

Portrait of Economic Academia: Evidence from Top 20 Economics Journals from 2012 to 2022

Collaboration Network & Topic Modeling Analysis



Netrunners
Jerry Cheng
Hongzhang Xie
Yicheng Zhang

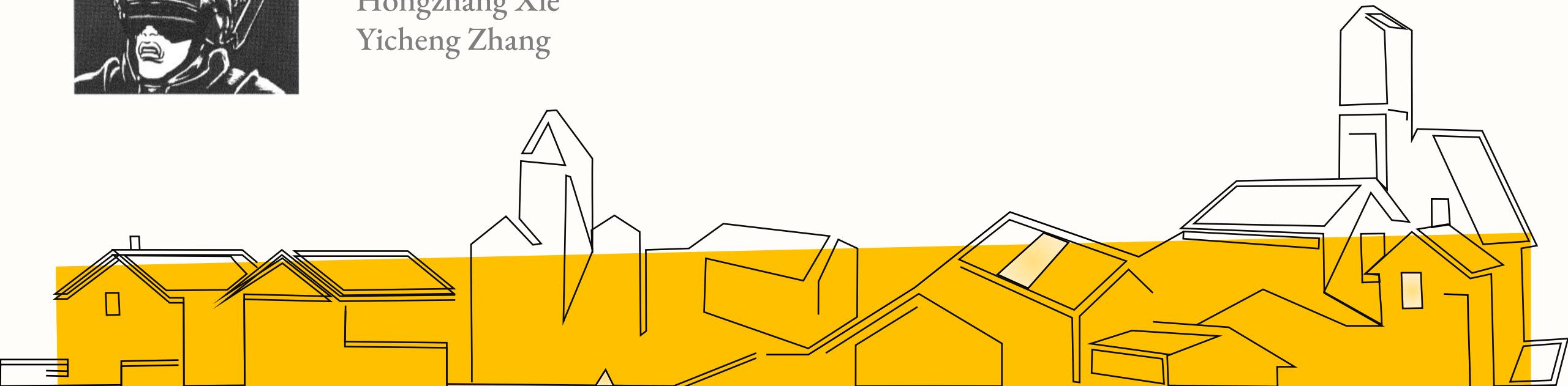
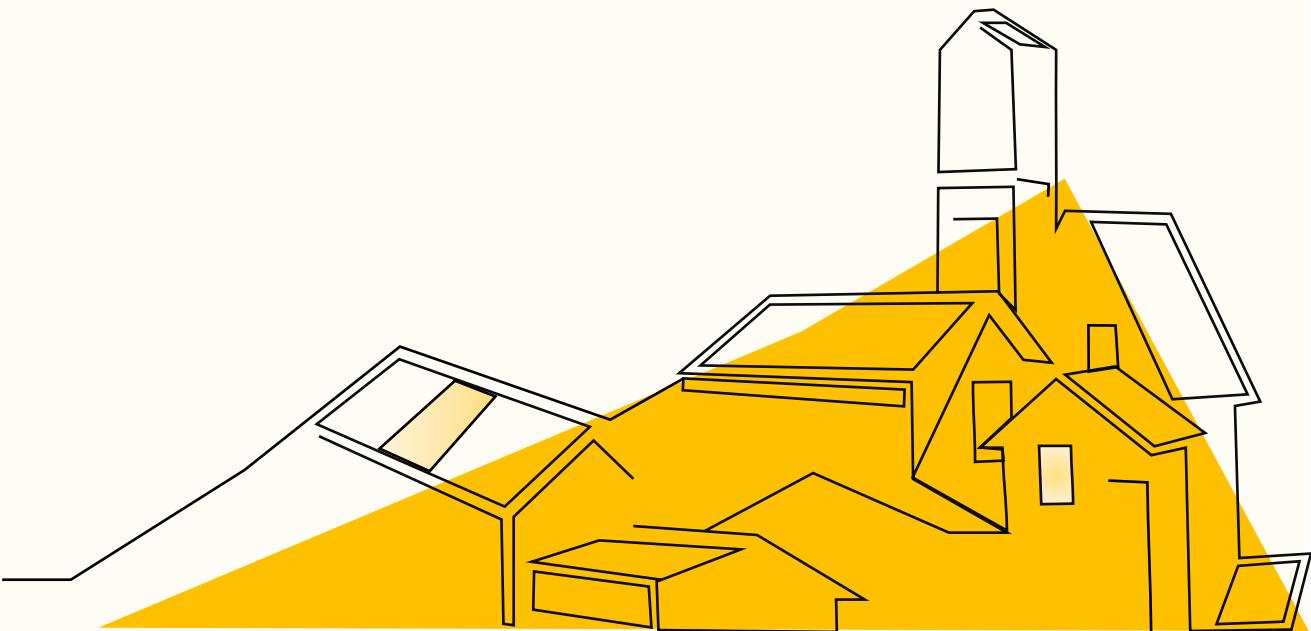


Table of Contents

1. Topic Overview
2. Research Questions and Social Science Relevance
3. Data Sources and Descriptions
4. Data Collection
5. Data Cleaning and Wrangling
6. Collaboration Network Analysis
7. Topic Modeling Analysis
8. Combining SNA and TM
9. Take-home Points



Research topic:

Examine the characters of the economic research collaboration and topics

Question 1

What are the network features and collaboration patterns at the institutional and tier level? Density, centrality, cross-rank freedom?

Question 2

How research topics vary and evolve across time and institutions?
For each topic, who are the most important contributors?

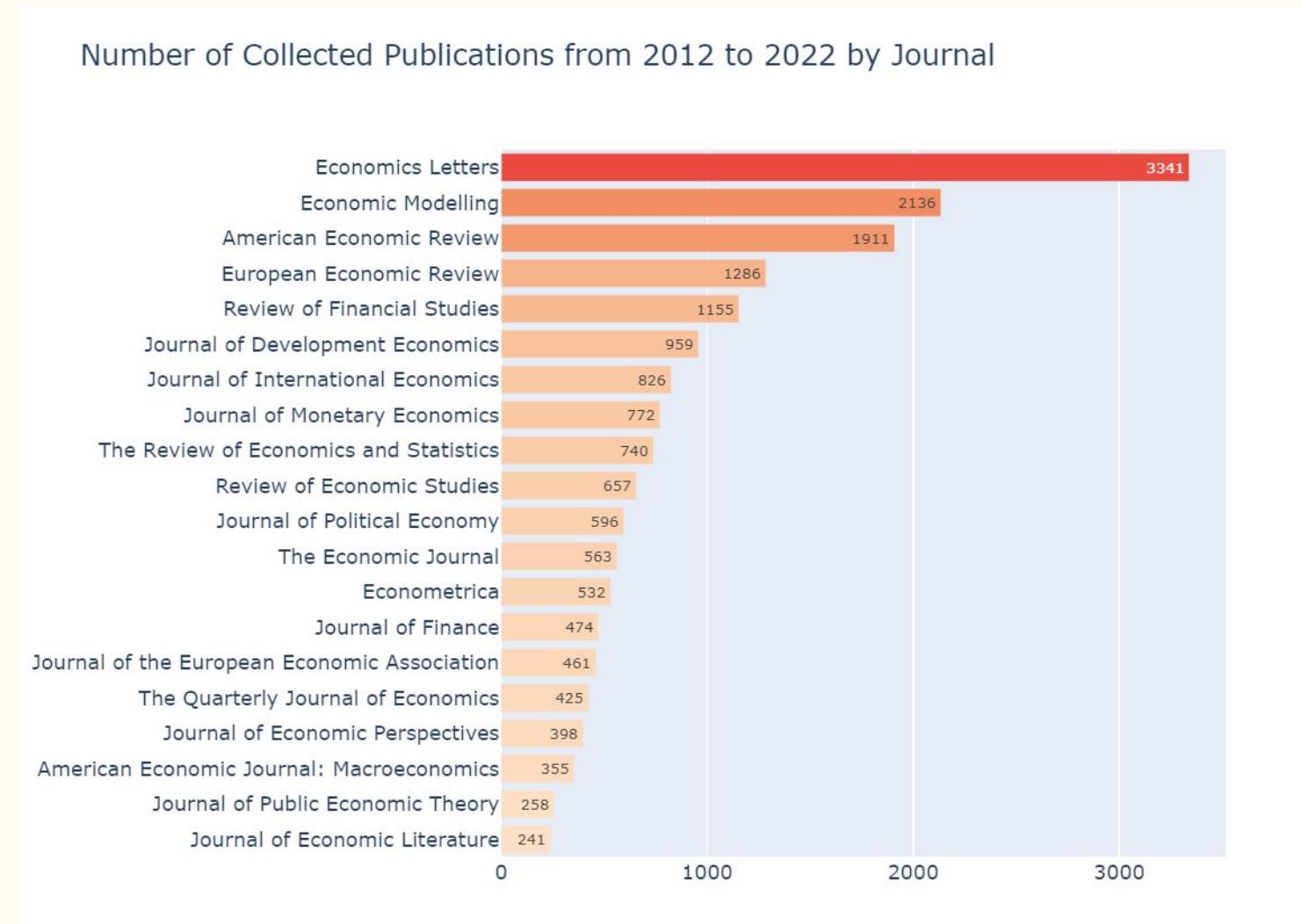
Question 3

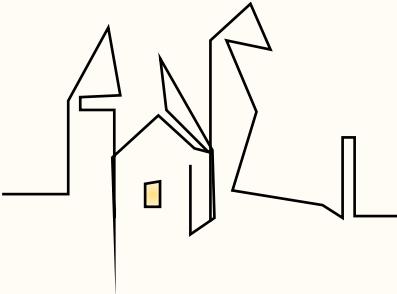
Do centralities in the network research the same topics? For a centrality, who are parallel peers, who are distinct peers?



Data Description -Top 20 Economics Journals Data

- We collected the data of top 20 economics journals ranked by h5-index (2017 – 2021).
- Our raw data have 79,565 publications from 1886 till now.
- We select 18,086 publications from 2012 to 2022 as the research sample.
- We have three databases:
 - Top 20 Economics Journals Data
 - Google Scholar Author Data
 - Authors' Affiliations Data





Data Description

Top 20 Economics Journals Data

Data sources

Official Websites:

- American Economic Association
- Oxford Academic
- Science Direct

Third-party database :

- ideas.repec.org
- econpapers.repec.org

Title ————— **Retail Pharmacies and Drug Diversion during the Opioid Epidemic**

Authors ————— **Aljoscha Janssen
Xuan Zhang**

Journals ————— **AMERICAN ECONOMIC REVIEW
VOL. 113, NO. 1, JANUARY 2023
(pp. 1-33)**

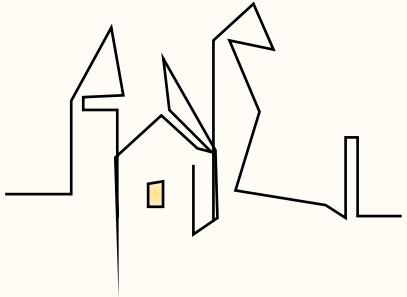
VOL ————— **Download Full Text PDF**

Published time ————— **Article Information**

Abstract ————— **This study investigates the role of retail pharmacy ownership in the opioid epidemic. Using data of prescription opioid orders, we show that compared with chain pharmacies, independent pharmacies dispense 39.1 percent more opioids and 60.5 percent more OxyContin. After an independent pharmacy becomes a chain pharmacy, opioid dispensing decreases. Using the OxyContin reformulation, which reduced nonmedical demand but not the legitimate medical demand, we show that at least one-third of the difference in the amount of OxyContin dispensed can be attributed to nonmedical demand. We show that differences in competitive pressure and whether pharmacists own the pharmacy drive our estimates.**

Citation ————— **Citation
Janssen, Aljoscha, and Xuan Zhang. 2023. "Retail Pharmacies and Drug Diversion during the Opioid Epidemic." *American Economic Review*, 113 (1): 1-33.**

DOI ————— **DOI: 10.1257/aer.20210357**



Data Description

Google Scholar Author Data

Links of two databases

- We input the **DOI** in **Top 20 Economics Journals Database** to search the **Authors** of the publications in Google Scholar.

Data sources

- Google Scholar Author Database
- Semantics Scholar Database
(For incompleteness data)

The screenshot shows a Google Scholar profile for 'Zhiguo He (何治国)'. The profile includes a photo, the name 'Zhiguo He (何治国)', and affiliations: 'Fuji Bank and Heller Professor of Finance, University of Chicago, NBER'. It also shows a verified email at 'chicagobooth.edu' and links to 'Homepage', 'Financial Economics', 'Chinese Economy and Fin...', and 'FinTech'. The profile has 9695 citations, an h-index of 35, and an i10-index of 45. A bar chart shows 'Cited by' counts from 2016 to 2023. Publications listed include 'Intermediary asset pricing' (2013), 'Blockchain Disruption and Smart Contracts' (2017), 'Rollover risk and credit risk' (2012), and 'Intermediary asset pricing: New evidence from many asset classes' (2017). A 'Public access' section is also visible.

DOI

Author id

Name

Affiliation*

Email

H-index

I10-index

Citations

* If the author has **Multiple Affiliations**, we use the **Affiliation of the Email**.

Data Description - Google Scholar Author Data

- Our raw data have 19,585 records of authors
- We select 13,654 authors who have publications from 2012 to 2022 as the research sample.

Number of Collected Publications from 2012 to 2022 by Authors



Data Sources and Description - Authors' Affiliations Data

Data sources

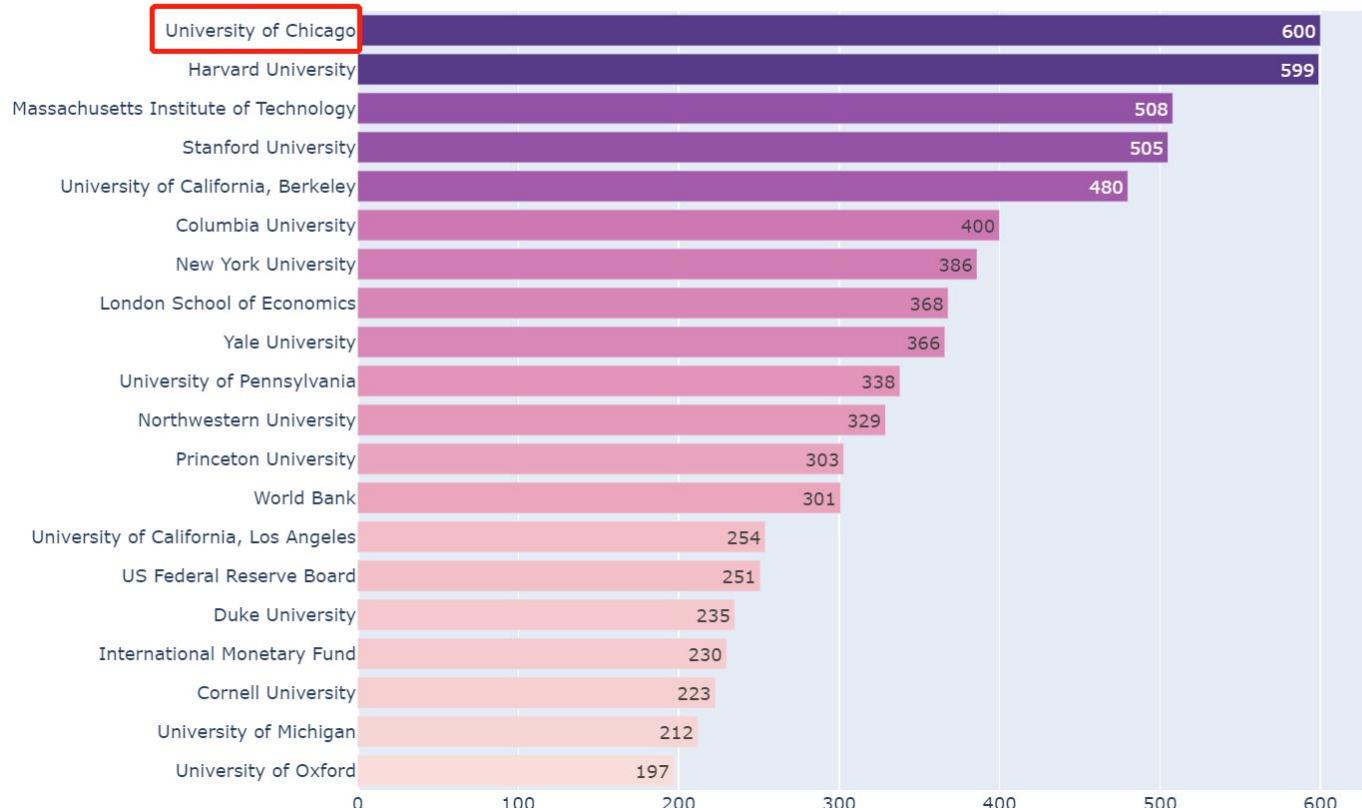
- Google Scholar Institutions Information

A screenshot of a Google Scholar search results page. The search query is "chicagobooth.edu". The results show a scholar profile with about 8,900 results found in 0.03 seconds. Below the scholar profile, it says "User profiles for: University of Chicago - uchicago.edu".

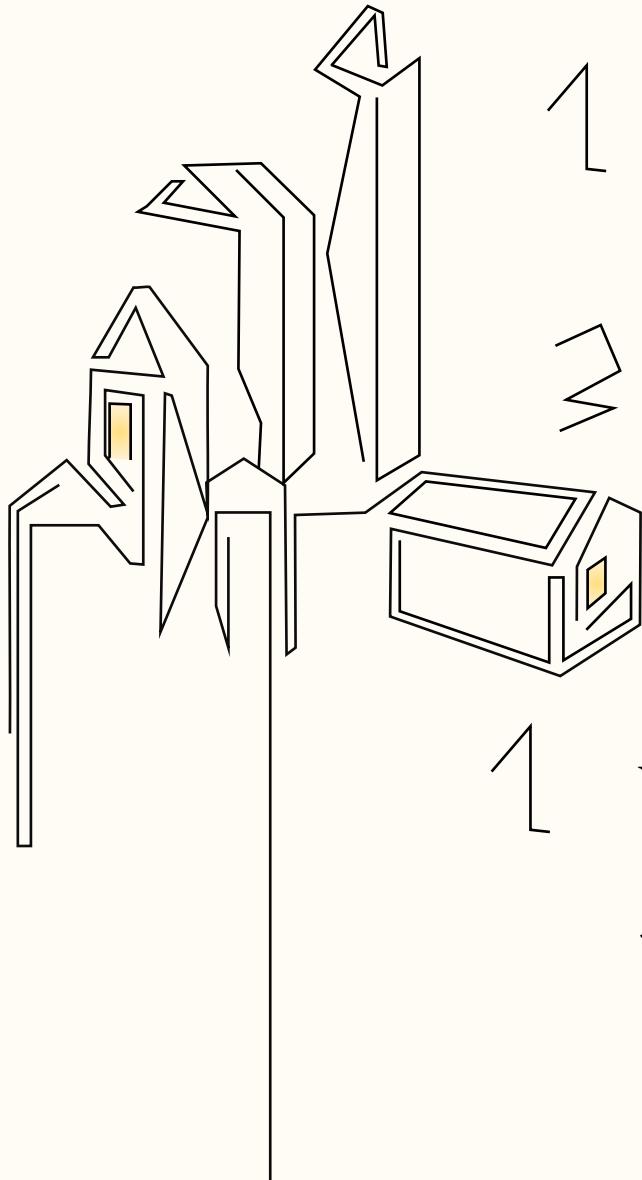
Data description:

- Our raw data have 1,392 records of institutions.
- We select 1,329 institutions that have authors who publish Top articles from 2012 to 2022 as the research sample.

Number of Collected Publications from 2012 to 2022 by Institutions



Advantages



1 Large sample size with multiple attributes.

Innovation, few studies have used similar data sources.

2 High data quality,
Crawl data from authoritative databases

3 Importance, analyzing those data can help us better understanding the research collaboration mechanisms, and improve efficiency.

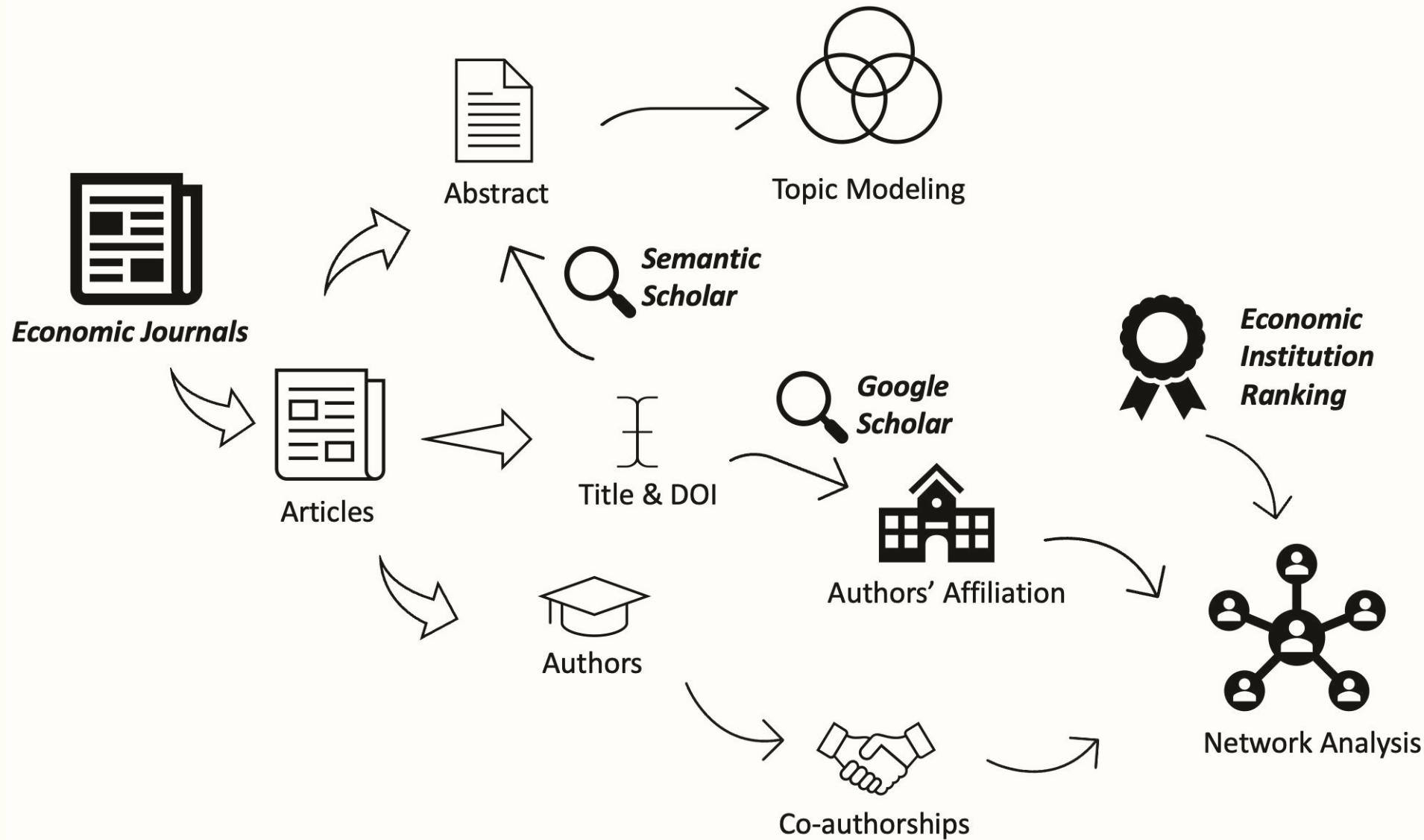
Limitations

1 We do not analyze the publication information outside the top 20 economic journals.

2 There are still some missing data in Publish time, DOI, and Authors in the publications before 2000.

3 We did not crawl the data for the authors who do not have Google Scholar profiles.

Data Collection Roadmap



Building Authors Database

(1) Retrieve Authors' Google Scholar Page

- Iterate through the article database. For every article, search the article's DOI (if available) in Google Scholar.
- Since the DOI is unique for every article, the researching result is either one or none.
- Scholars that has their own Google Scholar page will have the links embedded in their names.
- Collected *all the embedded links* to revisit later.

The screenshot shows a Google Scholar search result. At the top, there is a menu icon and the DOI: 10.1257/000282802320188961. Below that is a 'Scholar' section with a blue diamond icon. The main result is titled 'Bank bailouts and aggregate liquidity' by DW Diamond, RG Rajan, published in the American Economic Review in 2002. A red box highlights the authors' names. The abstract discusses government bailouts of banks. Below the abstract are options to 'Save', 'Cite', 'Cited by 82', 'Related articles', and 'All 9 versions'. At the bottom, it says 'Showing the best result for this search. See all results'.

Building Authors Database

(2) Collect author's info from Google Scholar page

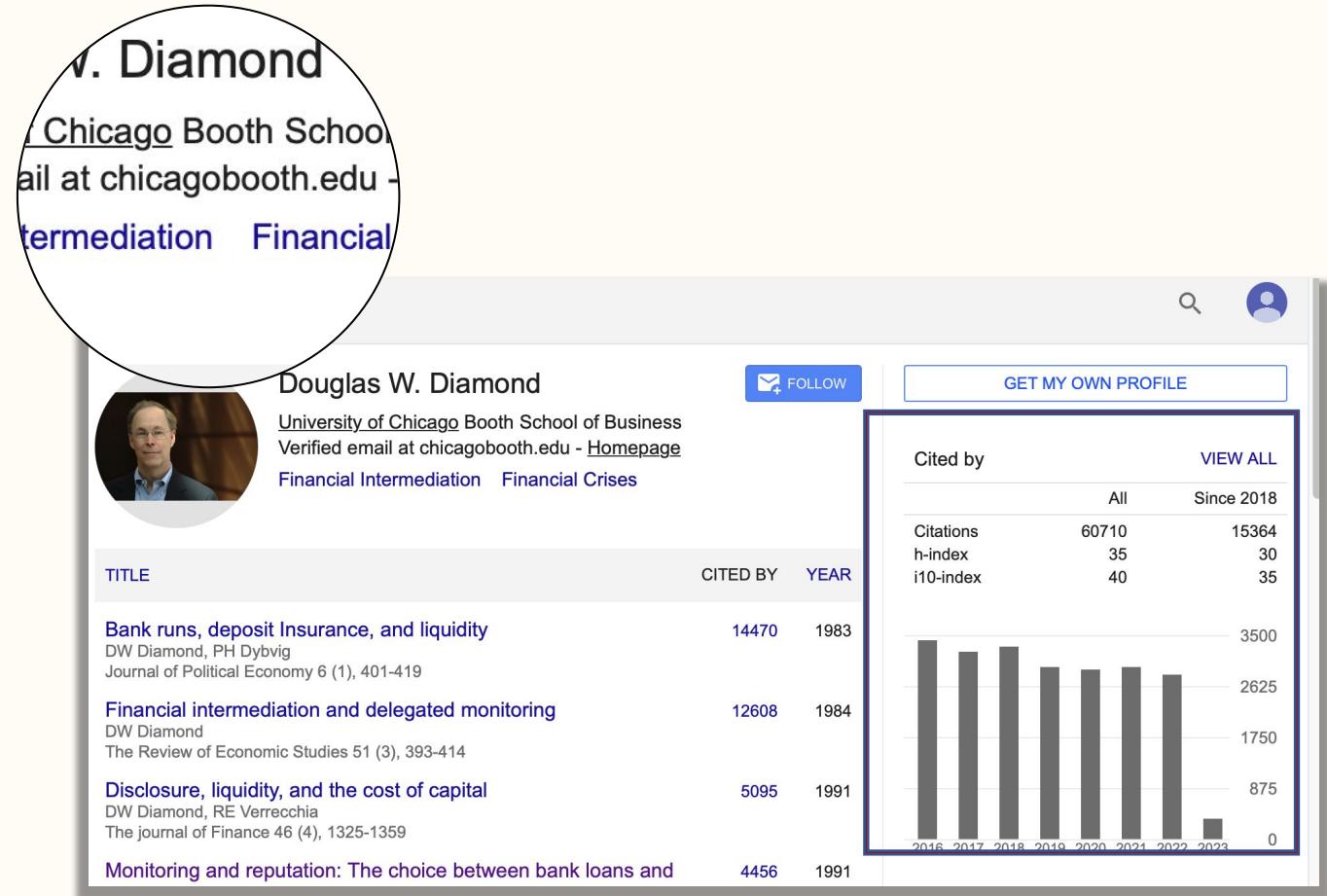
- Revisit the embedded URLs collected.
- Collected information includes
 - the *email address* for this scholar.
(Google Scholar affiliation section is messy)
 - Citations by year
 - h-index and i10-index listed
- Finally, search the email address collected directly in the Google Scholar.
- The affiliation match with this address will appear in the top box.

chicagobooth.edu

Scholar About 8,900 results (0.03 sec)

User profiles for:

University of Chicago - uchicago.edu



Building Institution Database

Top 10% Economic Institutions (Last 10 Years Publications), as of January 2023 (with details)

[Explanations](#) [Ranking](#) [More](#)

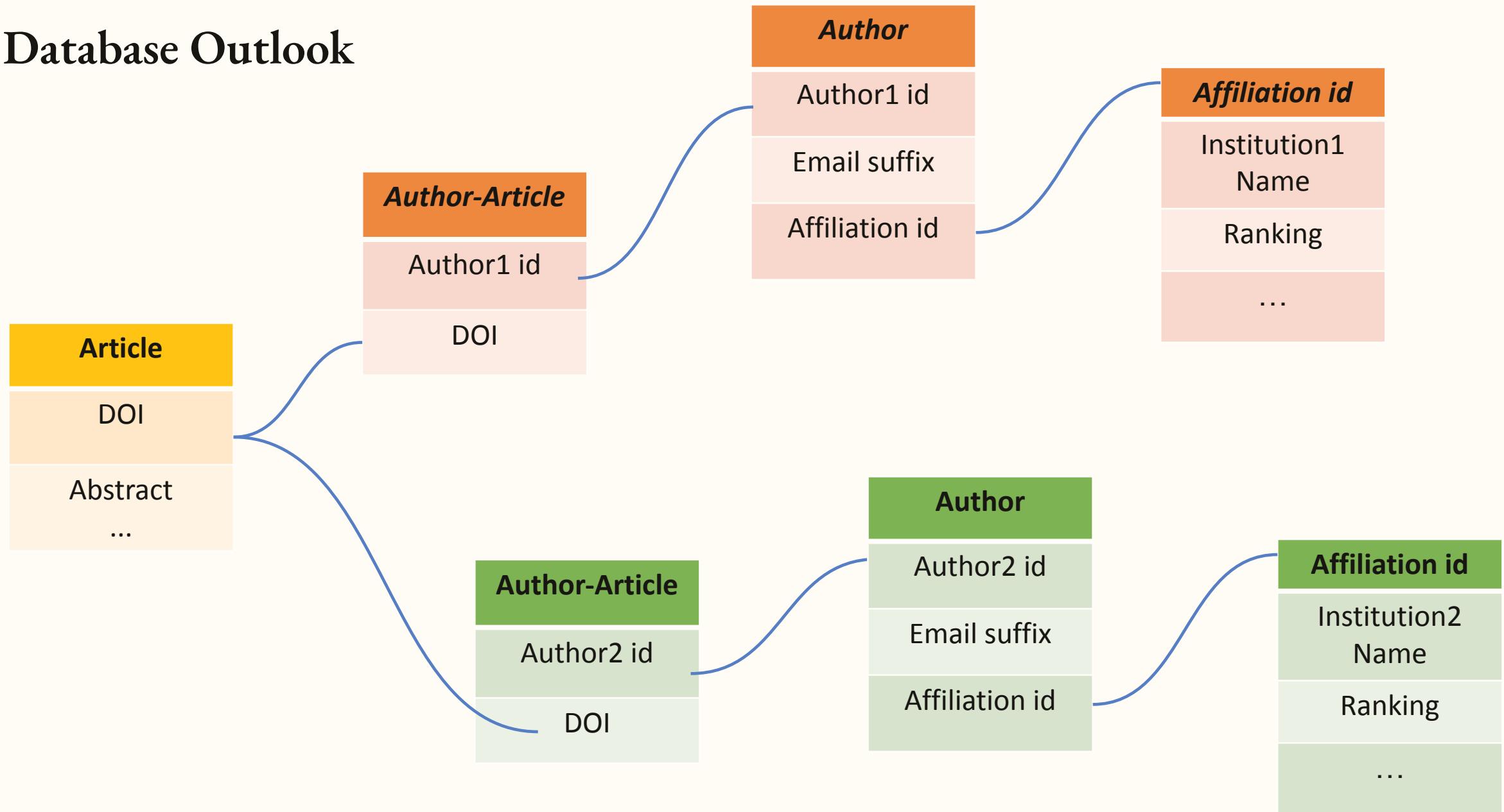
The rankings

Rank	Institution	Score	Nb Works	DNb Works	Sc Works	WSc Works	AI Works
			Score	Nb Works	DNb Works	Sc Works	WSc Works
1	London School of Economics (LSE)	2.04	3	5	1	1	
2	World Bank Group	3.25	2	1	2	4	
3	National Bureau of Economic Research (NBER)	3.77	20	26	4	2	
4	International Monetary Fund (IMF)	4.25	5	2	3	3	
5	Department of Economics, Harvard University	4.65	55	91	9	8	
6	European Central Bank	6.74	8	11	6	11	
7	Department of Economics, University of California-Berkeley	8.35	39	77	10	10	
8	Economics Department, Massachusetts Institute of Technology (MIT)	9.56	48	65	11	9	
9	Department of Economics, University of Chicago	9.81	37	61	8	6	
10	Federal Reserve Board (Board of Governors of the Federal Reserve System)	10.02	18	18	7	7	

- IDEAS ranking
 - Academic Oriented
 - Comprehensive ranking that include non-university institutions
- Full record linkage
 - Common word is removed (“university of”)
 - Take the highest score
 - Exactly 200 institutions were matched

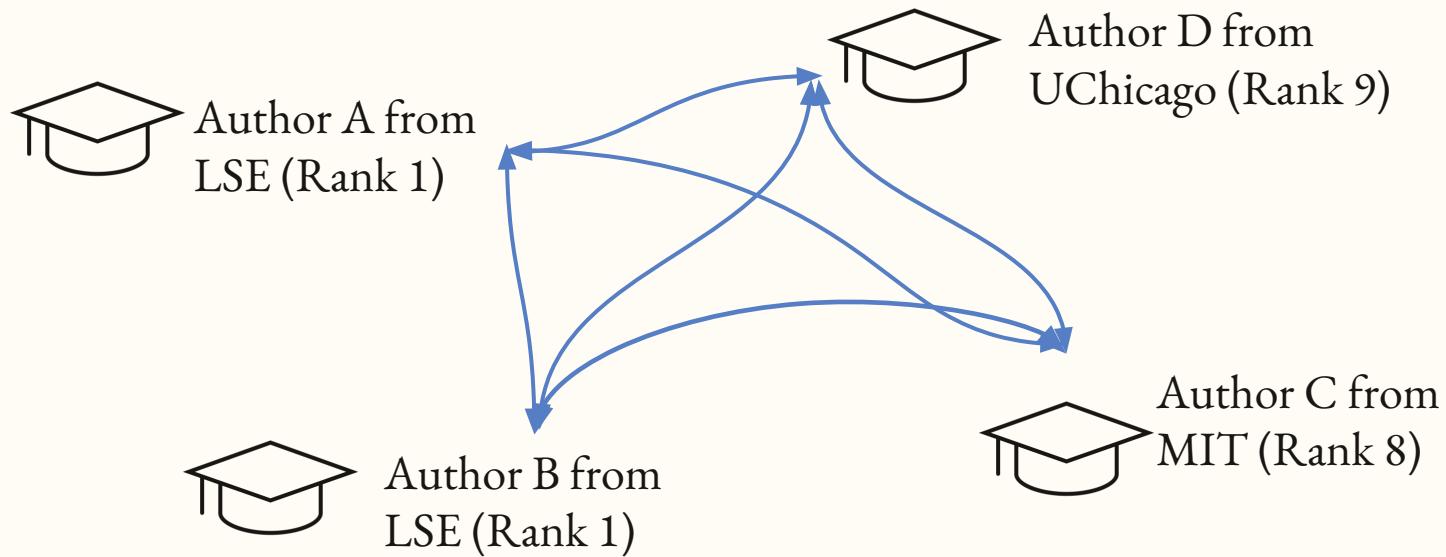
affiliationid	name	email	match_...	univer_ins	Rank
au.aalto.fi	Aalto University	aalto.fi	0.7962962...	Carleton University	309
kriot.kth.se	KTH Royal Institu...	kth.se	0.7470084...	Beijing Institute of ...	78
uocla.ucla.edu	University of Calif...	ucla.edu	0.682900...	Boston College	100
aei.aei.org	American Enterpr...	aei.org	0.7599879...	Economic and Soci...	212
miot.mit.edu	Massachusetts In...	mit.edu	1	Massachusetts Inst...	8
ciot.caltech.edu	California Institut...	caltech.edu	0.8677729...	Beijing Institute of ...	78
hu.harvard.edu	Harvard University	harvard.edu	1	Harvard University	5
su.stanford.edu	Stanford University	stanford.edu	1	Stanford University	14
uop.upenn.edu	University of Pen...	upenn.edu	1	University of Penns...	55
ai.amazon.com	Amazon Inc.	amazon.com	0.6931818...	Chapman University	448
au.american.edu	American Univers...	american.e...	0.746666...	Academia Romana	129
ac.amherst.edu	Amherst College	amherst.edu	0.8002011...	Dartmouth College	83
cmu.cmu.edu	Carnegie Mellon ...	cmu.edu	1	Carnegie Mellon Un...	360
anu.anu.edu.au	Australian Nation...	anu.edu.au	1	Australian National ...	99
au.ariel.ac.il	Ariel University	ariel.ac.il	0.7460317...	Drexel University	196
uoaa.arizona.edu	University of Ariz...	arizona.edu	0.7925925...	University of Alberta	416
asu.asu.edu	Arizona State Uni...	asu.edu	1	Arizona State Unive...	177
frboa.atl.frb.org	Federal Reserve ...	atl.frb.org	1	Federal Reserve Ba...	257
auob.aub.edu.lb	American Univers...	aub.edu.lb	0.666379...	Central Bank of Irel...	386

Database Outlook



Constructing Co-authorship links

Article – Author – Institution – Ranking

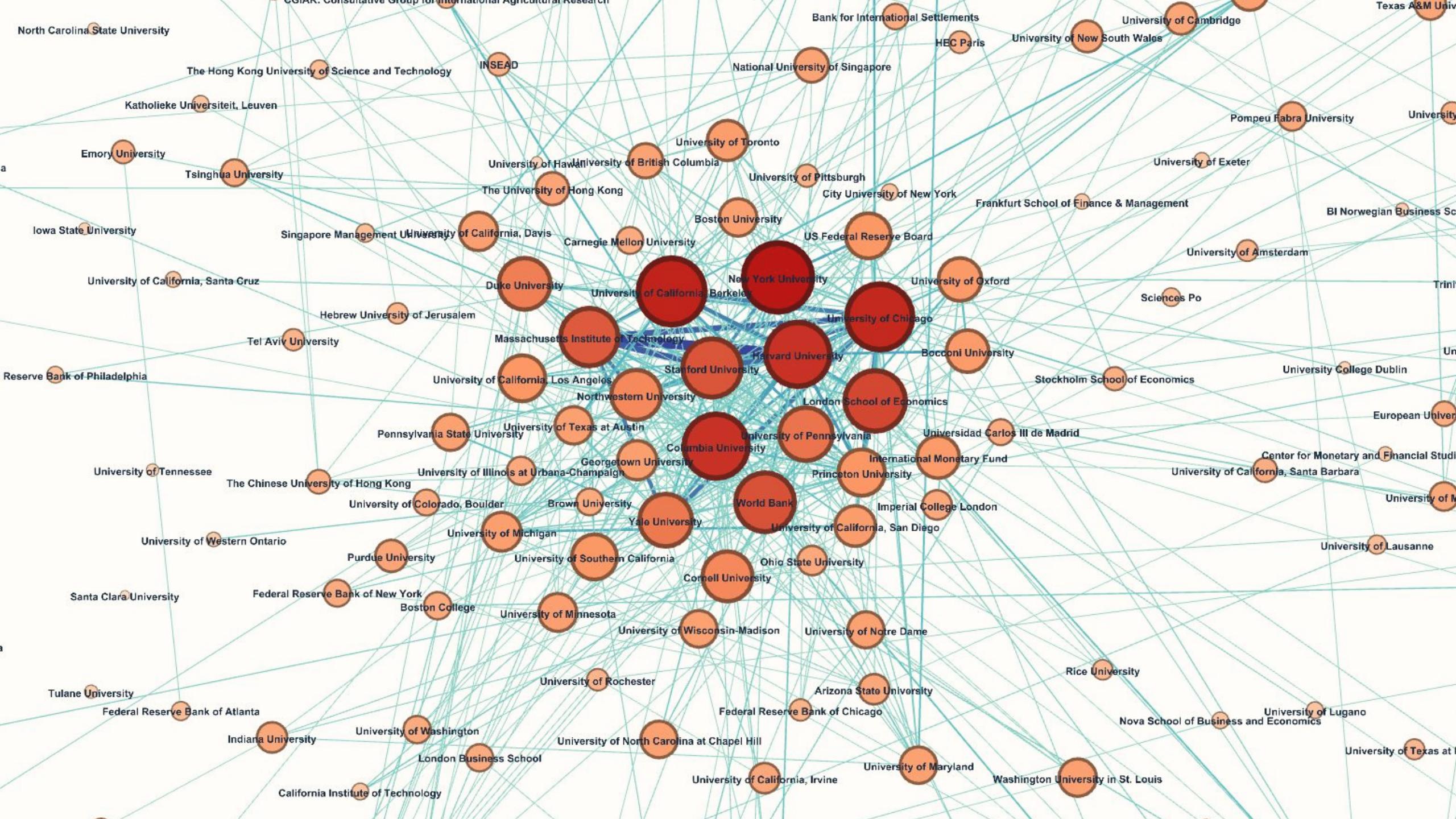


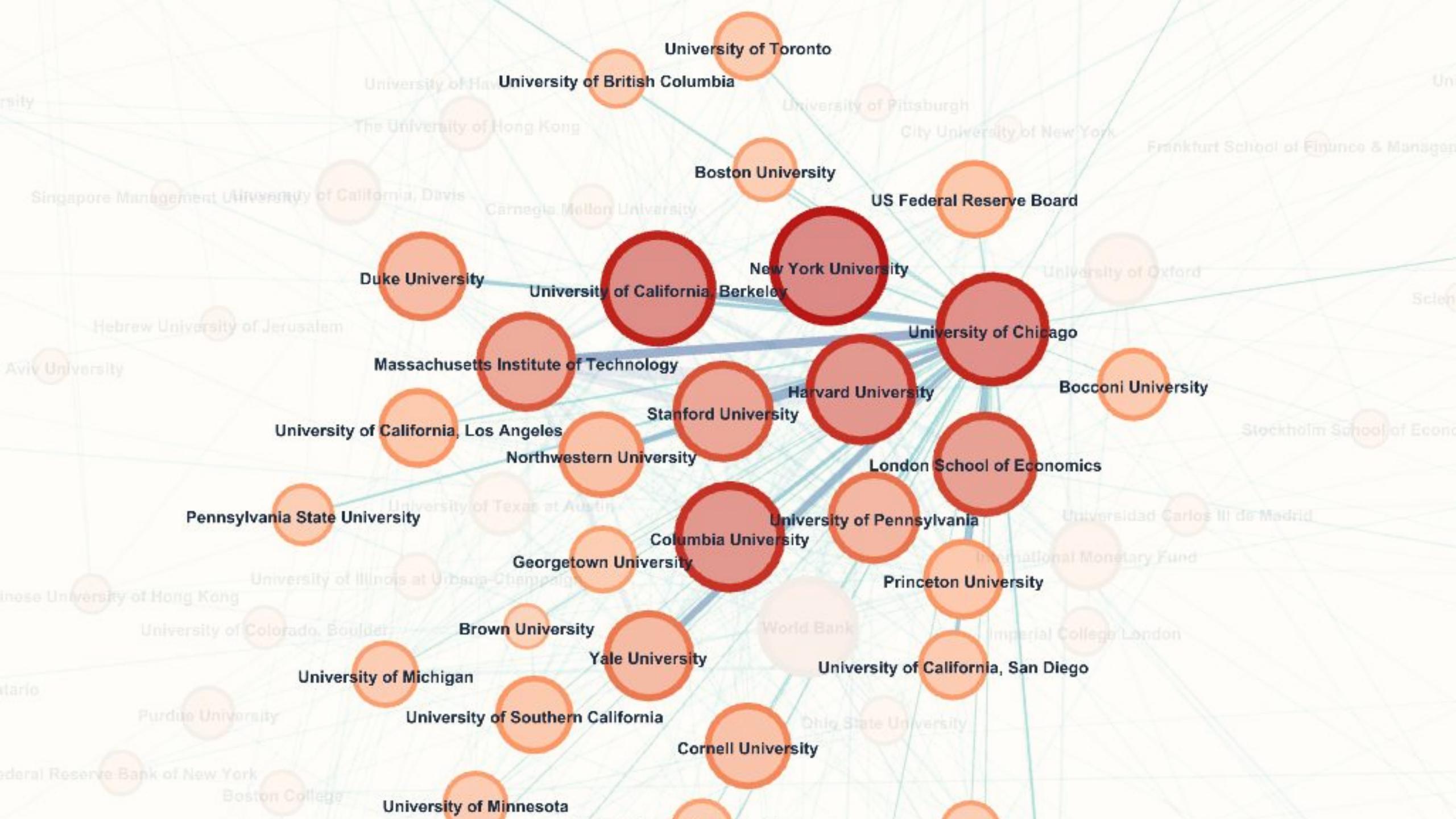
LSE (Whole Dataset)

Name	Number of links
LSE	10
UChicago	5
MIT	5

(example)

- Compile to Institution Level links



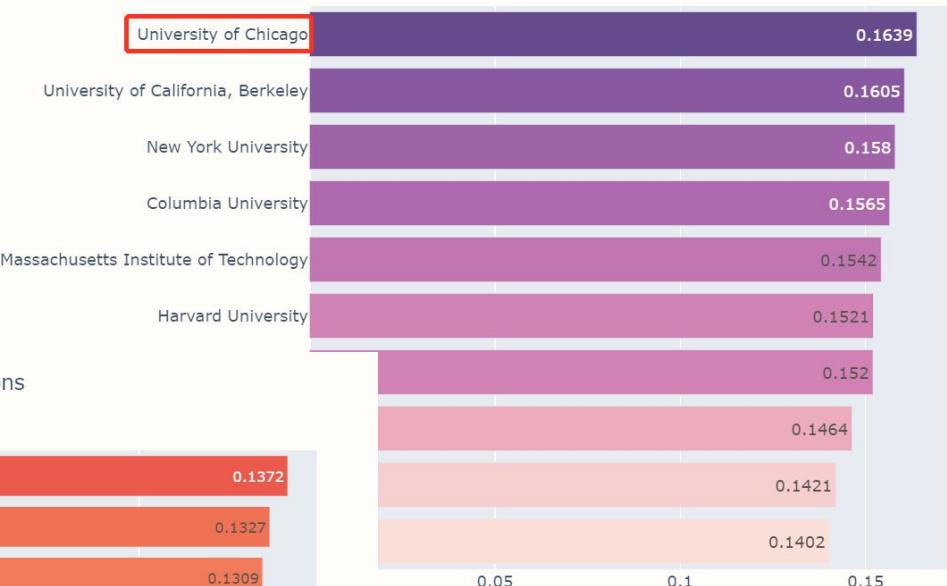


Centralities Analysis

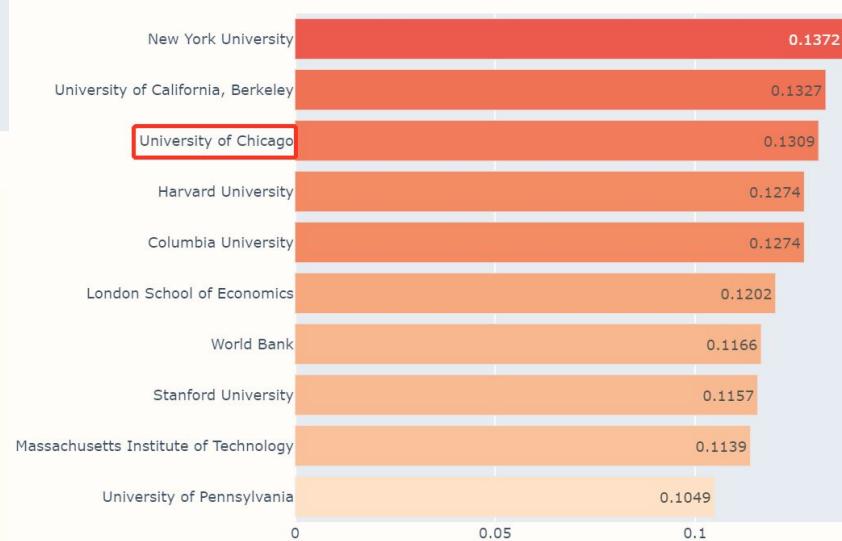
Top 10 Betweenness Centralities of Economic Institutions



Top 10 Eigenvector Centralities of Economic Institutions



Top 10 Degree Centralities of Economic Institutions



Quantifying Collaboration links

(1) Institution level analysis

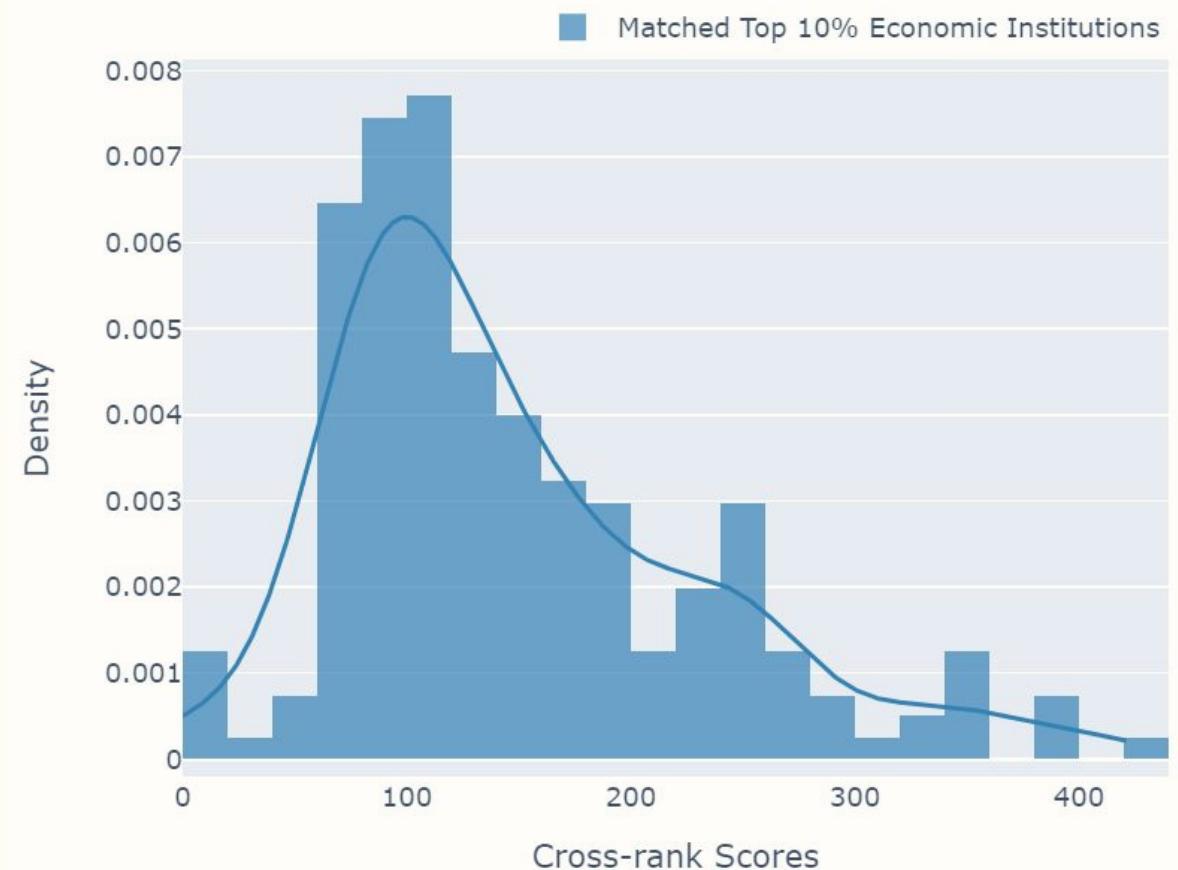
LSE (Whole Dataset)

Name	Number of links	Contribution To the score
LSE (Rank 1)	10	0
UChicago (Rank 9)	5	$(9-1) * 5$
MIT (Rank 8)	5	$(8-1) * 5$

(example)

$$\text{Cross-rank score} = (0 + (9-1) * 5 + (8-1) * 5) / 20 = 3.75$$

Cross-rank Scores Distribution

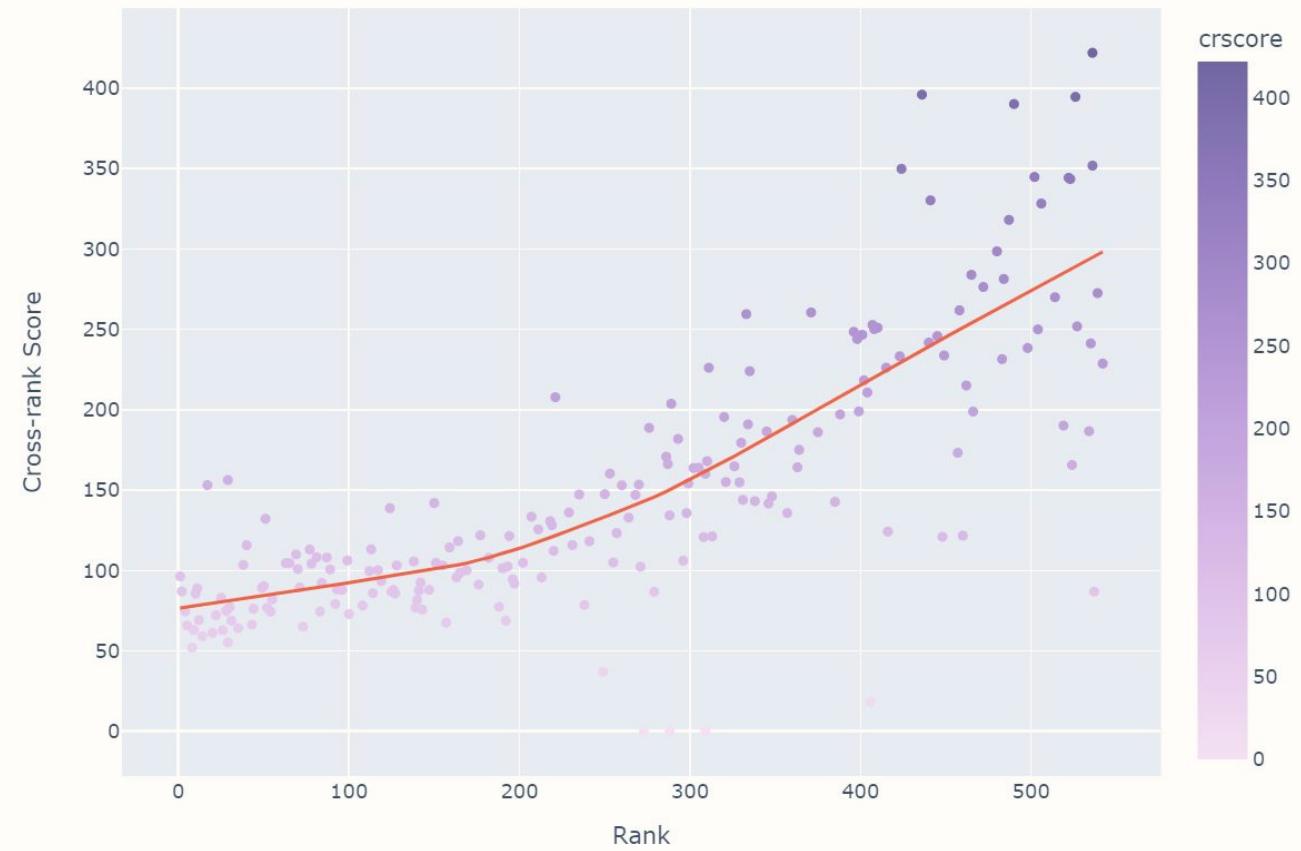


Quantifying Collaboration links

(1) Institution level analysis

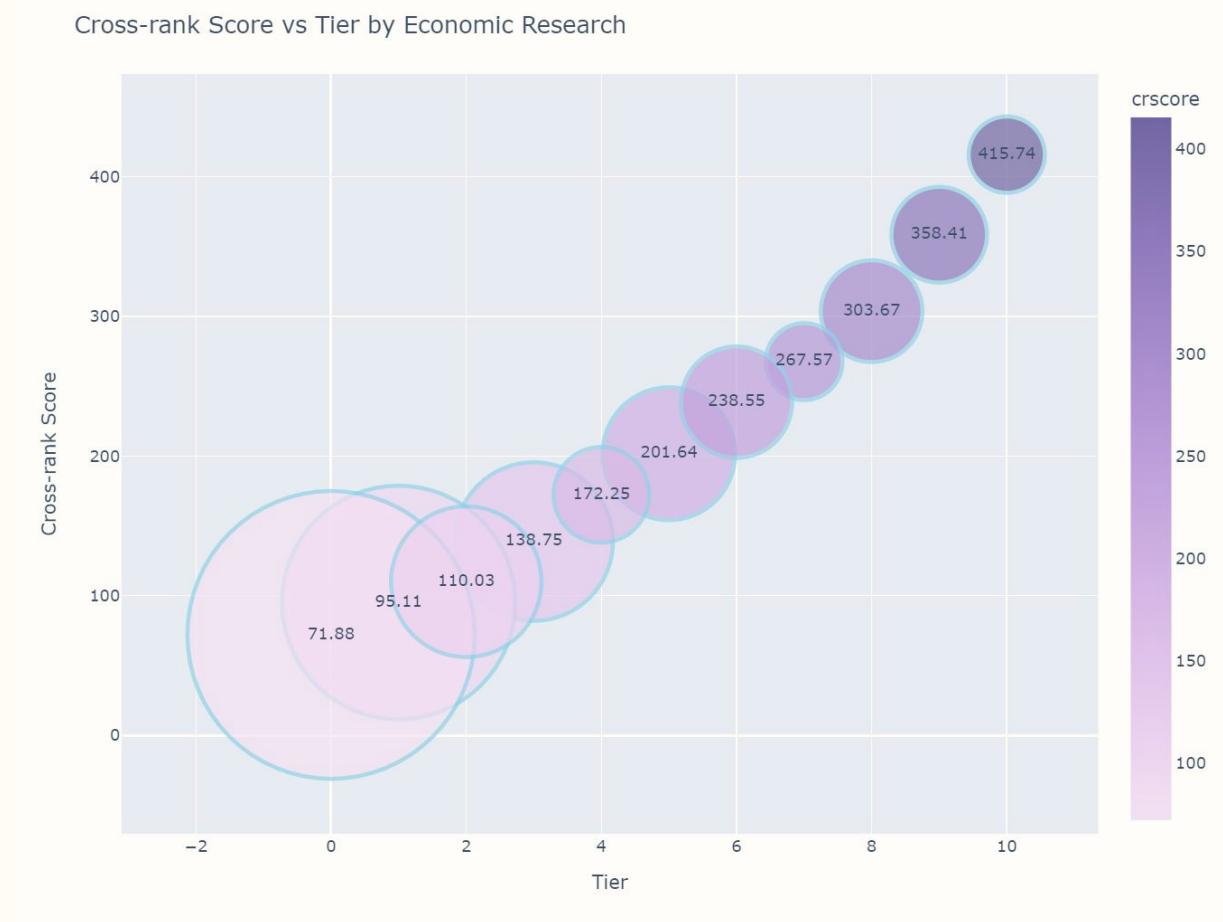
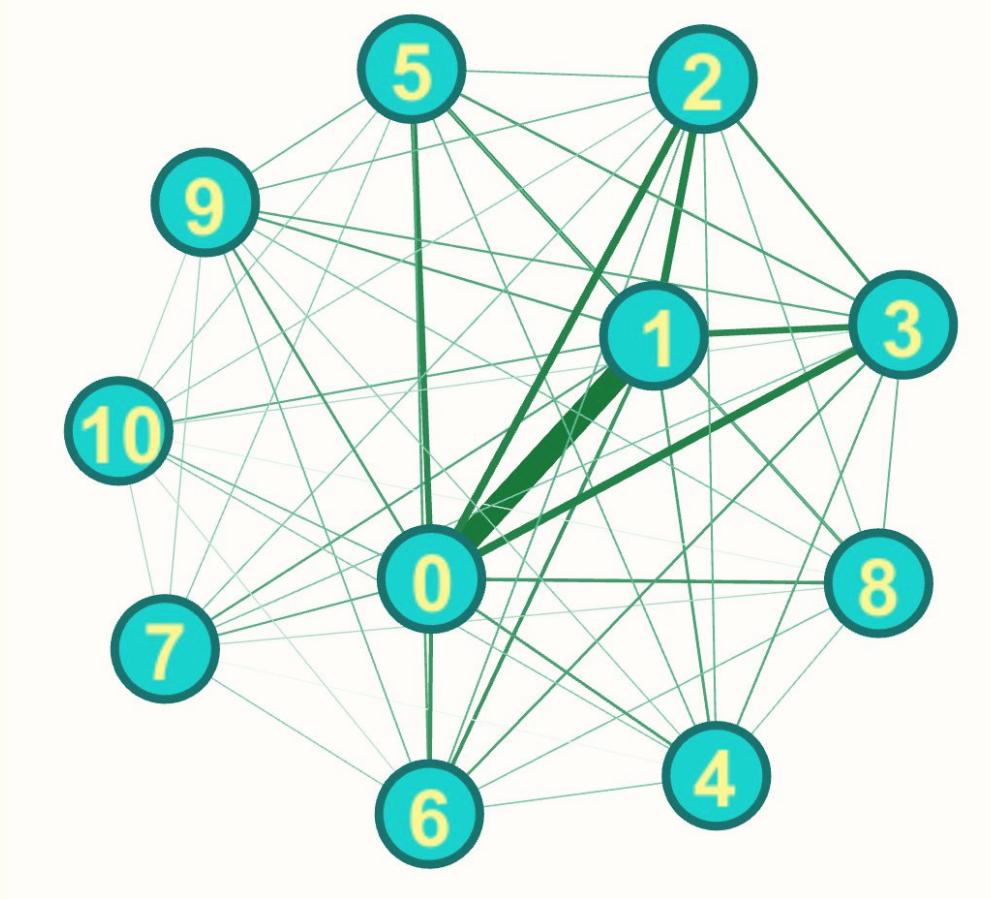
- Institutions that rank higher tend to collaborate more with each other - a low cross-rank score.
- Cross-ranking collaboration is more common for institutions that rank later.

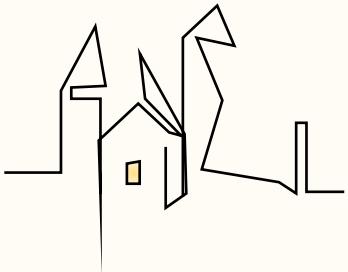
Cross-rank Score vs Rank by Economic Research



Quantifying Collaboration links

(2) Institution Tier level analysis





Topic Modeling Analysis

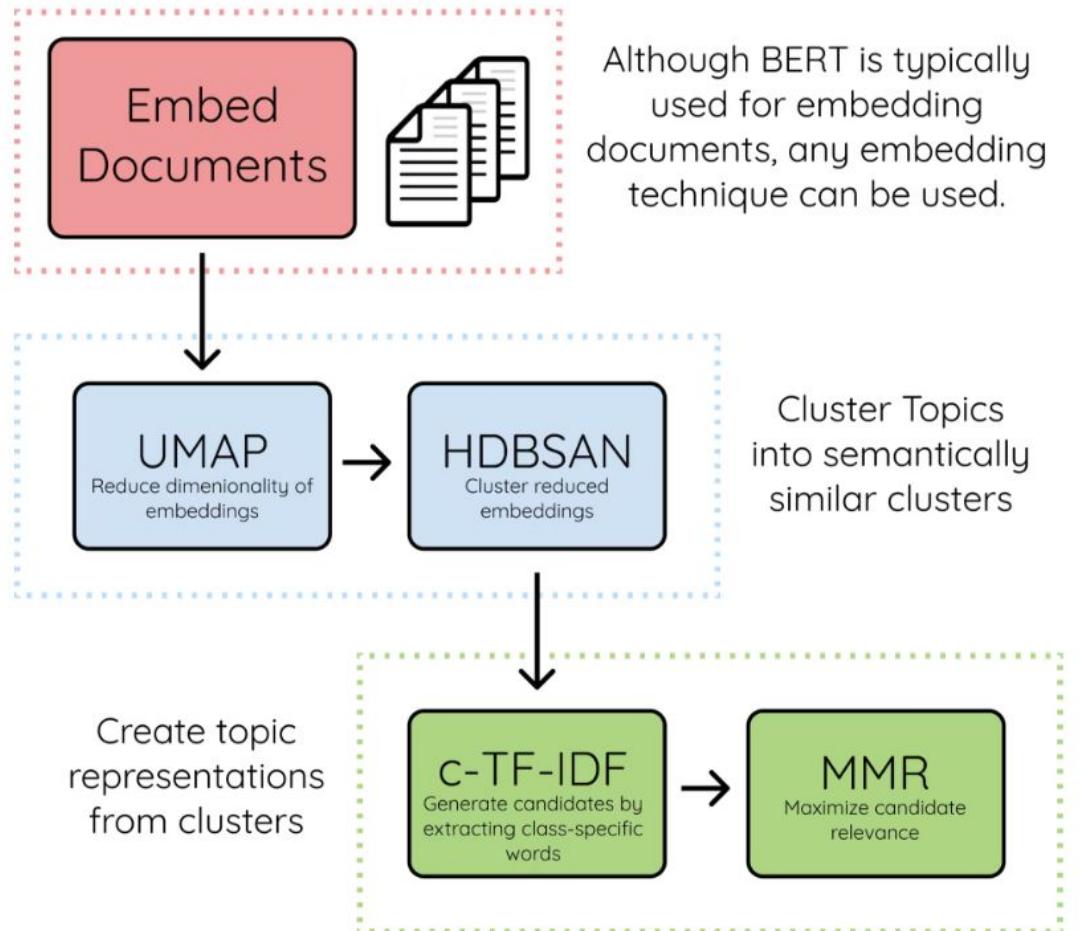
BERTopic Modeling

- An emerging TM technique that uses language embedding and class-based TF-IDF
- Original Paper (Grootendorst 2022):
<https://arxiv.org/abs/2203.05794>
- Documentation:
<https://maartengr.github.io/BERTopic/index.html>

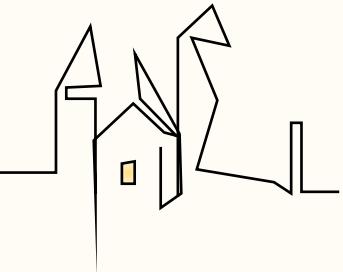
Advantages over LDA TM

- Easy to fit in model: only need to clear stopwords and re-concatenate into paragraphs
- Much more stable: results may vary a little across runs but not as much as LDA
- Much less training time: 15,000+ abstracts take less than 1 minute to train

BERTopic

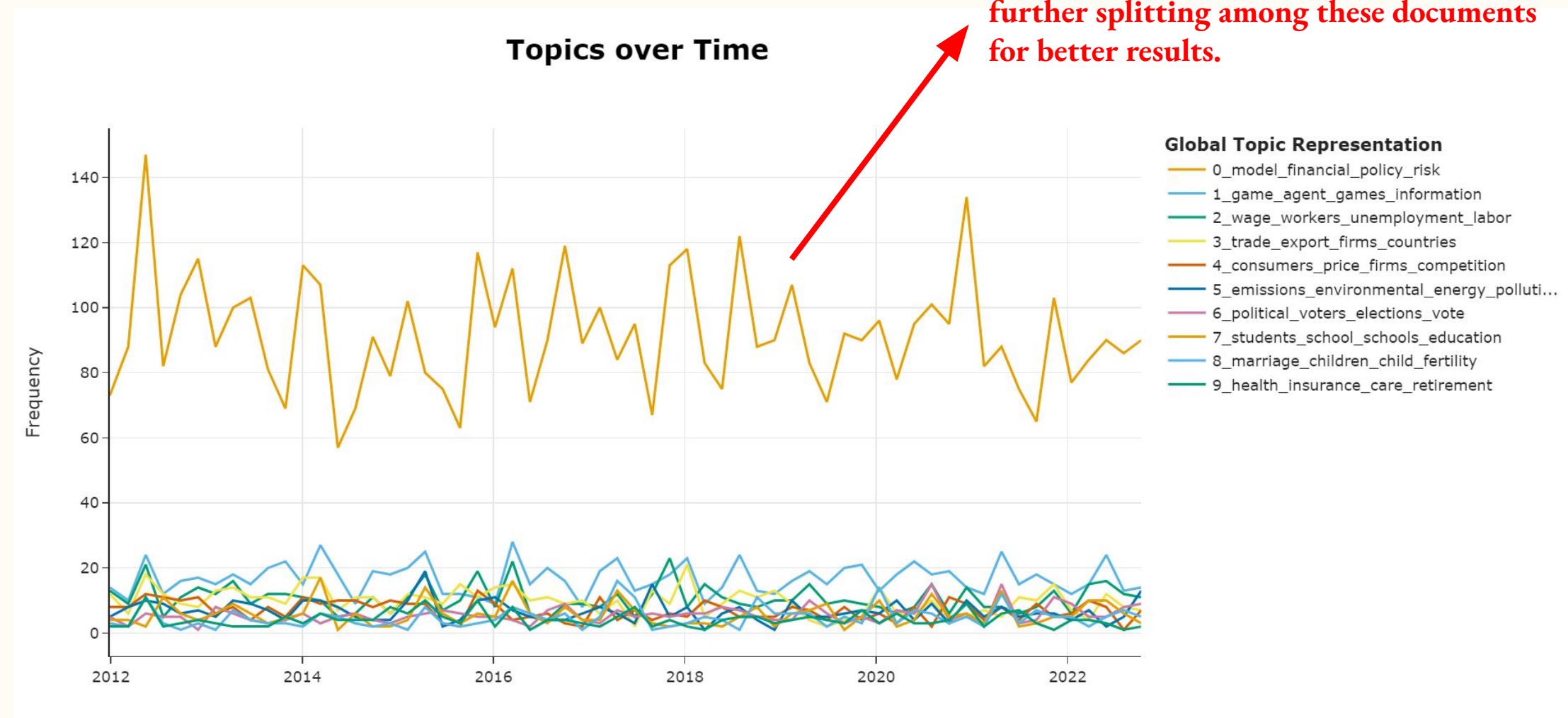


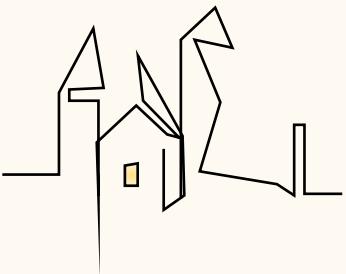
Source: <https://hackernoon.com/nlp-tutorial-topic-modeling-in-python-with-bertopic-372w35l9>



Topic Modeling Analysis

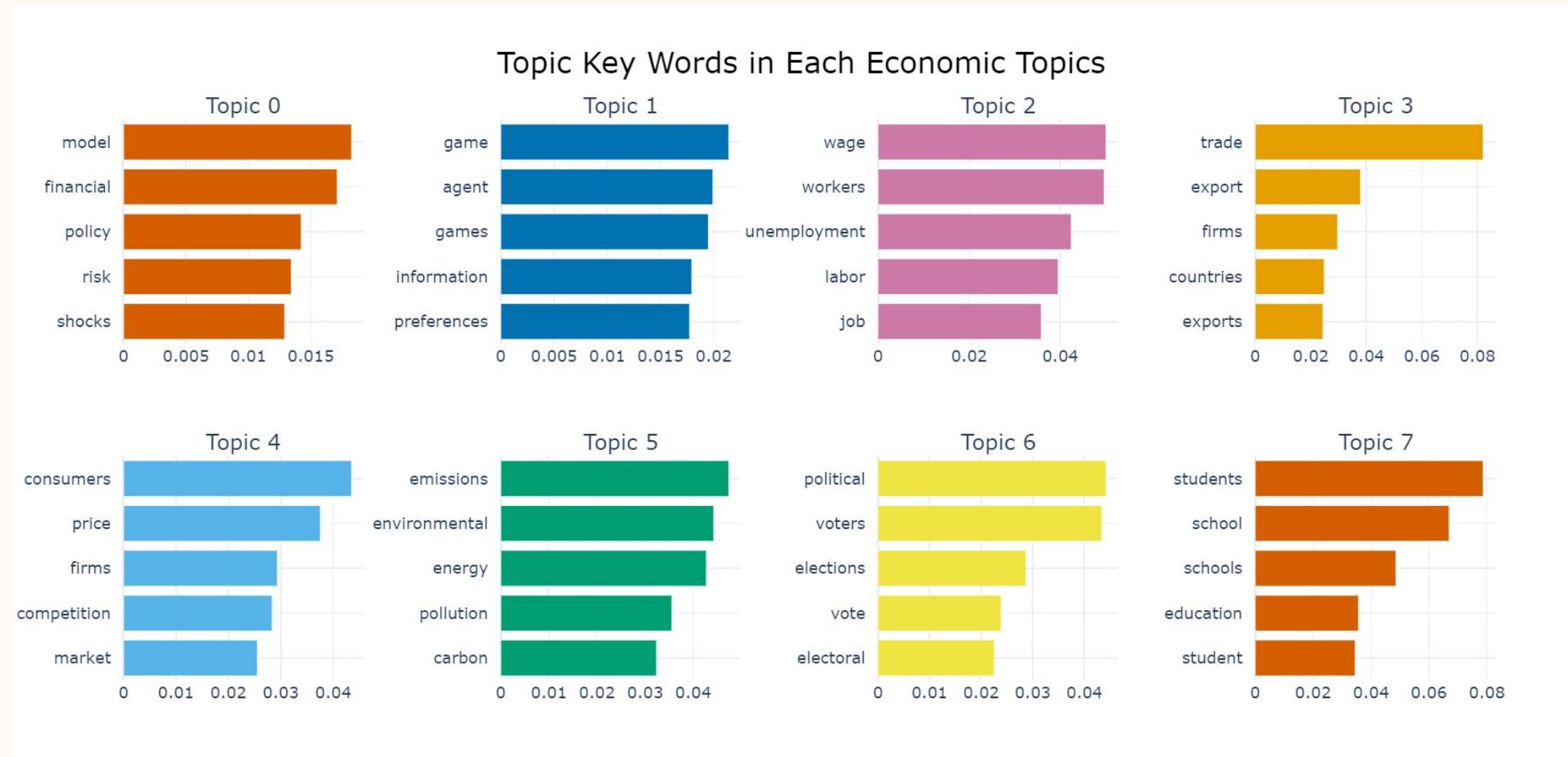
Dynamic Economic Research TM over Time

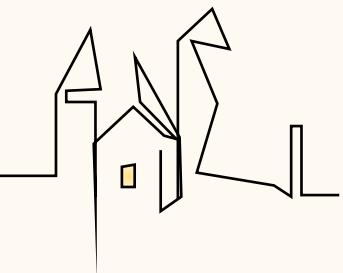




Topic Modeling Analysis

Top 5 Keywords for each Topic:

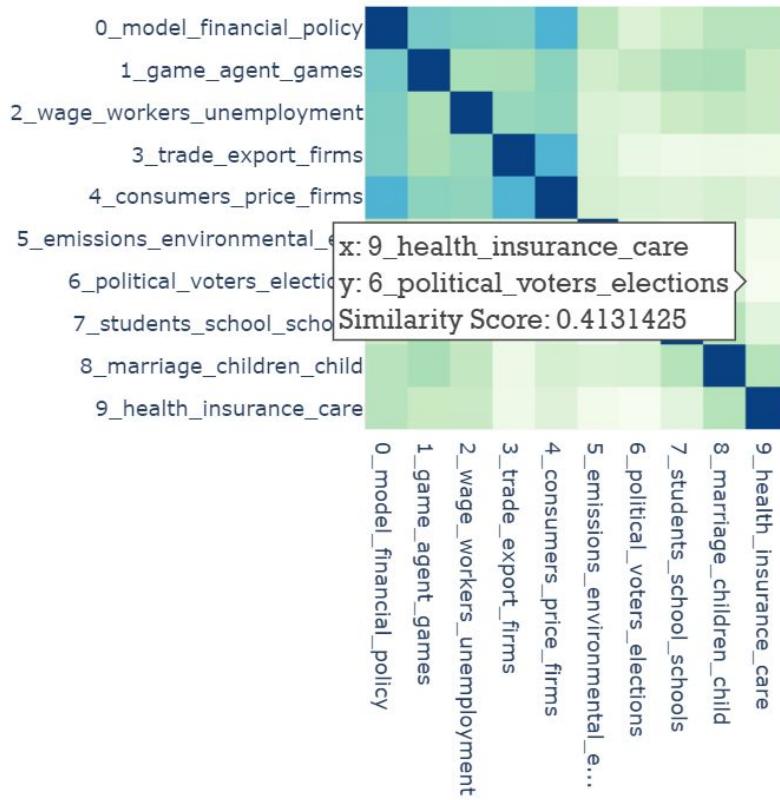




Topic Modeling Analysis

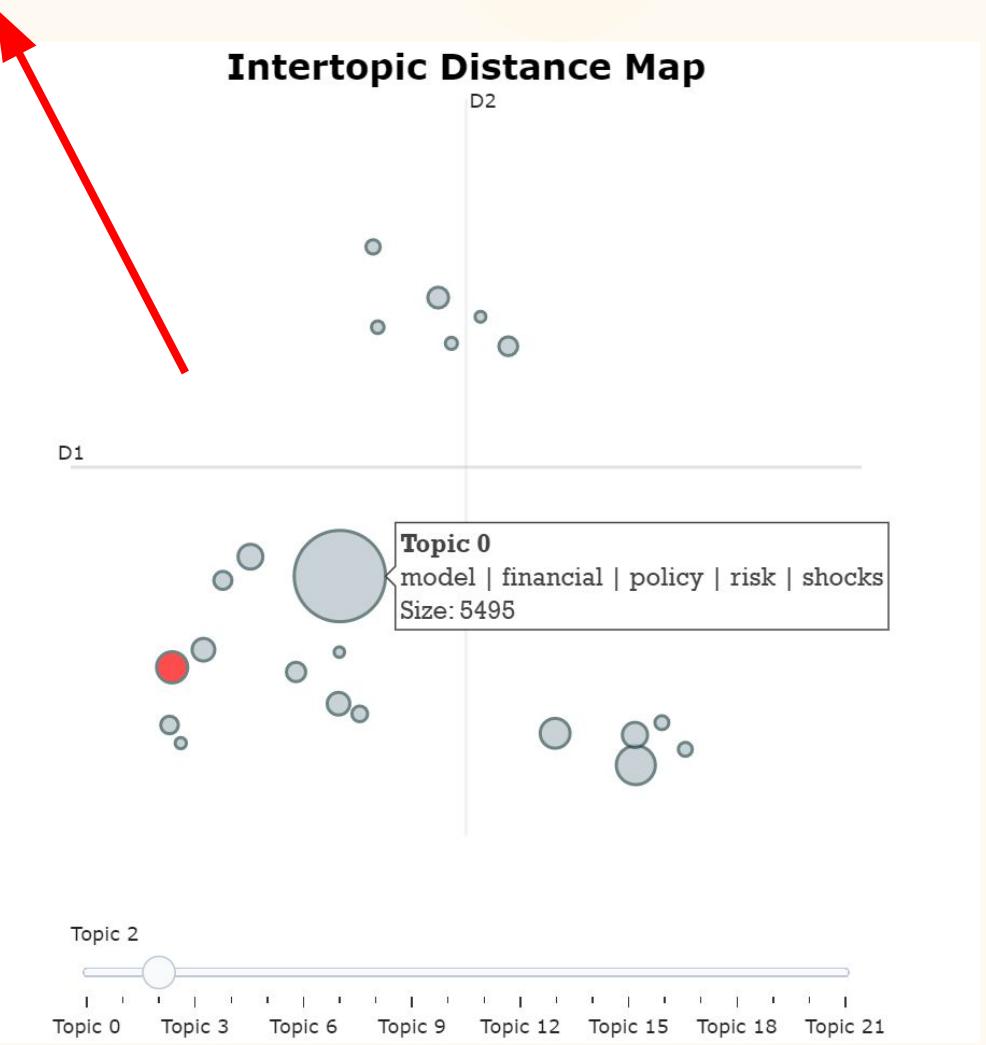
Similarity and Intertopic Distance

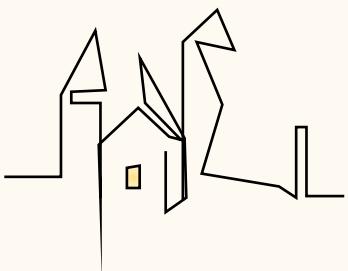
Similarity Matrix



All topics do NOT overlap on Intertopic Distance Map, so the training result should be fine!

Intertopic Distance Map



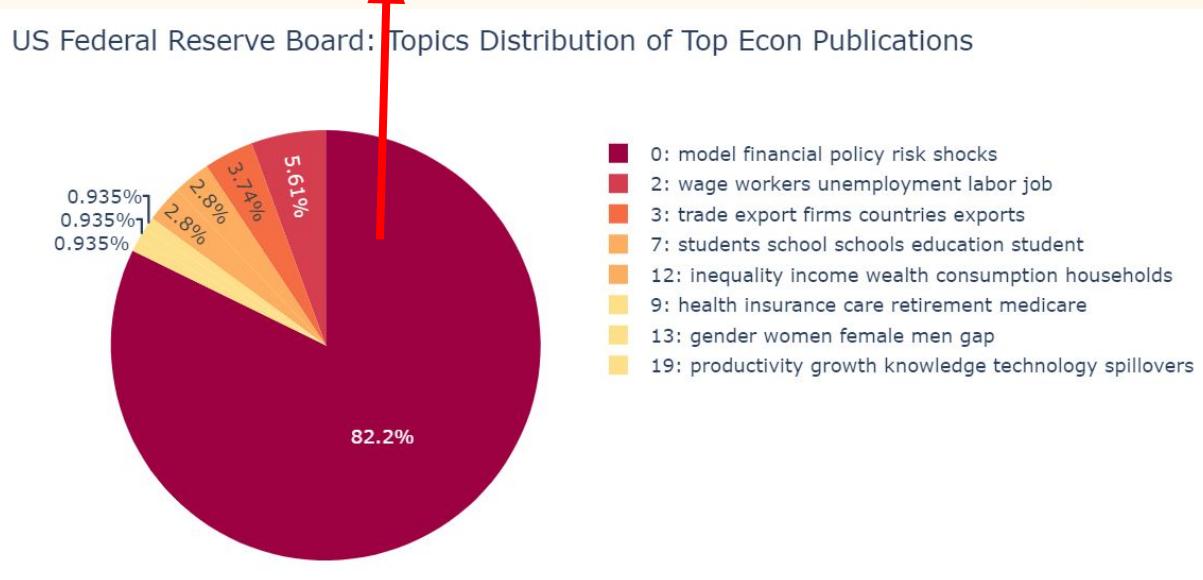


Topic Modeling Analysis

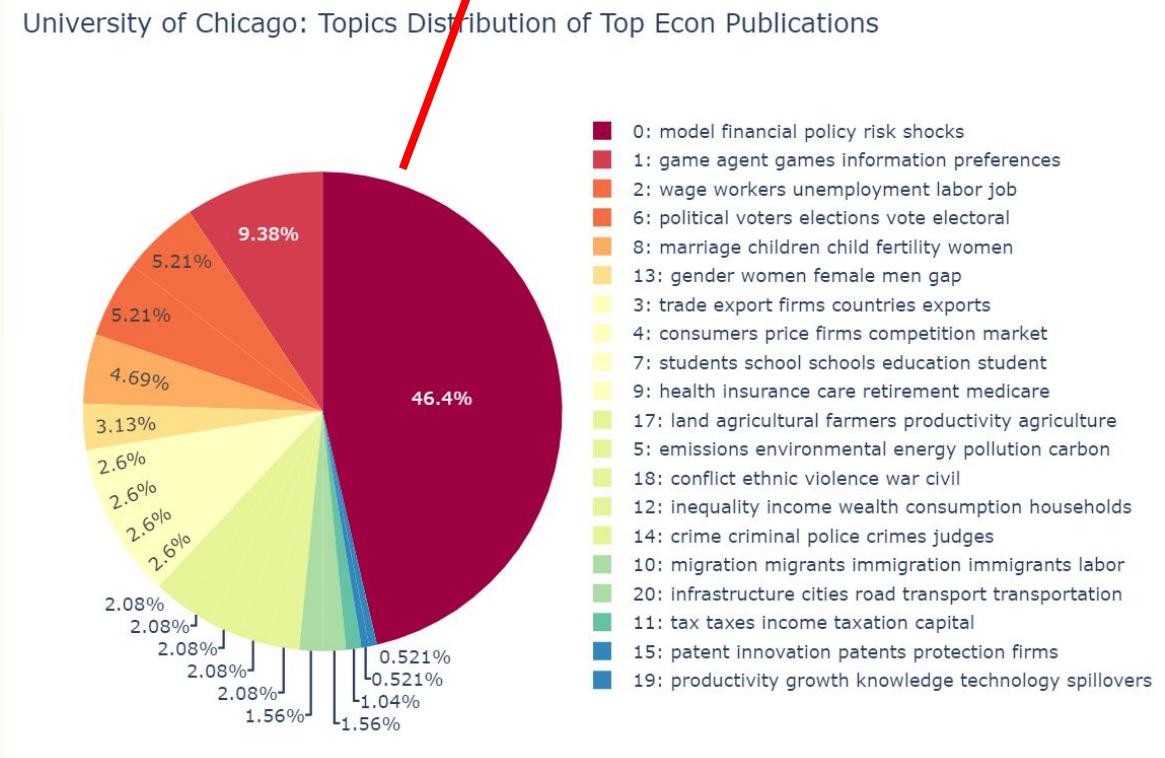
Specialization Areas of Different Institutions

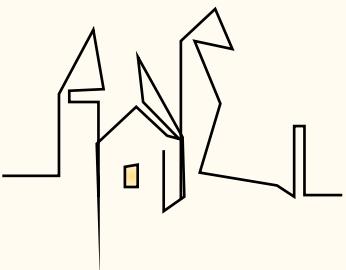
- Every document is assigned with a single MOST RELEVANT topic.
- We can use this result to see what topics an institution is focusing on.

FED is mostly researching about “financial policy, risk, shocks”, macro related topics. Make sense!



UChicago has a much broader interests of topics, almost covering all 22 topics we identify!

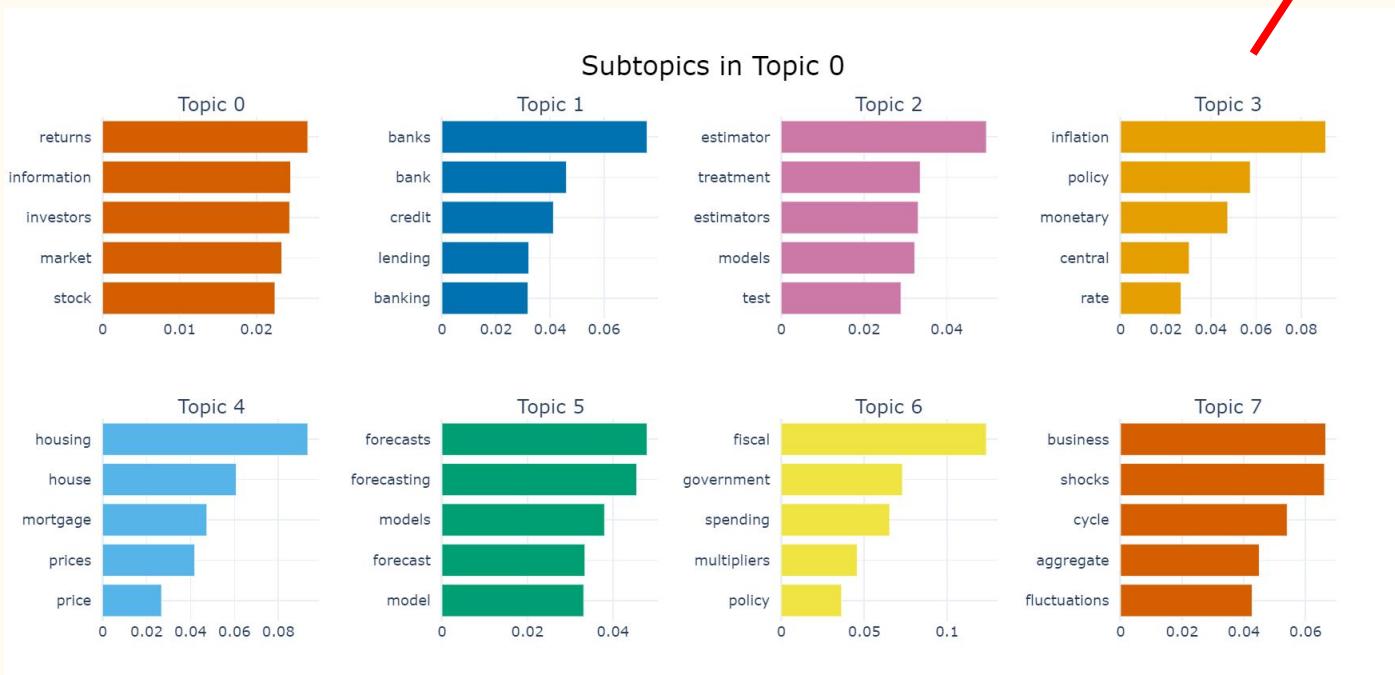




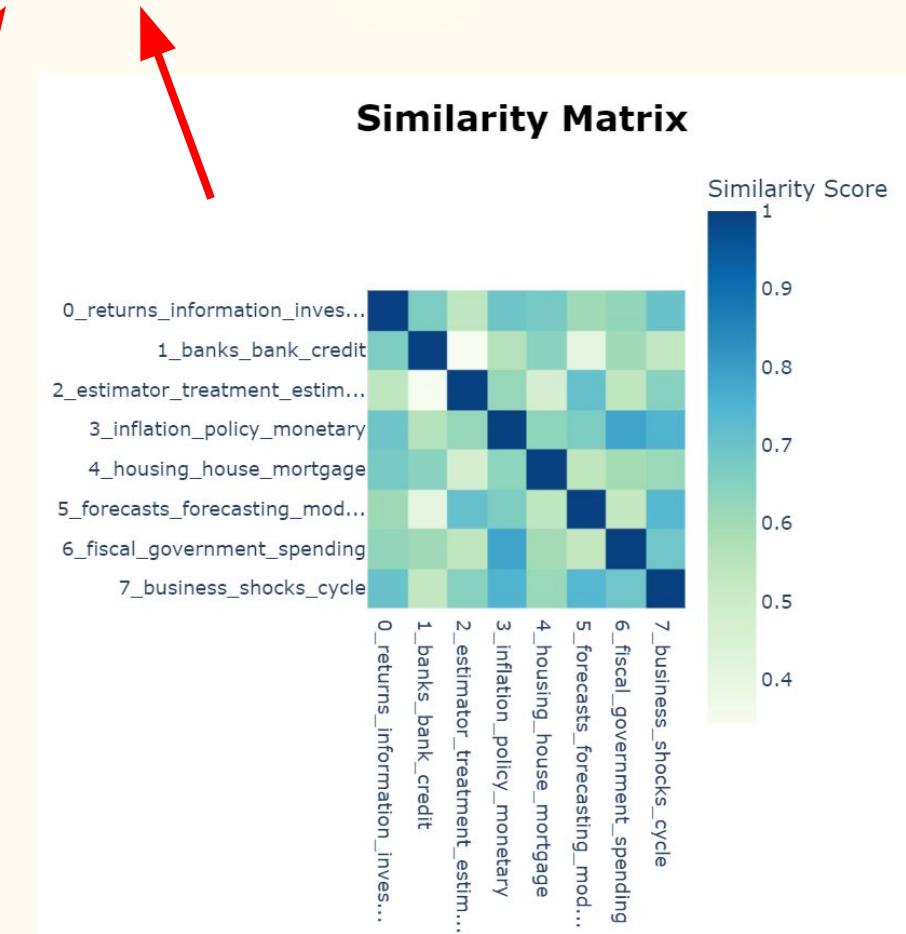
Topic Modeling Analysis

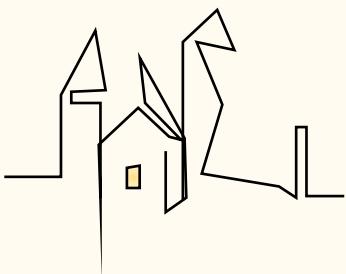
Further Decomposition of Dominating Topic 0

- Topic 0: *model, financial, policy, risk, shocks, ...*
- Occupy 5,000+ abstracts, need further splitting ...
- We re-run a TM on abstracts that belong to Topic 0.
- Show the results of Top 8 subtopics:



Though similarity scores are relatively high compared with the previous whole model, these subtopics do make sense and form separate economic fields.

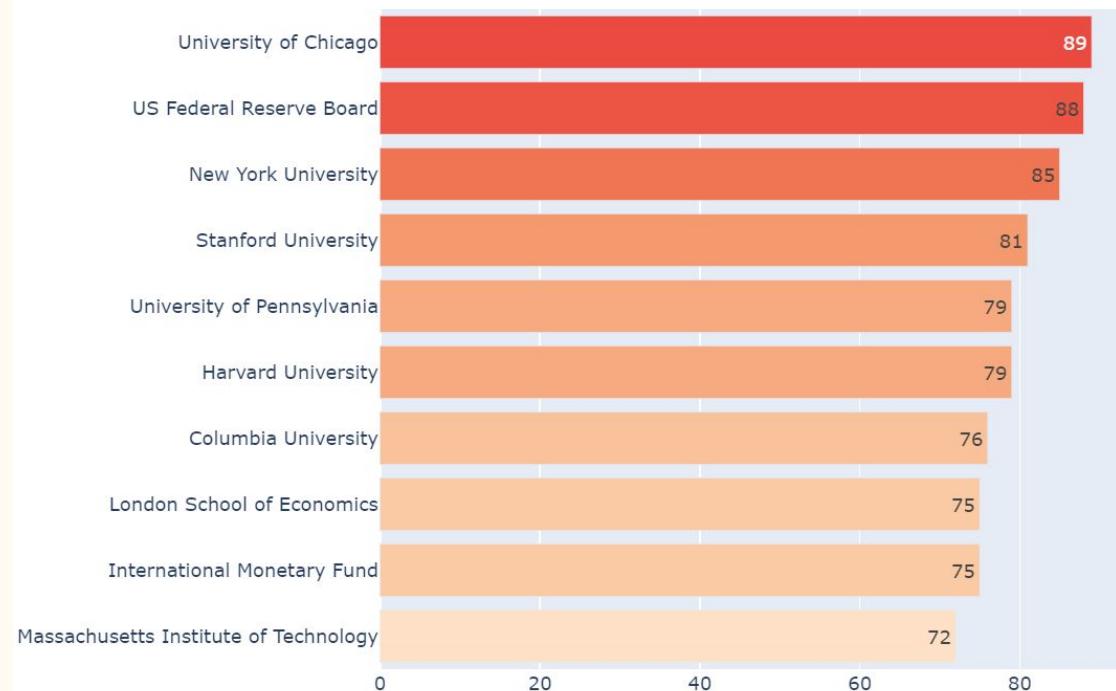




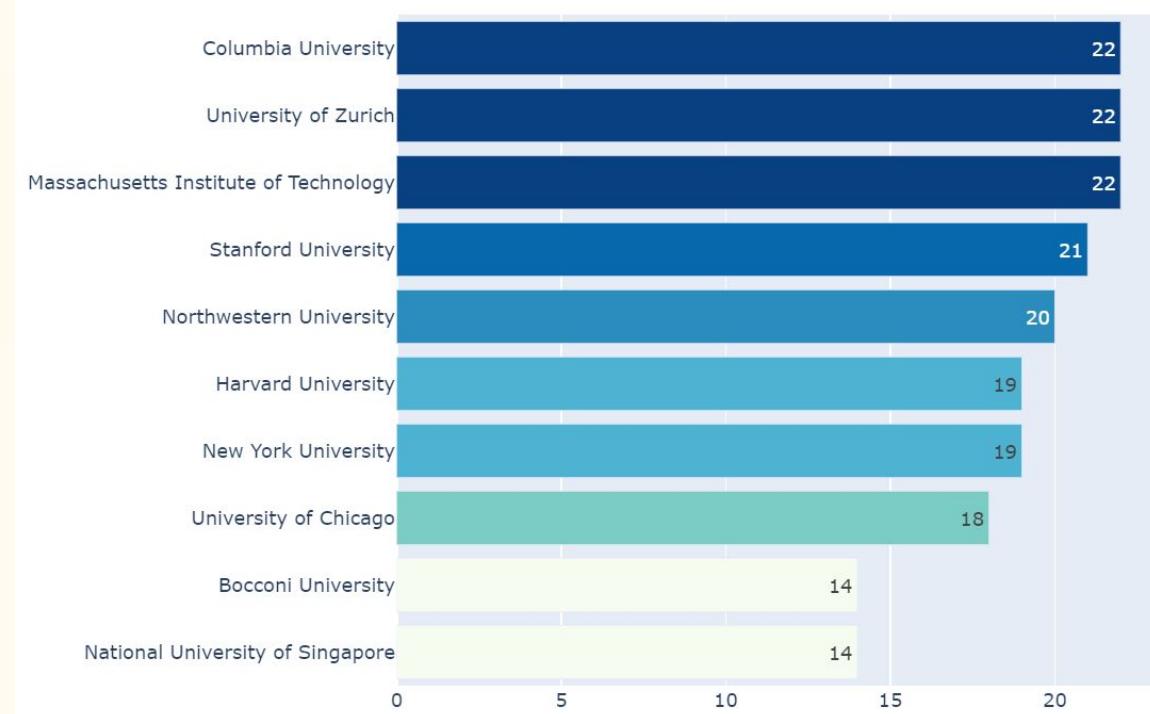
Topic Modeling Analysis

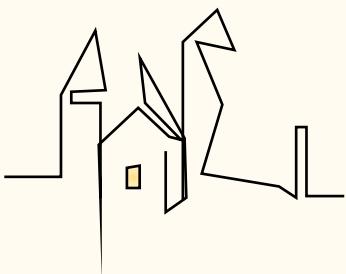
Top Institutions by Top 5 Research Topics:

Top 10 Institutions for Topic 0: model, financial, policy, risks, shocks



Top 10 Institutions for Topic 1: game, agents, information, preference





Topic Modeling Analysis

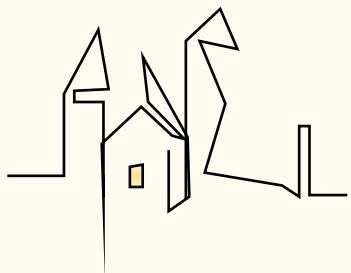
Top Institutions by Top 5 Research Topics:

Top 10 Institutions for Topic 2: wage, workers, unemployment, labor, job



Top 10 Institutions for Topic 3: trade, exports, firms, countries

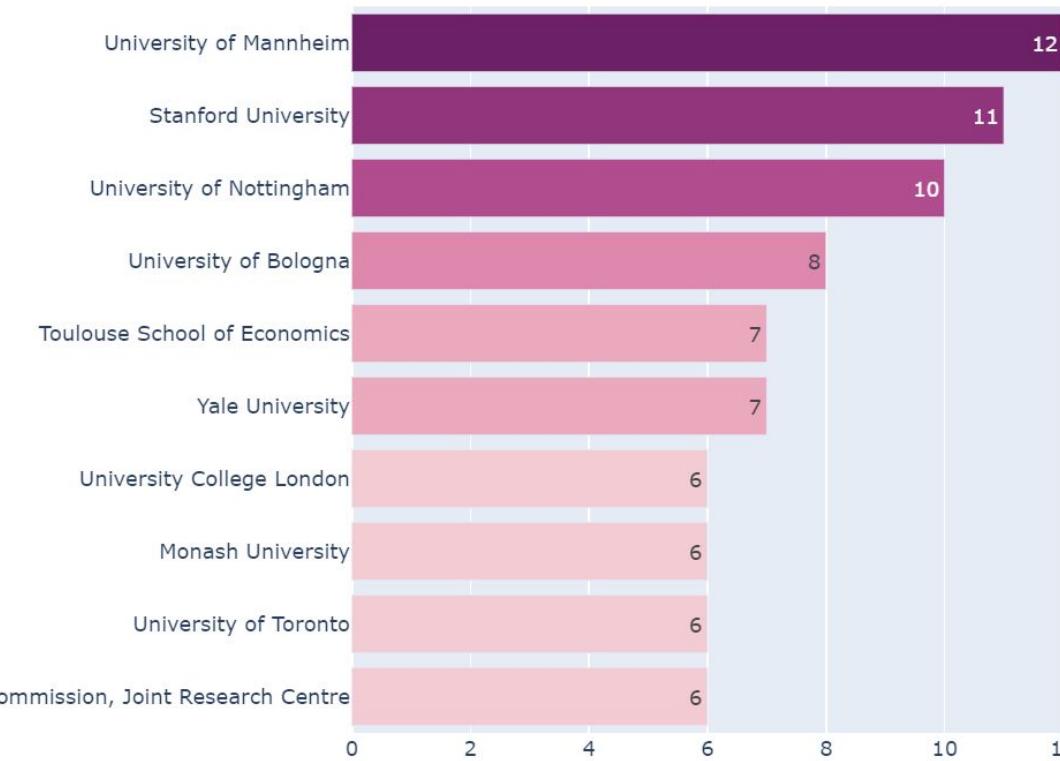




Topic Modeling Analysis

Top Institutions by Top 5 Research Topics:

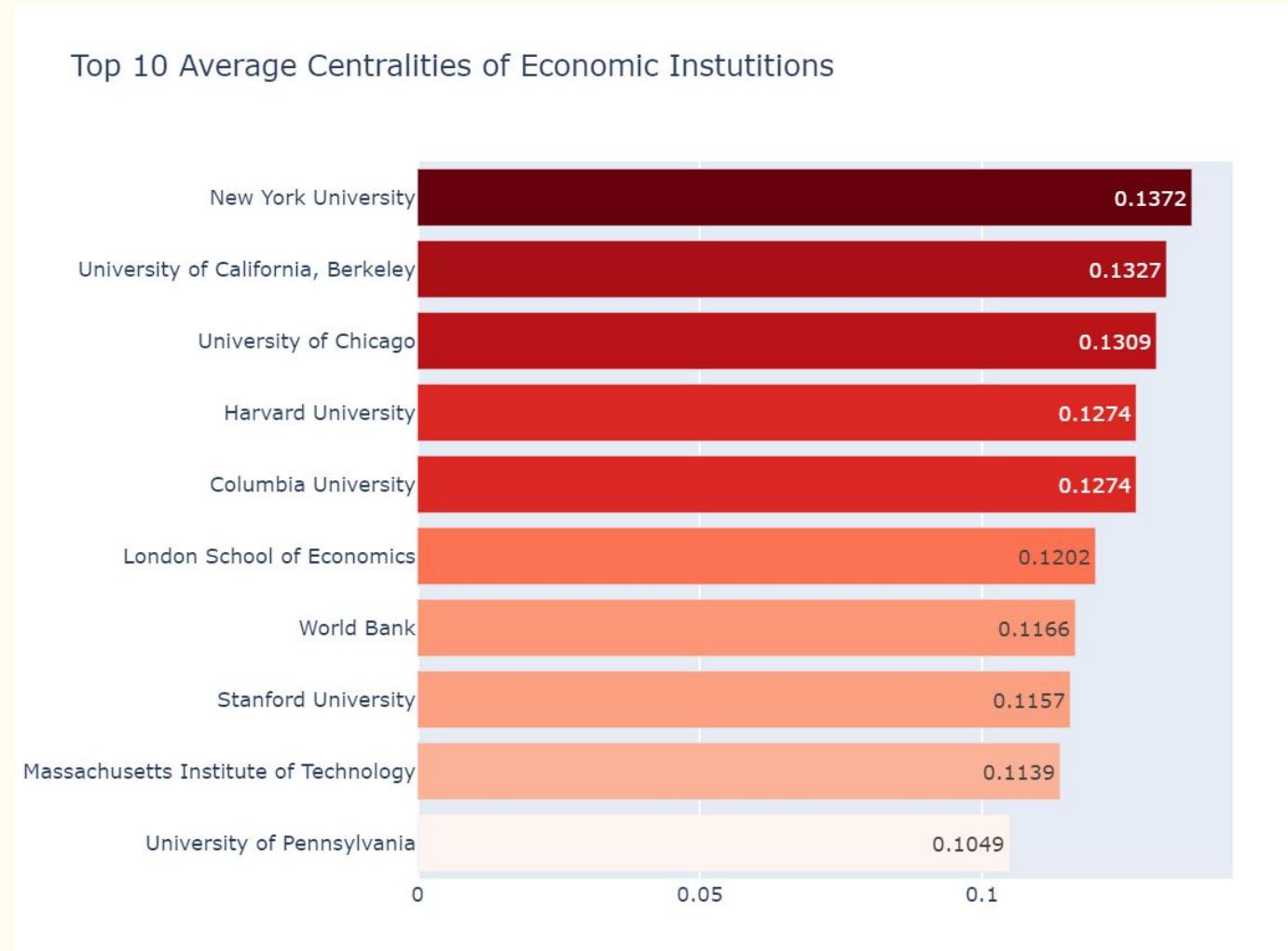
Top 10 Institutions for Topic 4: consumers, price, firms, competition, market

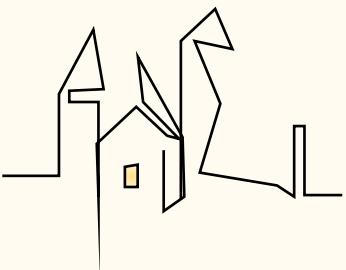




Do Centralities Research the Same Topics?

Combining SNA and TM First Recall the Top 10 Centralities:



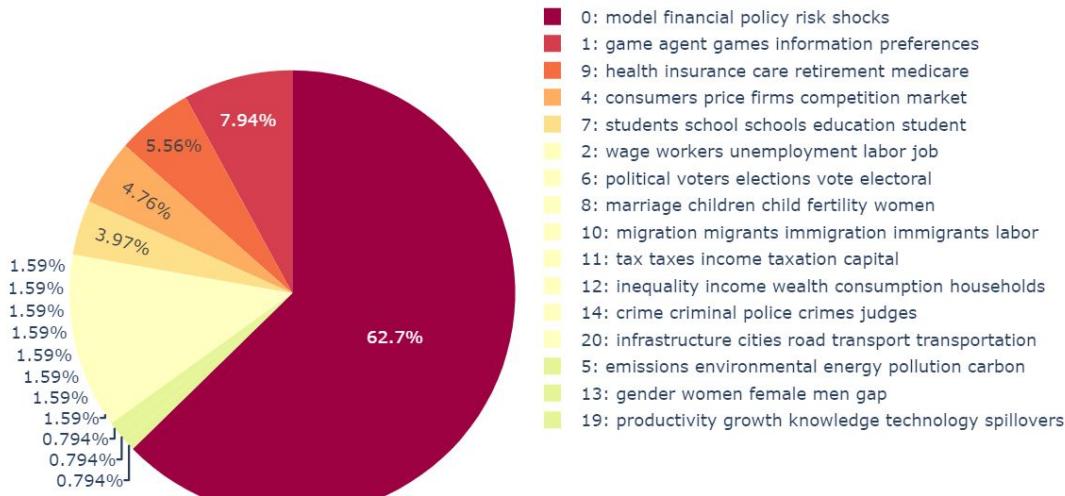


Do Centralities Research the Same Topics?

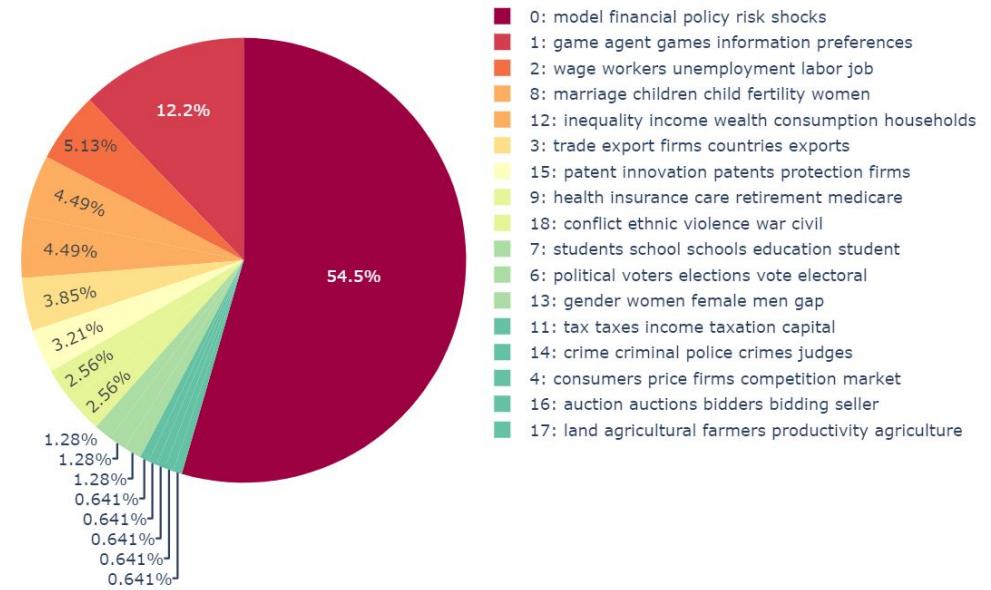
Combining SNA and TM Show Centralities' Topics Distributions:

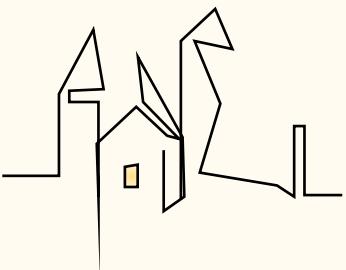
- Some are mainly focusing on the Topic 0 about Financial Policy, Risks, and Shocks (50%+ of publications are about Topic 0)
- e.g., LSE, NYU, UPenn, ColumbiaU

University of Pennsylvania: Topics Distribution of Top Econ Publications



New York University: Topics Distribution of Top Econ Publications



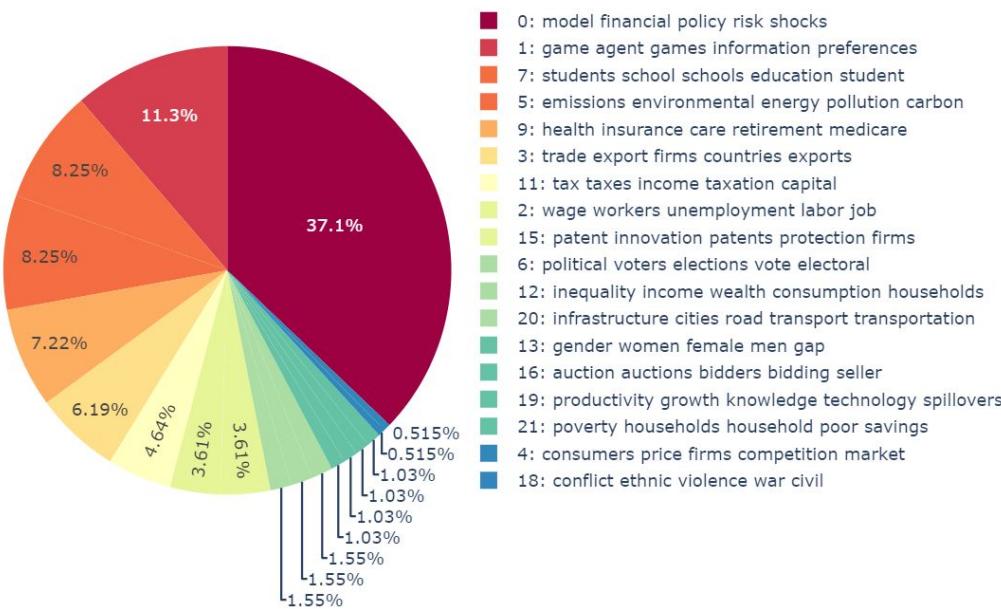


Do Centralities Research the Same Topics?

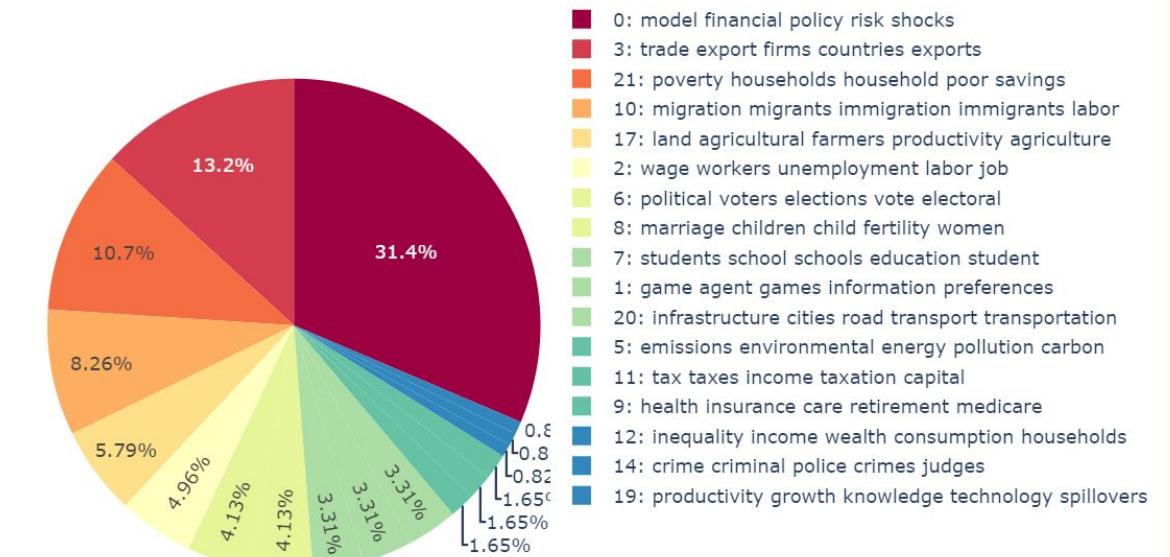
Combining SNA and TM Show Centralities' Topics Distributions:

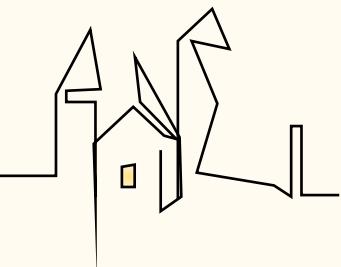
- Others have more balanced distribution of topics.
- e.g., MIT, UCB, Harvard, UChicago, Stanford, World Bank

MIT: Topics Distribution of Top Econ Publications



World Bank: Topics Distribution of Top Econ Publications



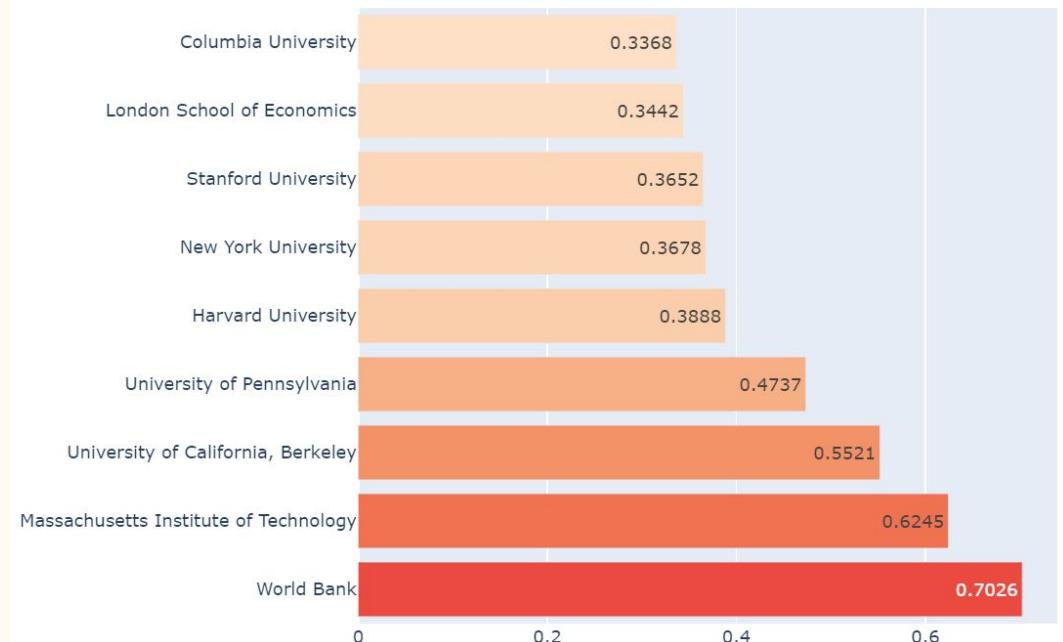


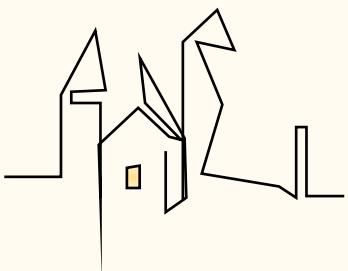
Do Centralities Research the Same Topics?

Combining SNA and TM Discrepancy Score between 2 Institutions' Topics:

- We want to measure the (dis)similarity levels of research topics among central institutions.
- Specifically, we would like to see the discrepancy in proportions of research topics. The **higher discrepancy** is, the **less similar** two centralities' research topics are.
- Suppose we have 2 institutions A and B, with n attributes X_{ij} .
 - ★ n is the number of possible topics which is 22 in our case.
 - ★ i denotes institution i , and j denotes Topic j .
 - ★ Each attribute is the proportion of a certain Topic j among all.
 - ★ Discrepancy Score between A and B := sum of $|X_{Aj} - X_{Bj}|$ over j
- **For example: UChicago vs Other Centralities**
 - ★ UChicago is very similar as ColumbiaU and LSE, but not so similar as MIT and World Bank.

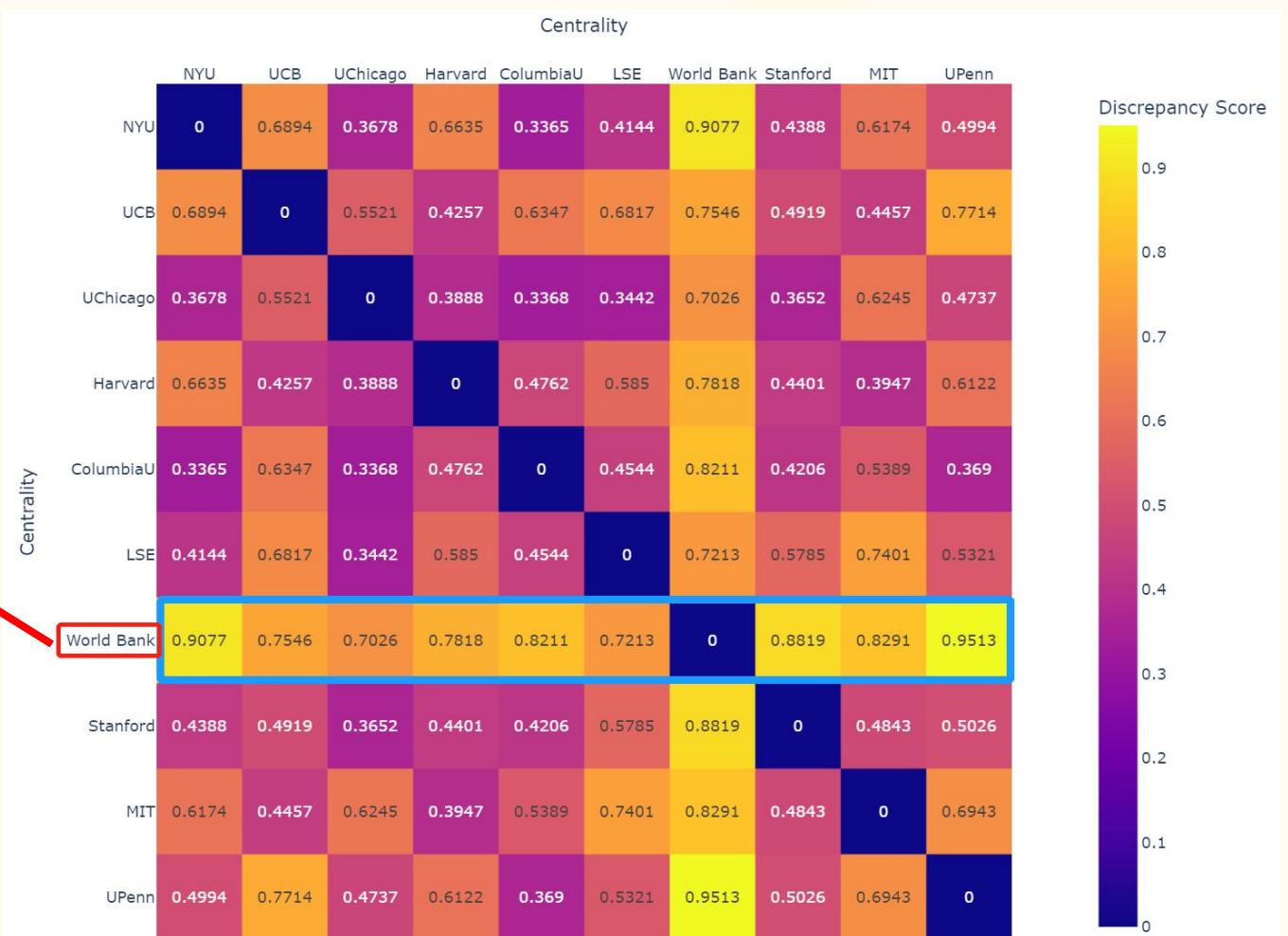
Discrepancy Score: UChicago vs Other Institutions



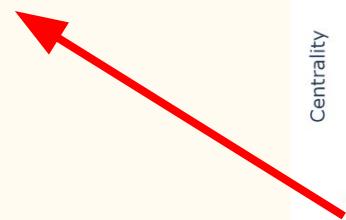


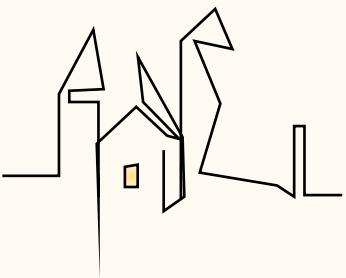
Do Centralities Research the Same Topics?

Combining SNA and TM Discrepancy Scores between 2 Institutions' Topics:



World Bank has large discrepancy scores with other university-institutions! While among university-institutions, the discrepancy levels are relatively lower in general.





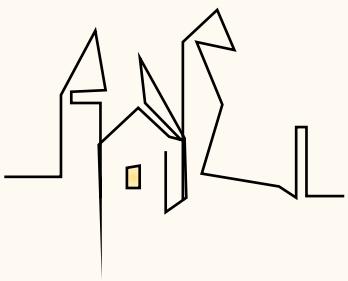
Take-home Points

From Technical Experience:

- **SQL** is really convenient to structure datasets and extract relevant information!
- Except for speed, **Selenium** is way better than Requests for web scraping!
- **Plotly** is fantastic for fancier visualization and interactive plots!
- When doing Topic Modeling, try out **BERTopic** package! Do NOT stick to Gensim LDA.

Social Science Significance:

- We form a whole picture of the current Economic academia by 1): identifying its collaboration patterns on Top Econ Journals through network analysis 2): research topics evolvement across time and institutions through dynamic topic modeling.
- **From SNA**, Top tier institutions are more likely to collaborate with institutions of similar levels; but for lower tier institutions, they often collaborate with those from very different tiers. In other words, an institution's cross-rank/tier collaboration likelihood decreases as its tier/rank gets higher.
- **From TM**, we see topics are stable in trends across time with some fluctuations but can vary a lot across different institutions.
- **From both SNA and TM**, we find centralities of university-institutions are relatively similar in what they research, but World Bank shows a large discrepancy against university-institutions.



This is the end of Presentation.
Thanks for Listening!

Q & A

