

Do self-archiving platforms enhance research impact?

Evidence from bioRxiv

Liu, Hongxu

Tongji University

Hu, Guangyuan

Shanghai University of Finance and Economics

Li, Yin

Fudan University (yinli@fudan.edu.cn)

Presenter: Liu, Hongxu

2022-10-08

Research Background

- Self-archiving platforms → increased dissemination of new research
- Depositing preprints →? enhanced research impact (citations)
- Potential selection bias: researchers deposit higher quality papers?
- Causal link between submitting preprints online **and** increased research impact → **unclear**



Research Background

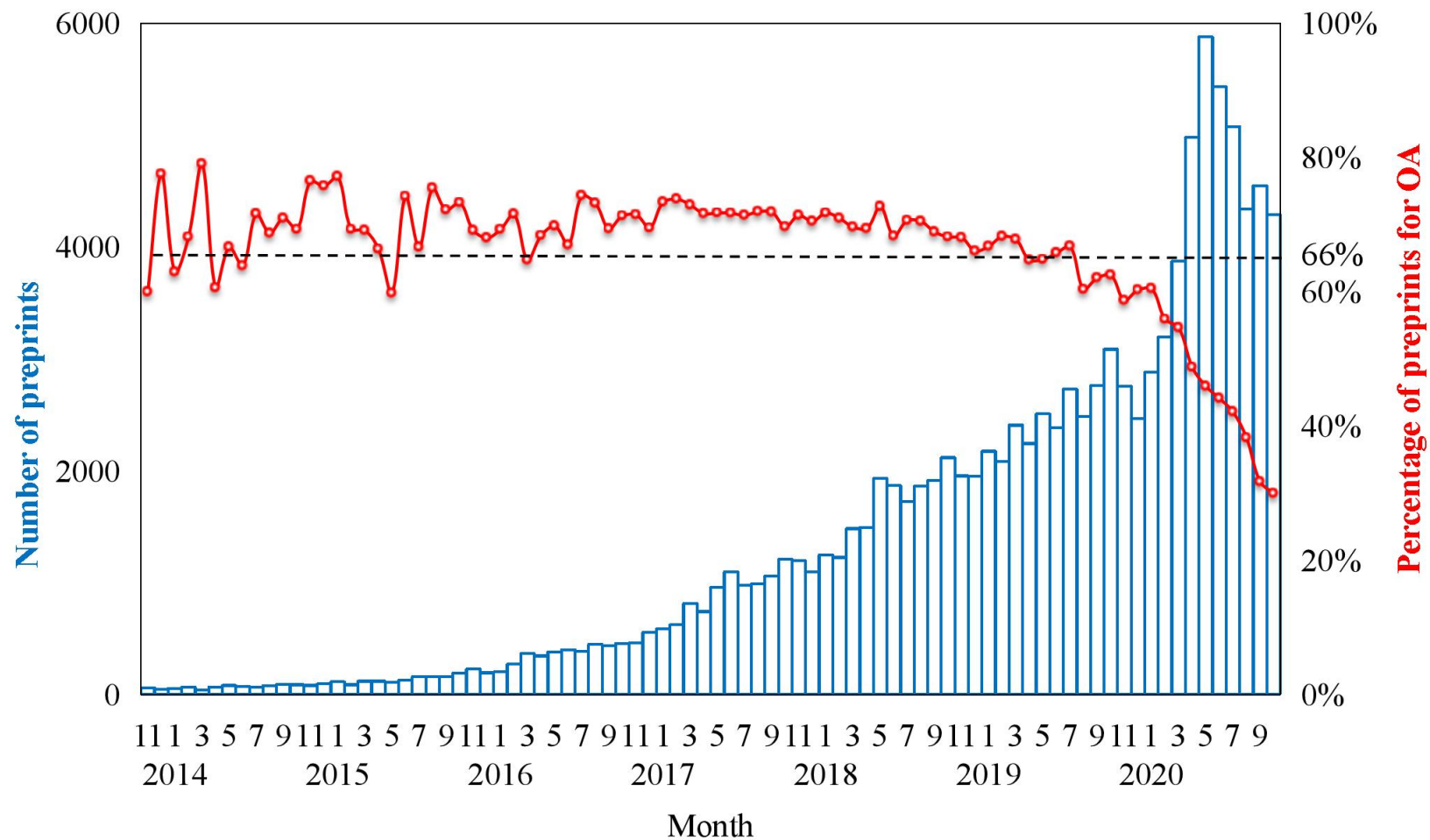


Fig 1 Number of preprints submitted and percentage of preprints published on the bioRxiv platform

What We Do

- Examine the causal link between preprint platforms and research impacts
 - ✓ treatment group: 5,423 published articles with preprints deposited on bioRxiv in 2018.
- Using text-mining algorithm for matching similar articles
 - ✓ control group: 7,862 similar articles without depositing preprints.
- Modeling the differences in citations in subsequent years
 - ✓ establish a positive causal effect of self-archiving platforms on forward citations.

Independent Variable	preprint = 1 (treatment), or 0 (control)
	Inpubcited
	Inpubcited(t)
Dependent Variable	Inpubcited(t+1)
	Inpubcited(t+2)
	Inpubcited(t+3)
	OA or not
	International collaboration or not
	First author from a top 100 university or not
	Corresponding author from a top 100 university or not
	Number of authors
	Number of countries
	Country of first author
	Country of corresponding author
Control Variable	Online year
	Preprint citations
	Preprint Mendeley reads
	Preprint Blog mentions
	Preprint Twitter mentions
	Abstract monthly reads (Abstracti、Abstract6all、Abstract12all、Abstractall)
	Full-text HTML monthly reads (Full-text HTMLi、Full-text HTML6all、Full-text HTML12all、Full-text HTMLall)
	PDF monthly reads (PDFi、PDF6all、PDF12all、PDFall)

for treatment and control groups

for treatment group

Descriptive Statistics

VARIABLES	Preprint=0					Preprint=1				
	N	mean	sd	min	max	N	mean	sd	min	max
Number of authors	7,862	6.077	9.874	1	436	5,423	8.364	13.75	1	390
First author from a top 100 university or not	7,862	0.225	0.418	0	1	5,423	0.311	0.463	0	1
Corresponding author from a top 100 university or not	7,862	0.220	0.414	0	1	5,423	0.302	0.459	0	1
OA or not	7,862	0.795	0.403	0	1	5,423	0.999	0.0304	0	1
International collaboration or not	7,862	0.328	0.470	0	1	5,423	0.448	0.497	0	1
Number of countries	7,862	1.566	1.275	1	32	5,423	1.821	1.543	1	30

Sampling: built a dataset of research articles with preprints matched with similar articles published in the same year.

- **Step 1:** draw a sample of research articles with preprints on bioRxiv.
 - ✓ **treatment group:** 5,434 articles archived on bioRxiv in 2018 and published in peer-reviewed journals.
- **Step 2:** Use a text-mining algorithm, the PubMed Related Citations Algorithm (PMRA), to identify the most similar research articles.
 - ✓ a method developed by the National Library of Medicine to identify the neighbors of a document, or the most similar ones in the database, by measuring the similarity between documents through the words they have in common (Kim et al., 2001).

$$\eta(\mu/\lambda) = P(E^-) \mu / P(E) \lambda = 1 \quad w_t = (1 + (\mu/\lambda)^{(k-1)} e^{-(\mu-\lambda)t})^{-1} \sqrt{\text{idf}_t}$$

Method

- **Step 3:** Restrict the control group to neighboring papers of the treated sample in PubMed that are published in the same year but without a preprint.
- **Step 4:** Use the PSM radius matching method to eliminate the 168 samples that were “Off support”.
- **Control group:** 7,862 research articles.
- Bibliometrics information: Web of Science.
- Altmetrics information: Dimensions citations, Mendeley reads, Blog mentions, and Twitter mentions.

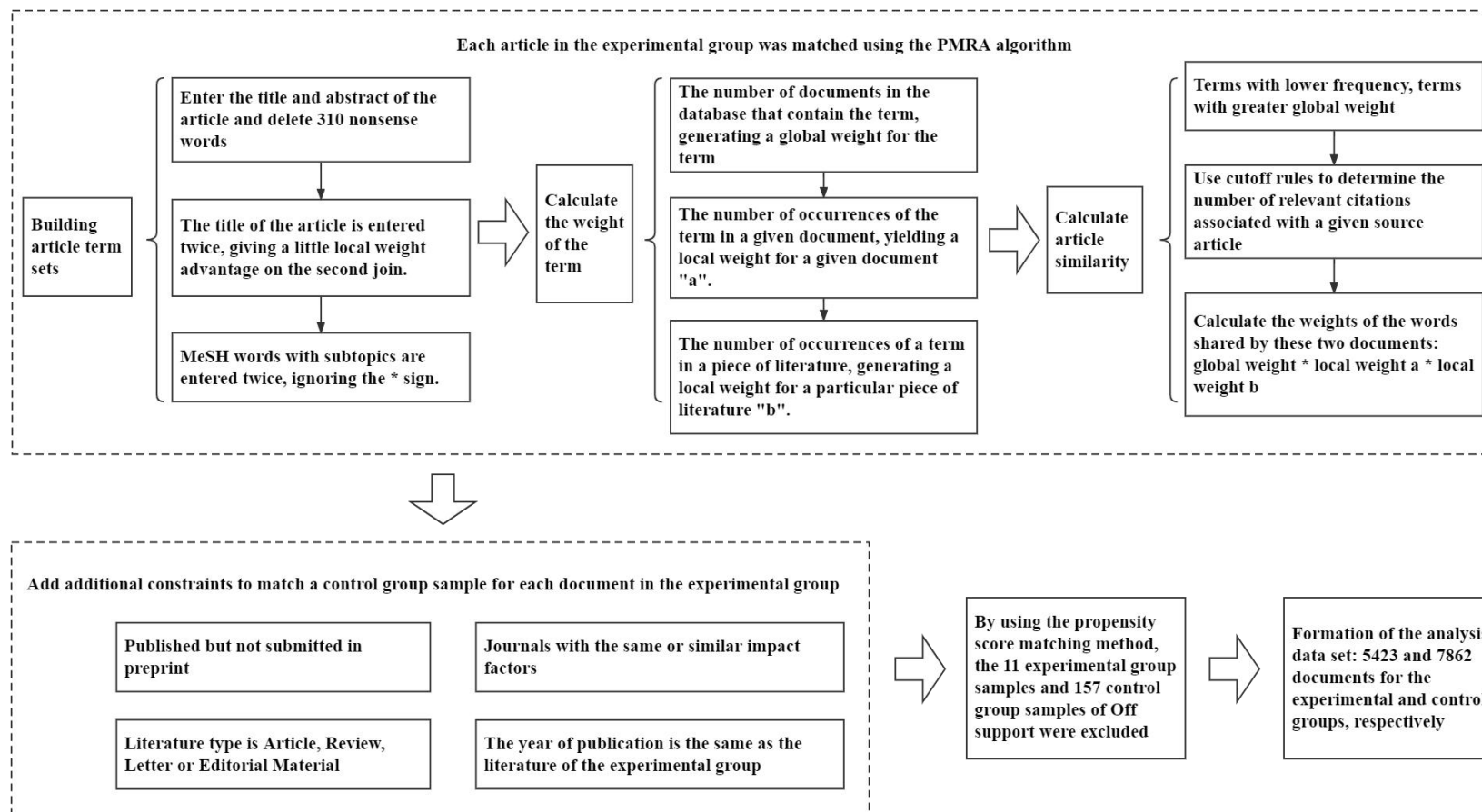


Fig 2: Flow chart of constructing the control group with the PMRA algorithm

- We use an econometric model to estimate the effects of preprints on forward citations

$$\ln(Y_{ij}) = \beta \text{Preprint}_i + \gamma Z_i + S_j + \varepsilon_{ij}, \quad (3)$$

- $\ln(Y_{ij})$ is the natural logarithm of the number of forward citations
- $\text{Preprint}_{ij} = 1$ for treated articles, or 0 for control articles
- Unit of analysis is a paper, for all articles similar to paper ij clustered in group j
- Z_i is a vector of control variables for author level and paper level characteristics
- S_j is a paper group fixed effects
- The equation is estimated with a fixed effect OLS model

Findings 1: Positive impact of preprints on forward citations

VARIABLES	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Inpubcited	Inpubcited	Inpubcited	Inpubcited	Inpubcited	Inpubcited	Inpubcited
preprint	0.322*** (0.0188)	0.322*** (0.0183)	0.302*** (0.0198)	0.289*** (0.0180)	0.290*** (0.0185)	0.294*** (0.0180)	0.187*** (0.0191)
OA or not			0.100*** (0.0337)				0.00592 (0.0341)
International collaboration or not				0.274*** (0.0204)			0.178*** (0.0305)
Number of authors					0.0139*** (0.00259)		0.0132*** (0.00299)
Number of countries						0.111*** (0.00986)	0.00804 (0.0144)
Constant	2.298*** (0.0134)	2.298*** (0.0149)	2.219*** (0.0303)	2.208*** (0.0166)	2.214*** (0.0213)	2.125*** (0.0215)	2.239*** (0.0364)
Group fixed effect	No	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effect	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	13,285	13,285	13,285	13,285	13,285	13,285	12,930
R-squared	0.020	0.020	0.021	0.034	0.041	0.039	0.048

Robust standard errors in parentheses *** p<0.01, ** p<0.05, * p<0.1

Findings 2

- The positive effect diminishes over time.....

VARIABLES	(1) lnpubcited	(2) lnpubcited(t)	(3) lnpubcited(t+1)	(4) lnpubcited(t+2)	(5) lnpubcited(t+3)
preprint	0.187*** (0.0191)	0.190*** (0.0143)	0.0790*** (0.0169)	0.0496*** (0.0184)	0.0132 (0.0327)
OA or not	0.00592 (0.0341)	-0.0850*** (0.0222)	0.0377 (0.0287)	0.0662** (0.0314)	0.168*** (0.0507)
International collaboration or not	0.178*** (0.0305)	0.0742*** (0.0219)	0.146*** (0.0279)	0.154*** (0.0293)	0.163*** (0.0483)
Number of authors	0.0132*** (0.00299)	0.00660*** (0.00163)	0.0115*** (0.00282)	0.0110*** (0.00259)	0.0198* (0.0111)
Number of countries	0.00804 (0.0144)	0.0196* (0.0109)	0.0202 (0.0138)	0.0112 (0.0138)	-0.0226 (0.0253)
Constant	2.239*** (0.0364)	0.573*** (0.0242)	1.328*** (0.0313)	1.550*** (0.0341)	1.500*** (0.0605)
Group fixed effect	Yes	Yes	Yes	Yes	Yes
Country fixed effect	Yes	Yes	Yes	Yes	Yes
Observations	12,930	12,930	12,930	11,870	4,318
R-squared	0.048	0.043	0.045	0.034	0.038

Findings 2: The impact of preprints on citations over time

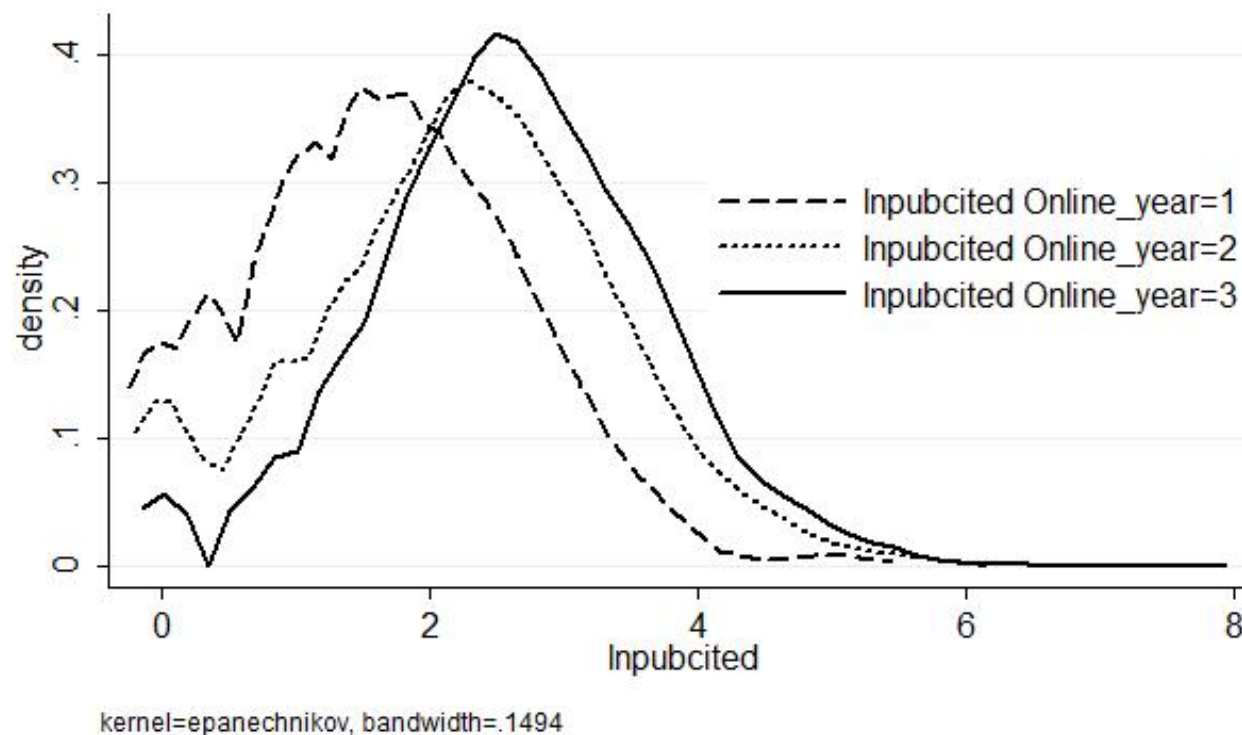


Fig 3: Citation distributions of the treatment and control groups for 1, 2 , 3 years appearing in public

Findings 3: Mechanisms

- How do preprints enhance citations?
- Dissemination via online social networks
- Positive effect on citations
 - ✓ Citations to the preprints
 - ✓ Mendeley reads
 - ✓ Twitter mentions

VARIABLES	(1) lnpubcited	(2) lnpubcited	(3) lnpubcited	(4) lnpubcited	(5) lnpubcited	(6) lnpubcited
preprint	0.0918*** (0.0252)	0.0627** (0.0300)	0.0632** (0.0285)	0.0596** (0.0262)	0.0596** (0.0262)	0.0653** (0.0283)
Preprint citations	0.0549*** (0.0104)	0.0535*** (0.00981)	0.0531*** (0.00995)	0.0525*** (0.0103)	0.0525*** (0.0103)	0.0609*** (0.00909)
Preprint Mendeley reads	0.00675*** (0.00234)	0.00578** (0.00231)	0.00551** (0.00228)	0.00594*** (0.00216)	0.00594*** (0.00216)	0.00585*** (0.00196)
Preprint Blog mentions	0.0487 (0.0373)	0.0358 (0.0358)	0.0294 (0.0362)	0.0202 (0.0372)	0.0202 (0.0372)	0.0252 (0.0369)
Preprint Twitter mentions	0.00247*** (0.000724)	0.00145 (0.000923)	0.00163* (0.000851)	0.00193** (0.000758)	0.00193** (0.000758)	0.00114 (0.000884)
Abstract _{full}		6.71e-05 (4.55e-05)				3.87e-05 (9.24e-05)
Full-text HTML _{full}		9.66e-06 (0.000129)				0.000219 (0.000559)
PDF _{full}		-1.89e-05 (6.18e-05)				3.85e-05 (0.000158)
Abstract _{12all}			6.07e-05* (3.10e-05)			-1.65e-05 (8.38e-05)
Full-text HTML _{12all}			-5.43e-05 (0.000171)			-0.000260 (0.000537)
PDF _{12all}			-3.21e-05 (3.94e-05)			0.000200 (0.000161)
Abstract _{all}				4.52e-05*** (1.51e-05)	4.52e-05*** (1.51e-05)	3.46e-05* (2.02e-05)
Full-text HTML _{all}				7.09e-06 (4.30e-05)	7.09e-06 (4.30e-05)	3.17e-05 (6.42e-05)
PDF _{all}				-3.68e-05 (2.41e-05)	-3.68e-05 (2.41e-05)	-0.000178*** (6.29e-05)
Constant	2.298*** (0.0149)	2.298*** (0.0149)	2.298*** (0.0149)	2.298*** (0.0149)	2.298*** (0.0149)	2.298*** (0.0149)
Group fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Country fixed effect	Yes	Yes	Yes	Yes	Yes	Yes
Observations	13,285	13,285	13,285	13,285	13,285	13,285
R-squared	0.083	0.085	0.085	0.085	0.085	0.086

Conclusions

- We apply a **novel method** based on a text-mining algorithm, the PMRA, to match the sample in the treatment group, allowing us to estimate causal effects from observational and bibliometrics data.
- We show the **causal effect** of depositing on bioRxiv on citations is smaller than those in previous studies (18.7% vs. 36%) (Fraser et al., 2020).
- The positive effect on citations **diminishes over time**.
- **Online visibility** might have contributed to the positive effect of self-archiving platforms.
- Lessons for researchers:
 - 1) Deposit preprints online;
 - 2) Increasing online visibility;
 - 3) Hope for the best (*effect is larger than zero, but not as big as previously thought*).

Limitations

- bioRxiv only
- Depositions on more than one preprint platform *not considered*
- Revisions after depositing a preprint online *not considered*
- *Limited* time window for citations to accumulate
- *Limited* data on author-level factors (e.g., gender, age, reputations)
- *Limited* data on journal-level factors (e.g., journal impact factors)

Reference

- Akbaritabar, A., Stephen, D. & Squazzoni, F. (2022). A study of referencing changes in preprint-publication pairs across multiple fields, *Journal of Informetrics*, 16(2), 101258.
- Buehling, K, Geissler, M, & Strecker, D (2022). Free access to scientific literature and its influence on the publishing activity in developing countries: The effect of Sci-Hub in the field of mathematics. *Journal of the Association for Information Science and Technology*.10.1002/asi.24636.
- Fang, Z., Dudek, J., & Costas, R.. (2020). The stability of twitter metrics: a study on unavailable twitter mentions of scientific publications. *Journal of the Association for Information Science and Technology*,71(12), 1455-1469.
- Fraser, N., Momeni, F., Mayr, P., & Peters, I. (2020). The relationship between bioRxiv preprints, citations and altmetrics. *Quantitative Science Studies*, 1(2), 618-638.
- Kim W, Aronson AR. & Wilbur WJ. (2001). Automatic MeSH term assignment and quality assessment, *Proceedings of the 2001 AMIA Symposium*, 319-323.
- Kwon, D. (2020). How preprint servers are blocking bad coronavirus research, *Nature*, 581(7807), 130.
- Moed, H F. (2007). The effect of "open access" on citation impact: an analysis of arxiv's condensed matter section, *Journal of the American Society for Information Science and Technology*, 58(13), 2047-2054.
- Rotolo, D., & Leydesdorff, L. (2015). Matching medline/pubmed data with web of science (wos): a routine in r language, *Journal of the Association for Information Science and Technology*, 66(10), 2155-2159.

Thanks !

Liu, Hong-xu

Tongji University

Hu, Guang-yuan

Shanghai University of Finance and Economics

Li, Yin

Fudan University (yinli@fudan.edu.cn)