube, Twitter, Facebook all blocked.
- 2010
  - Google hacked by Operation Aurora.
  - Google moves its news service to Hong Kong. google.hk is largely unavailable in China.
  - China retaliates, google.com blocked
  - Other Google-related sites are unavailable including sites.google.com.
- 2011-present
  - 2017 Cambridge University Press self-censors hundreds of published articles.
  - 2021 Springer Nature (Nature & Scientific American) self-censors 000's of articles.
  - Economics papers on China?

Data Construction

1. Assemble a list of academic staff at economics departments ranked in the top 50 by Tilburg University in 2019, **2991 staff**.
2. Retain those with Google Scholar pages, **1804 unique authors**.
3. All papers on the author profile pages scraped from Google Scholar, **145,590 papers**.
4. Deduplication, both within- and across-authors, yields English language papers in social sciences, **94,299 papers** have at least one citation.
5. 70,995 author names extracted from the papers, grouped where necessary using fuzzy matching algorithms yielding **61,439 author**s in total.
6. Scrape search results to find personal websites. **7,024 authors** had personal websites hosted on GS.

Sample

Starting from the broad set of 94,299 papers

- Papers with a start date 2000-2008, inclusive
- Citations by paper by year from 2004-2020
- **27,918 unique papers**

Google Sites

We measure GS usage at the paper level

- AtLeastOne = 1 if at least one author has a GS personal page
- 85% of papers with AtLeastOne=1 have just one GS author.

Summary Statistics

Papers

		N	Share	Mean	Median
Total Papers		94299			
Baseline Sample		27918			
Cites - flow	2009			9.24	3
	2019			11.44	1
Cites - total	2009			40.57	10
	2019			149.27	31
1+ author on the paper has a GS		5779	0.20		
Papers about China		434	0.02		
Other papers by China authors		9427	0.34		

Cumulative Citations

Table 1: Cumulative Citations and Google Sites, 2020

	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

	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

Google Site Usage and Selection

GS usage may not be random

- Authors at top departments less likely to use personal webpages to advertise papers.
  - 3 groups of papers based on maximum rank of the authors: 1-10, 11-20, and 21-50.
- MaxRank and AtLeastOne
  - ▶ 27.8% for rank 21-50
  - ▶ 20.3% for rank 11-20
  - ▶ 11.4% for rank 1-10
- Multi-author papers more likely to have at least one author with a GS personal page.
  - 4 groups of author team size: 1, 2, 3 , and 4+
- Author Number and AtLeastOne
  - ▶ 11.8% for 1 authors
  - ▶ 21.6% for 2 authors
  - ▶ 25.0% for 3 authors
  - ▶ 29.4% for 4+ authors

We include interactions of paperage fixed effects with MaxRank and NumAuths groups.

Baseline Specification

The analysis is done at the level of the paper (p).

Dependent var: Citations (annual flow) of a paper
The baseline specification is

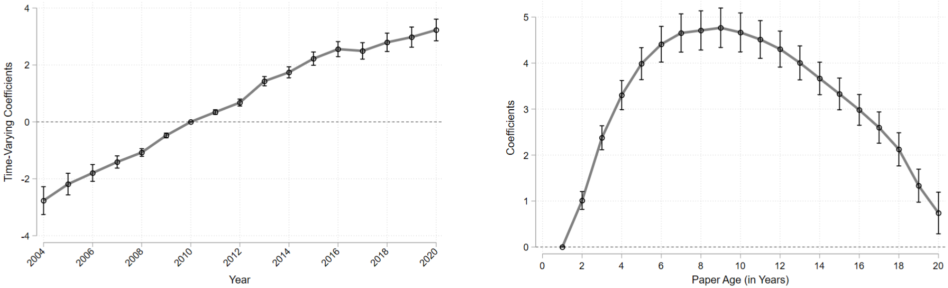
$$\begin{aligned}
C_{pt} &= \sum_t \beta_{Gt}(D_t \times \mathbf{I}(AtLeastOne_p)) + \\
&\quad + \sum_n \sum_a \beta_{na}\mathbf{I}(NumAuths_p = n)\mathbf{I}(PaperAge_{pt} = a) \\
&\quad + \sum_m \sum_a \beta_{ma}\mathbf{I}(MaxRank_p = n)\mathbf{I}(PaperAge_{pt} = a) \\
&\quad + D_t + D_p + \varepsilon_{pt}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

with $\beta_{G2010} = 0$, errors clustered at the paper level, including paper and year fixed effects.
$AtLeastOne_p = 1$ if at least one of the authors has a webpage hosted on GS in 2019.

Baseline

Trends, Paper Age

Figure 1: Baseline Specification

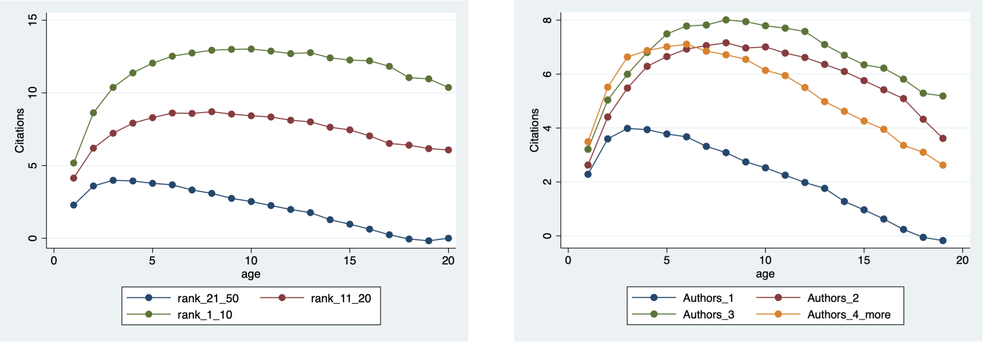


(a) Time

(b) Paper Age

Paper citations rise over time. Citations rise, then fall, with paper age.

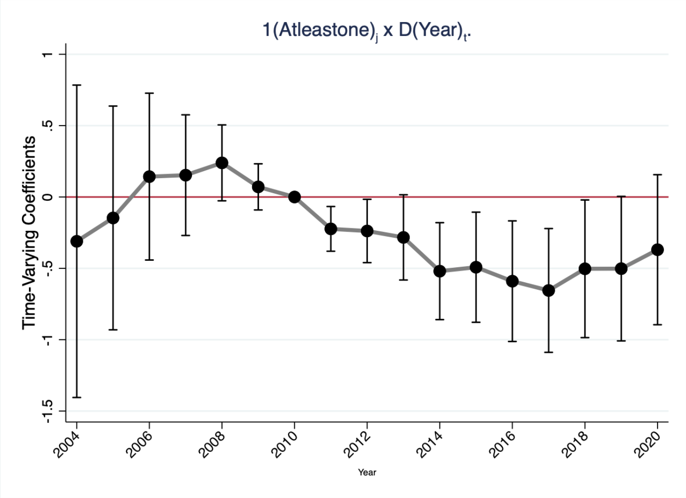
Paper Age by Group



Lifecycle of paper citations varies by maxrank and authornumber group.

Google Sites Effects

Effects of At Least One Author Hosting on GoogleSites



Papers on GS have 4.3 fewer citations after 2010.

Papers on China

- China does not just block large domains: Google Facebook, NYTimes, etc
- Sites containing information deemed unacceptable may also be blocked.
- There is no direct evidence that economics papers about China are deemed unacceptable.
- We enhance the baseline specification to include an indicator if a paper has China in the title or abstract but is not on GS, $ChinaRef_p$, interacted with year dummies.

China Paper Specification

Dependent var: Citations (annual flow) of a paper

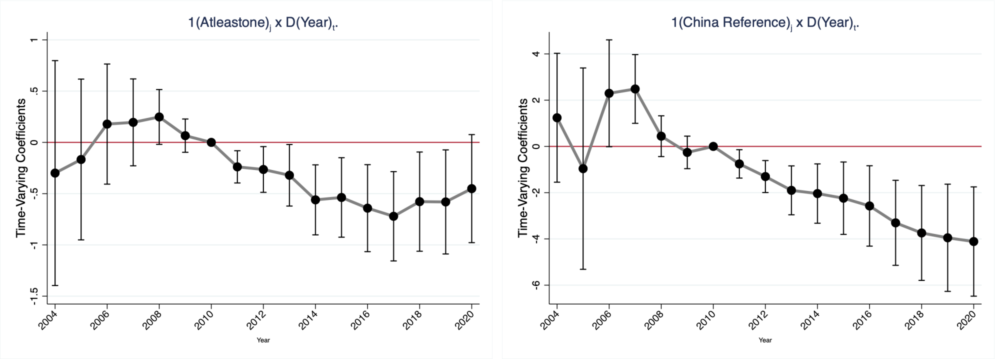
$$\begin{aligned} C_{pt} \quad = \quad & \sum_t \beta_{Gt}(D_t \times \mathbf{I}(AtLeastOne_p)) + \sum_t \beta_{Ct}(D_t \times \mathbf{I}(ChinaRef_p)) + \\ & + \mathit{Paper\ Age\ Interaction\ groups} + D_t + D_p + \varepsilon_{pt}. \end{aligned}$$

$AtLeastOne_p = 1$ if at least one of the authors has a webpage hosted on GS in 2019.

$ChinaRef_p = 1$ if the paper has China in the title or abstract and is not on GS.

Baseline + China

Effects on Papers that Reference China



Papers about China are cited less after 2010.

GS effect is slightly larger (4.5), smaller se.s.

Baseline Specification + China Paper + China Author

Dependent var: Citations (annual flow) of a paper

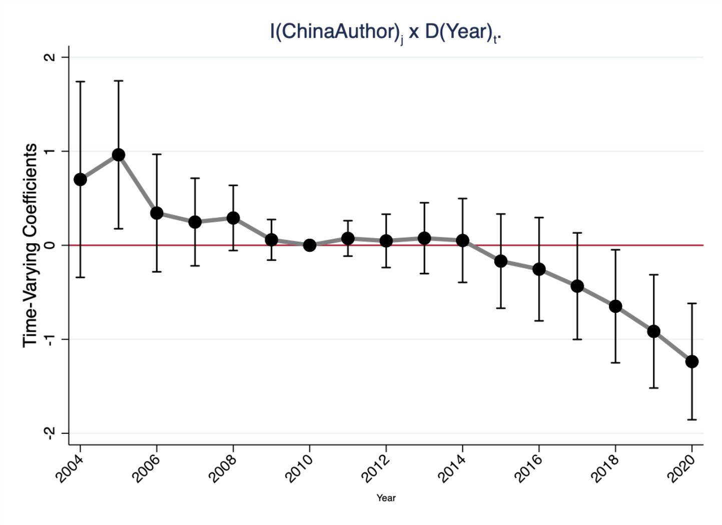
$$\begin{aligned} C_{pt} \quad = \quad & \sum_t \beta_{Gt}(D_t \times \mathbf{I}(AtLeastOne_p)) + \sum_t \beta_{Ct}(D_t \times \mathbf{I}(ChinaRef_p)) \\ & + \sum_t \beta_{CAt}(D_t \times \mathbf{I}(ChinaAuthor_p)) + \\ & + \textit{Paper Age Interaction groups} + D_t + D_p + \varepsilon_{pt}. \end{aligned}$$

$AtLeastOne_p = 1$ if at least one of the authors has a webpage hosted on GS in 2019.

$ChinaRef_p = 1$ if the paper has China in the title or abstract and is not on GS.

$ChinaAuthor_p = 1$ if the author has another paper on China but this paper does not have China in the title or abstract and is not on GS.

Baseline + China Paper + China Author



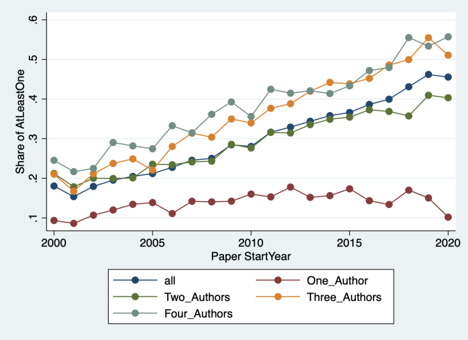
Authors writing about China are cited less in their other papers.

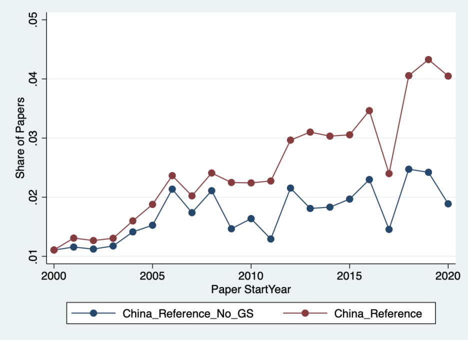
Growing Over Time

The importance of this (lack of) knowledge flow is rising over time.

- More researchers are using Google Sites.
- More papers are being written about China.
-
-

Growing Importance of GS and China Papers in Recent Years





Among papers with a Top 50 author:

Share of papers hosted on GS rising rapidly.

Share of papers on China also rising.

Citing Data

Web of Science - *published* papers citing other *published* papers

- We scrape Web of Science for every published paper from 2000 to 2020 that cites a published paper in our original sample by Top 50 authors.

- Previous baseline sample from Google Scholar included all papers starting in 2000-2008 with at least one citation from 2004-2020 had **27,918** cited papers.
- Using these same criteria, Web of Science (WoS) yields **13,507** papers.

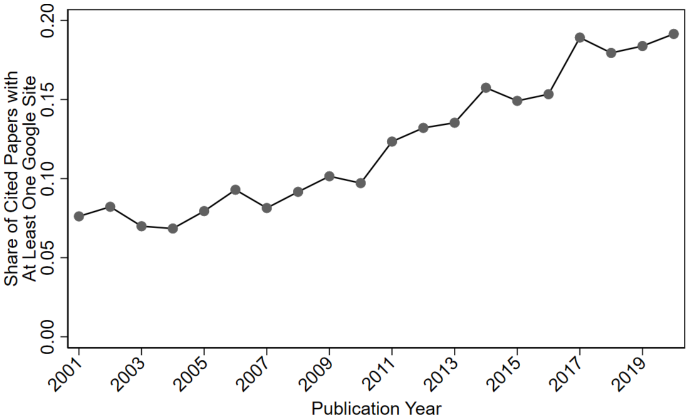
- 236,198 citing papers with pub date from 2016-2020 in Business/Economics journals.
- By construction, every citing paper cites at least one paper by a Top 50 author.
- 25,797 Top 50 papers are cited; 3,277 GS papers are cited

Growing Over Time

The importance of this (lack of) knowledge flow is rising over time.

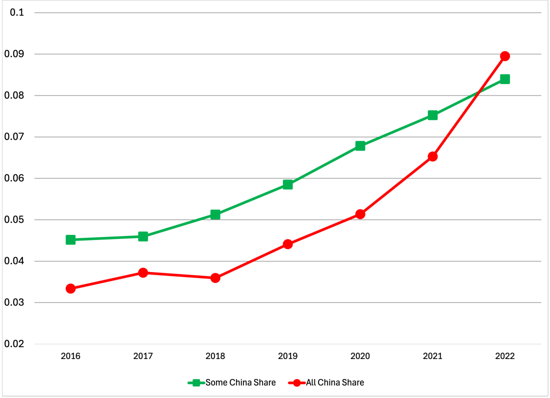
- More researchers are using Google Sites.
- More papers are being written about China.
- More published papers are being written by researchers in China.
- Share of cited published papers with at least one author using GS is rising.

Growing Importance of GS Papers on Web of Science (published papers)



GS papers are increasing as a share of cited published papers by top 50 authors.

Growing Importance of China-based Authors on Web of Science (published papers)



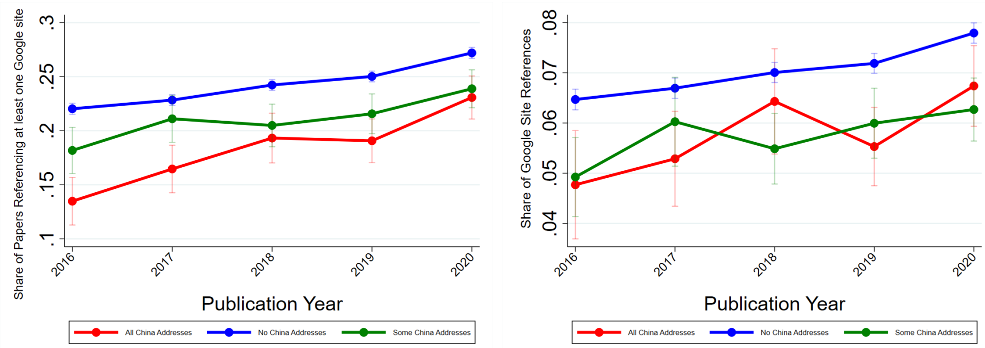
Published papers by China-Based authors are an increasing share of an increasing total.

Citations by China-based Authors

Web of Science - *published* papers citing other *published* papers

- Two measure of GS citation frequency
  - Average share of cited GS papers out of all Top 50 papers cited.
  - Share of papers that cite at least one GS paper.
- Three groups of papers
  - No China-based authors.
  - Some China-based authors.
  - All China-based authors.

Google Sites and Citations by Authors in China


At Least One Citation

Average Citation Share

China-based authors are 17% less likely to cite a GS paper than authors outside China.

Share of GS paper cites is 13% lower for China-based authors.

Conclusions

Internet restrictions by China are limiting the flow of knowledge to Chinese researchers.

- The Great Firewall blocks websites hosted by Google (and others) after 2010.
- Citations are lower after 2010 for papers hosted on Google sites, and for papers about China
- Published papers by China-based authors are less likely to cite published papers by authors who have a Google site.

Further Questions

What might mitigate these effects? NBER, CEPR affiliations
Which cited researchers are hurt most? less connected; younger; women