MACS30122 Final Project Report

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

1. Team Name

Netrunners

2. Team Members (first and last name, email)

Jerry (Guo) Cheng	chengguo@uchicago.edu
Yicheng Zhang	zycheng@uchicago.edu
Hongzhang Xie	hongzhangxie@uchicago.edu

3. GitHub Repository Link

https://github.com/macs30122-winter23/final-project-netrunner

4. Literature Reviews

Academic research plays an important role in our society. Mansfield (1991) found that the social rate of return from academic research between 1975 and 1978 was 28%. Jones et al. (1998) also believed that the rate of return on research was around 30% and the optimal level of research funding was four times the current actual investment. From 1956 to 2020, research and development funding in the United States increased 82 times. There are substantial resources invested in academic research. It is essential to do the research in academic research itself.

Lazega et al. (2008) found different research capabilities of French cancer researchers (big fish/small fish) in different levels of scientific research institutions (big pond/small pond) have different competitive advantages and competitive strategies. This finding suggests that the pattern of research network structures can impact academic research productivity.

We can also find that with the development of modern science, the connections between academic research networks are increasing and deepening. Uddin et al. (2012) examined publications in an emerging research field: 'steel structures'. They found that the number of collaborators per paper and collaborators per author increases as the research field grows. They also find that the degree centrality and betweenness centrality of the ICNs network has a significant positive correlation with the average efficiency of the ICN papers. This finding shows the growing importance of collaborative networks as the academic research system evolves.

In addition, Larivière et al. (2018) conducted a study on the cooperative relationship between universities and industrial research institutes. They found that the monopoly level of journal publishing is increasing. Universities become major publishers of journals, while industry research institutions are hard to publish their findings in academic journals. Then they empirically prove the increased trend of monopoly will hinder the ability to translate basic knowledge into technological innovation. Based on their findings, we can expect institutions to institutions and researcher to researcher monopolies may also have a negative impact on research outputs. Previous studies have pointed out that the existence of a monopoly in academia is widespread. Ramsden (1994) collected and analyzed the publication data of Australian researchers and found that "Most papers are produced by few academic staff." It is important to analyze collaborative networks in academic research.

Network analysis has a wide range of applications in social science research. In the field of academic research collaborative networks, there are some studies that focus on the cooperation network between colleges and industrial institutions, such as Larivière et al. (2018), Calero et al. (2007) and Nicolaou et al. (2003). And some studies focus on the social network of the researchers. For example, the researcher's social platforms (Meishar-Tal & Pieterse, 2017).

However, there are a few studies that analyze academic research collaborative networks using journal publication data. One of the important and interesting studies is Tang et al. (2008). They introduce a system to extract and mine the academic social network data. At the time they published the paper, they extracted 448,470 research profiles. They discussed how to extract the profile, solve the name disambiguation problem, model the academic network, and do the topic models. In Google Scholar, this article has been cited 2,279 times now. We found that even though they did much deeper than us, the ideas of the research process designed in our final project are similar to their design (we did not read that paper when setting up our research strategy). We think this shows that our final project design is correct and meaningful.

Academic research has an important impact on social productivity, and the academic network also has an impact on the productivity of academic research. Therefore, understanding the characteristics of the economic research collaboration structure is important. This is one of the reasons our group decided to do this topic.

Another reason is that after entering the era of big data, we now have high-quality academic research publication data. We believe that by understanding deeper of these data, we can predict the expected benefits of collaboration between specific researchers and research institutions. In the future, we may be able to build the model to help researchers make better research collaboration strategies (like Zhou et al. (2018)).

5. Project Description

In this project, we aim to **examine the characteristics of the economic research collaboration structure based on articles scraped from the top 20 economic journals from 2012 till now**.
Our main research questions are:

(1) What are the network features and collaboration patterns at the institutional and tier level? What are the features of density, centrality or cross-rank freedom?
(2) How do research topics vary and evolve across time and institutions? For each topic, who are the most important contributors?
(3) Do centralities in the network research the same topics? For a centrality, who are parallel peers, who are distinct peers?

To answer the first question, we investigated which authors and institutions are leading the economics research, and how researchers and institutions collaborate with each other by conducting network analysis. Are there any authors in centrality in the networks? Do giant institutions (e.g., Harvard, MIT, Uchicago) persistently dominate Top economic journals? Are universities in the same tier more likely to collaborate with each other or is cross-tier collaboration common? Insights to the above questions can give us a clear structure of how economic academia functions.

Strategy:

1) Data visualization method: We conducted network analysis to intuitively visualize the distribution of collaboration between tiers.
2) Quantitative method: We design the metric “cross-rank score” to quantitatively analyze the level of cross-ranking / cross-tier collaboration

Hypotheses:

1) We expect to observe a dominant share of the paper published in the top 20 journals by elite (top 20) economic research institutions / universities.
2) We expect that collaborations among researchers are more common among those from the same tier / close tier of universities - intuitively, there should be a negative relationship between the probability density of co-authorships and the gap of the institutions authors belong to.

To answer the second question, we are interested in how research topics vary and evolve across time and institutions. For each topic, who are the most important contributors? Do relatively new topics / emerging topics such as machine learning possess a relatively lower level of exclusiveness in collaboration? How does the density function of collaboration distance differ from that in well established / traditional topics, for instance supply and demand theory?

Strategy:

We used the BERTopic Modeling method to find the topics we are interested in. And a direct measurement is to use LDA modeling to separate papers into groups and apply the same quantitative method to retrieve the collaboration score for every topic group. To see the detailed steps to identify trending topics leveraging dynamic LDA modeling and other textual analysis tools, please check the data analysis section below.

Hypotheses:

1) We expect there are stable topics that embrace lasting popularity among economic researchers and also emerging topics from time to time. We expect to see fluctuations in these topics' frequencies.
2) We expect different institutions will have different specialties and their own research interests. We expect university-institutions will have broader research topics than commercial institutions or government's institutions.

To answer the third question, we aim to find the relationship between the centralities in the network and the research topics. Do centralities in the network research the same topics? For a centrality, who are parallel peers, who are distinct peers?

Strategy:

We measured the similarity of the research topics structure between two institutions as :

$$\Sigma\,|proportion\ of\ topic\ i\ in\ institution\ A\ -\ proportion\ of\ topic\ i\ in\ institution\ B|$$

The higher the discrepancy score, the lower the similarity of the research topics structure of two institutions.

Hypotheses

1) We expect that the intuitions in the center of the research collaboration network (i,e, the intuitions that are highly engaged in research collaborations) have similar research topics. Top economic research institutions like NYU, Berkeley, and UChicago may have similar structure of the research topics.
2) We expect to see non-academic centralities and academic centralities have large discrepancy in their research topics of interest.

6. Data Source

1. Top 20 Economics Journals database

We scraped article-level information from the top 20 economics journals ranked by the 2017 - 2021 h5-index. For every journal, we collected all available articles - from the journal's inception until February 2023 (or the latest issue). We collect a total of 79,565 publications from all 20 journals and the time span is 1886 to 2023.

We prioritized crawling data from the official website of each journal leveraging requests (if we are lucky) or selenium. However, there are cases that the website prohibits the use of crawlers (such as sciencedirect.com) or the website has a dynamic design that it is too tedious to use selenium, in which case we crawl data from some other authoritative third-party economic journal database websites (such as https://ideas.repec.org/ or econpapers.repec.org/). The limitation of using a third-party database is that some attributes of the paper are missing, most commonly the DOI of the paper and the JEL classification label. This does not prove to be a problem other than the DOI. We could also accurately identify a paper based on its full title. And since we are planning to implement Dynamic topic modeling on the abstract for the classification of the paper, JEL labels are not necessary.

List of info collected:
Title, Authors, Abstract of the paper, and
Cite: Citation format for the paper itself.
DOI: Digital Object Unique Identifier.
VOL: Volume of the journal.
URL: The link of the webpage of the paper.
Published time: Published year and month.
JEL (if available)

2. Google scholar author database

To locate the authors of the papers collected and gather their academic attributes, we take the following steps for each paper: (1) use *selenium* to input the DOI (or the articles' title if DOI is not available) into the search bar of Google Scholar; (2) retrieve the embedded href to the authors' Google Scholar page from the first result. Then for all the links of the authors we obtained, we requested the source page and applied an html parser with regular expressions to obtain author-level data (as listed below). From 79,565 articles we collected, we gathered information from almost 20,000 authors in the google scholar.

The major limitation of using Google Scholar is that there is a subset of authors who do not have a Google Scholar page, possibly those that do not have many publications. This would definitely bias the pattern we are hoping to investigate somehow. The main reason we continue using Google scholar data is that we couldn't find any other third-party database that contains a more comprehensive record of "affiliation" attributes than Google Scholar, which is one of the key pieces of information we are interested in.

Note: Please see the third data source on how we are planning to remedy the incompleteness of this data source

List of info collected:
DOI: The DOI of the paper we used for searching (matching purpose)
Name: The name of the author
Authorid: The Google Scholar user id of the author
Affiliations: The institution where the author currently works (normally a university)
Email address: the suffix of the email contains a hint of the author's affiliations
Citations: Number of citations on the author by year
H-index: Total h-index and the h-index since 2018
I10-index: Total i10-index and the i10-index since 2018

3.Semantics Scholar database (https://www.semanticscholar.org/)

To remedy the incompleteness of the Google Scholar database, we supplement our author-level data with Semantics Scholar, which is a third-party academic database providing free-to-use API for data retrieval. By passing the articles' DOI in the query, we could retrieve the h-index and I10-index for all the authors of the article. After experimenting on a small sample of our article data, we find this database more comprehensive (though still not complete) than Google Scholar on these two attributes. Unfortunately, Semantics Scholar has a very poor record of the authors' affiliations. Therefore, we still plan to use Google Scholar data as the main data source and supplement it with the h-index and I10-index for the authors that do not have a Google Scholar page.

List of info collected:
DOI: The DOI of the paper we used for searching (matching purpose)
Name: The name of the author
H-index: Total h-index.
I10-index: Total i10-index.

7. Data cleaning/wrangling

After scraping down all the articles from the top 20 economic journals, we performed the following steps to complete our database.

Constructing Affiliation Dataset:
We identify the author's email suffix and search the suffix in Google Scholar using Selenium. Google Scholar will return the official name of the institution that matches with the suffix. We generate an affiliation id using the first letter in each word of the affiliation name plus its official email address.

Matching Affiliation to an External Ranking:
We matched authors' affiliation to the ranking of economic institutions provided by IDEAS/RePEx using record linkage, and appended a column of "ranking" in our affiliation table. In this matching process, we find 200 exact matches of institution names.

Cleaning of Author, Article, Affiliation Dataset:
For most of the article attributes and authors attributes we collected, we clean the data following a standard procedure - removing N\As or duplicates and extracting titles/labels/index data originally in the form of list, tuple or dictionary to individual data frame columns. The issue date of an article is not directly available, so we use regular expressions to extract from the issue volume attribute. Finally, we order the columns in an easy-to-use way.

Building Database
We then write schema and query to load all the data frames into the SQL database. Four tables were formulated in our econtop.db: 1) author; 2) article; 3) affiliation; 4) author_article. The first 3 datasets are unique in each record. The author and article datasets are linked via the author_article dataset to generate author-article level linkage (We do not include article information in the author dataset and vice versa since it is a many-to-many relationship). The author dataset contains the affiliation id, which is directly linked to the affiliation dataset.

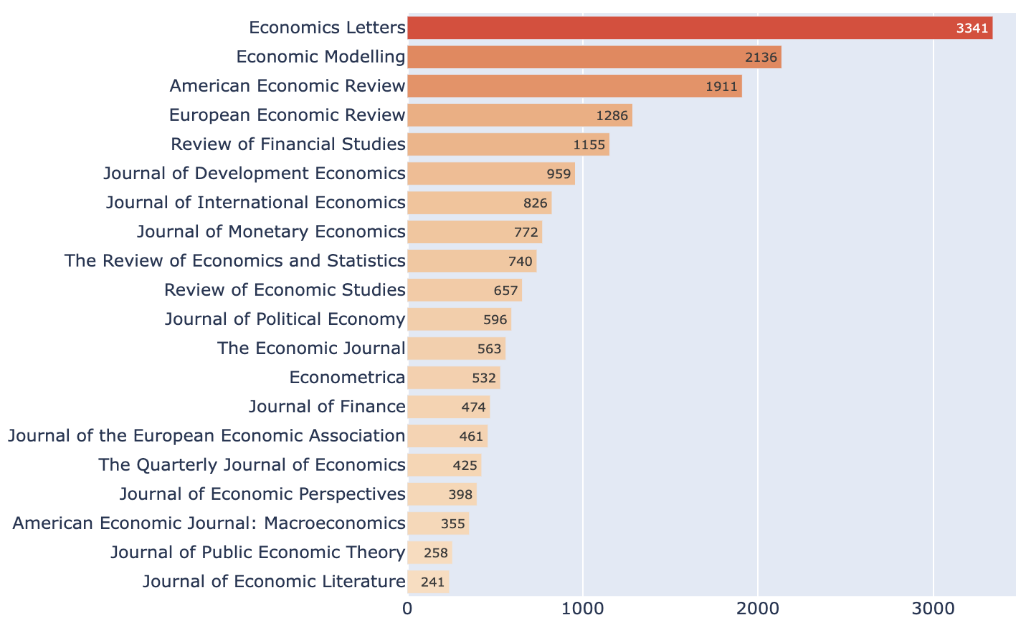
Supplementing missing data with the semantic database:
We then filter the authors that appear in the article table but not in the author table (due to the incompleteness of the Google Scholar database), and load their author-level attributes from the Semantic Scholar database into our econtop_author database.

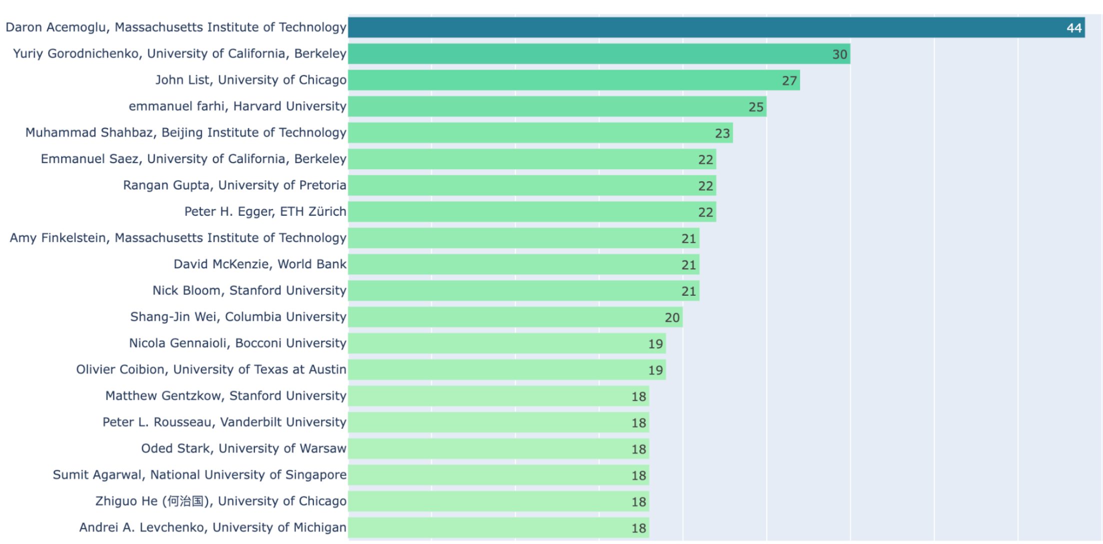
8. Data analysis, visualization and results

1. Descriptive Illustration of Data

Even after constraining the year time to between 2012 and 2022, the number of publications among the top 20 journals is not evenly distributed. This is because some journals only issue four volumes annually, such as AEJ Macroeconomics. And some other issues monthly, such as the Economic Letters. The top 20 economic scholars by a number of publications in top economic journals coincide with our educated guess - all are well-known economists in current times.

Number of Collected Publications from 2012 to 2022 by Journal





2. Collaboration Network Analysis

2.1 Collaboration link and Network Construction:

(a) Individual Level

For a particular article with multiple authors, a collaboration link - undirected linkage is formed between every two authors. For example, for an article with 4 authors A, B, C, D, there will be 6 collaboration links (A,B), (A,C), (A,D), (B,C), (B,D), (C,D). The frequencies of these links will be used to construct author-level network visualization.

(b) Institutional Level

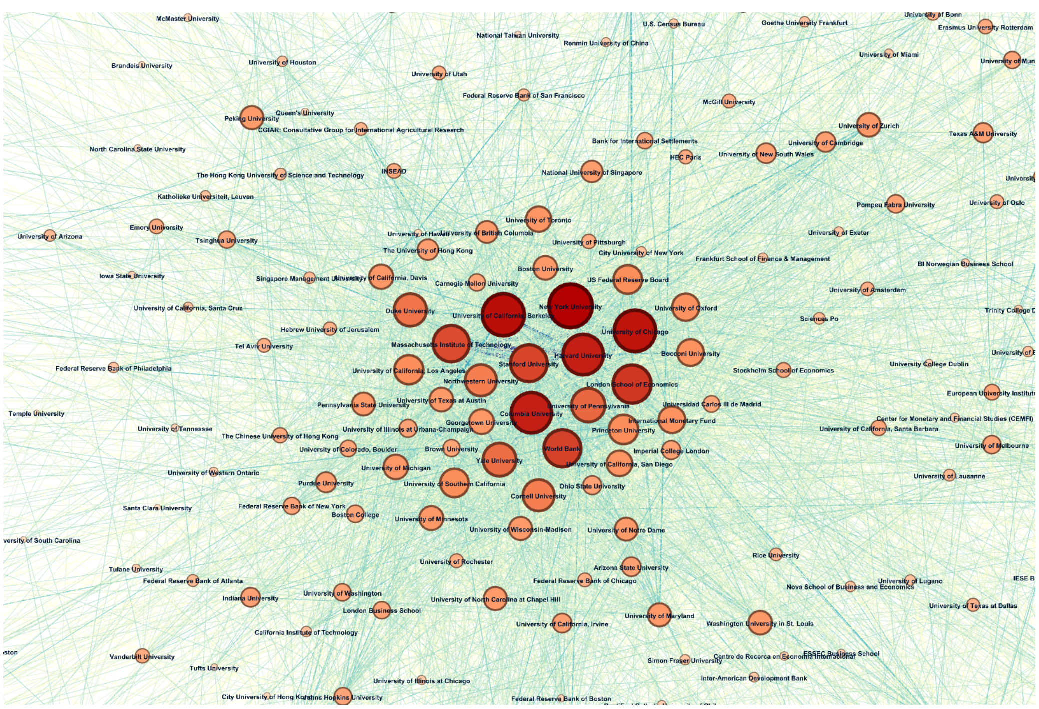
Similarly at institutional level, given any article, a collaboration link is an undirected linkage of two collaborating authors' institutions.These frequencies will be used to construct an institution-level network.

(c) Institution-Tier Level

We define the "tier" of institutions using the top 550 institutions in IDEAS Economic Institutions Ranking. We tentatively classify the institutions into 11 tiers based on their rankings: 1-50, 51-100, 101-150, 151-200… 451-500, 501-550.

2.2 Collaboration Network Overview

We could plot the co-authorship linkage of institutions as an academic network, using every institution as a node and the number of linkages/frequency as the weighted node. The larger and redder the dot is, the more "connected" the institution is with others. The results coincide with our previous knowledge that the central institutions feature New York University, the University of Chicago, Harvard, Columbia, and the London School of Economics.



Running centrality analysis on the network, we find across the three measurements of centrality - degree centrality, betweenness centrality, and eigenvector centrality, the institution that ranks top 10 are generally stable. An interesting finding is that the World Bank takes the first spot when it comes to the betweenness centrality but ranks relatively low in other measurements. This result is quite intuitive as betweenness centrality reflects the ability of the node to influence other nodes and the World Bank operates as a platform to connect scholars across institutions.

2.3 Cross-rank Score Measurement

To measure the openness to collaboration for every institution in a more quantitative way, we propose the following strategies to construct a metric called "cross-rank score".To put it simply, the metric is a weighted score of collaboration links strength for one institution or tier. More specifically:

(a) Institution Level

We focus on the institutions that are ranked by the IDEAS Economic Research. In total, we matched around 200 institutions. Within this community, we compute the cross-rank score for each institution. Given institution A, we calculate its collaboration number with institutions in the community. We allow within-institution collaboration, that is, we consider co-authors from the same institutions as valid collaboration. To compute the links, we go through each article that contains authors from institution A and count the number of institutional collaboration links of A with institutions in this single article. We finally obtained all the collaboration links involving A, with the format of [A, X, number of collaborations]. For each institution X, we obtain a score equal |Rank of A - Rank of X|*the number of collaborations. We sum all the scores and divide them by total collaborations to get the cross-rank score for institution A. For example, we have [A, A, 9], [A, B, 2], where A is ranked 1, and B is ranked 10. The cross-rank score for A is thus (|1-1|*9+|1-10|*2)/11=1.64.

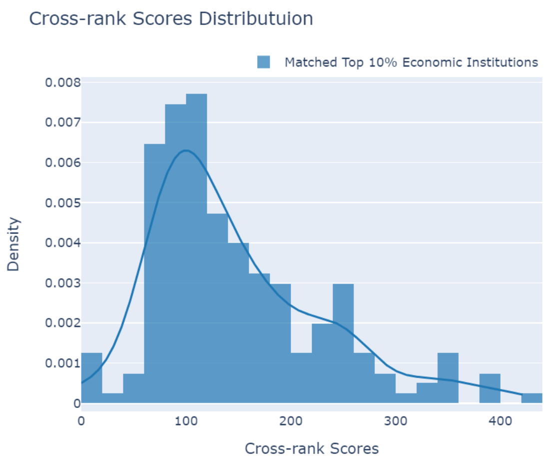
(b) Tier Level

At the tier level, the procedure is almost the same as at the institution level. The only difference is we further cluster these 200 institutions into tiers by every 50 ranks. We label the Top 1-50 as Tier 0, Top 51-100 as Tier 1, etc., and we have 11 tiers in total. Note that we do not have all institutions with ranks in our dataset simply because for these institutions, though their ranks are high, they may not have publications in the past 10 years in Top journals. Like before, we count all tier collaboration links, then calculate the cross-rank score for each tier. The rank difference we use is a Tier difference times 50. For example, we have [Tier 0, Tier 0, 100], [Tier 0, Tier 8, 8], then the cross-rank score for Tier 0 is (|0-0|*100+|0-8*50|*8)/108=29.63.

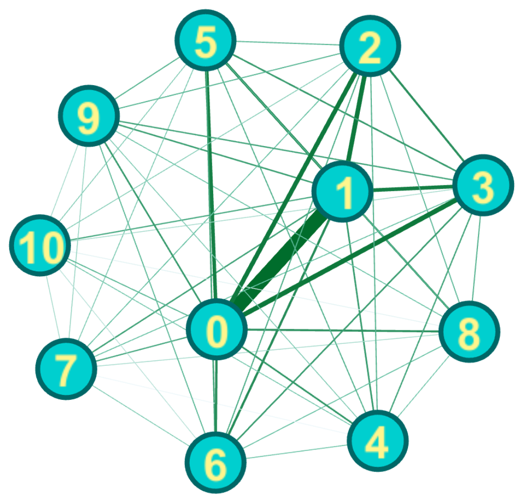
2.4 Quantitative Cross-rank Collaboration analysis

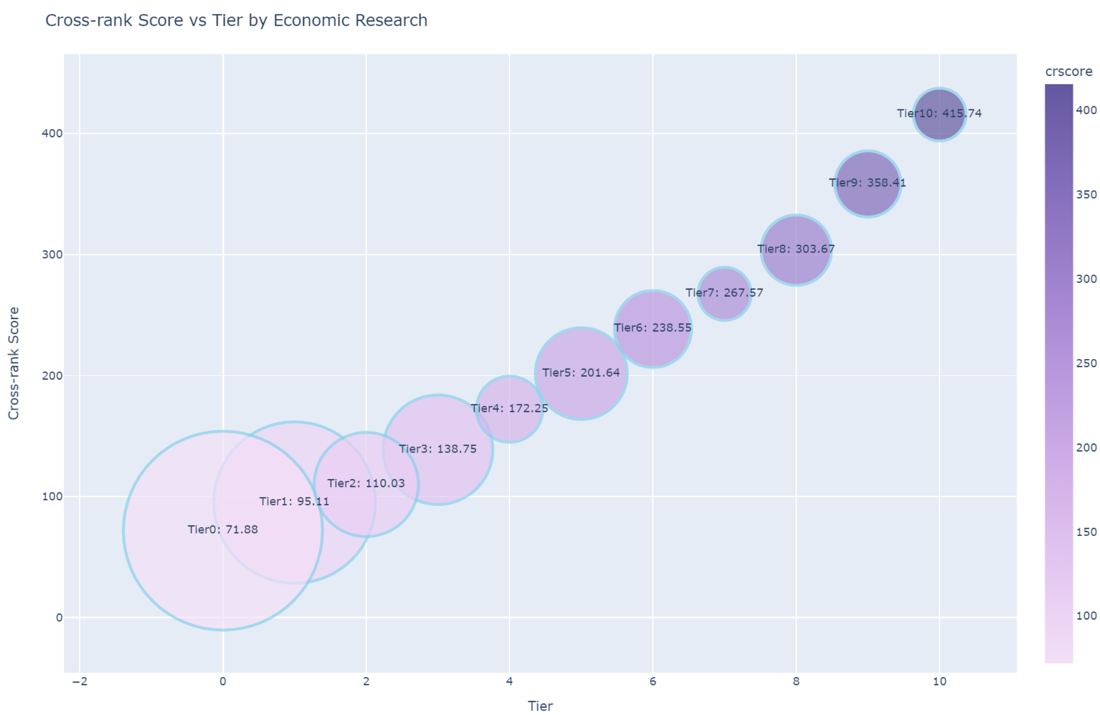
The density distribution for the collaboration score is given below. The mode is around 100, meaning that a larger portion of institutions is collaborating with others that are plus or minus 100 rankings. To see the relationship between the ranking of an institution and its tendency to collaborate across ranks, we plot the cross-rank score of every institution against the ranking of the institution itself. Note that the available data in this analysis step is constrained by the number of institutions that are matched with the external rankings from IDEAS, i.e. 200.

We could observe from the graph that there appears to be a negative relationship between the ranking status of the institutions (a smaller / higher rank) and the likelihood of collaborating with institutions that are far away from the ranking itself, resulting in a smaller cross-rank score. In other words, as far as our data is concerned, the more reputable institutions are more likely to collaborate with peer institutions that also rank high, whereas the institutions that rank later are more likely to collaborate with those that rank much higher. This finding reflects the pattern that scholars from less distinguished institutions commonly work with scholars from higher level institutions to get their paper published in top economics journals. The suggestion of this pattern coincides with the conventional belief that the publication of top-tier economic articles is dominated by scholars from top economic research institutions.



Similar analysis on institutional tier-level reflects a stable pattern. Tiers level network analysis shows a much larger amount of collaborations between the top 2 tiers than the collaborations between any other two tiers. The relationship between the tier level and the cross-rank collaboration tendency is identical to that for institutional level.



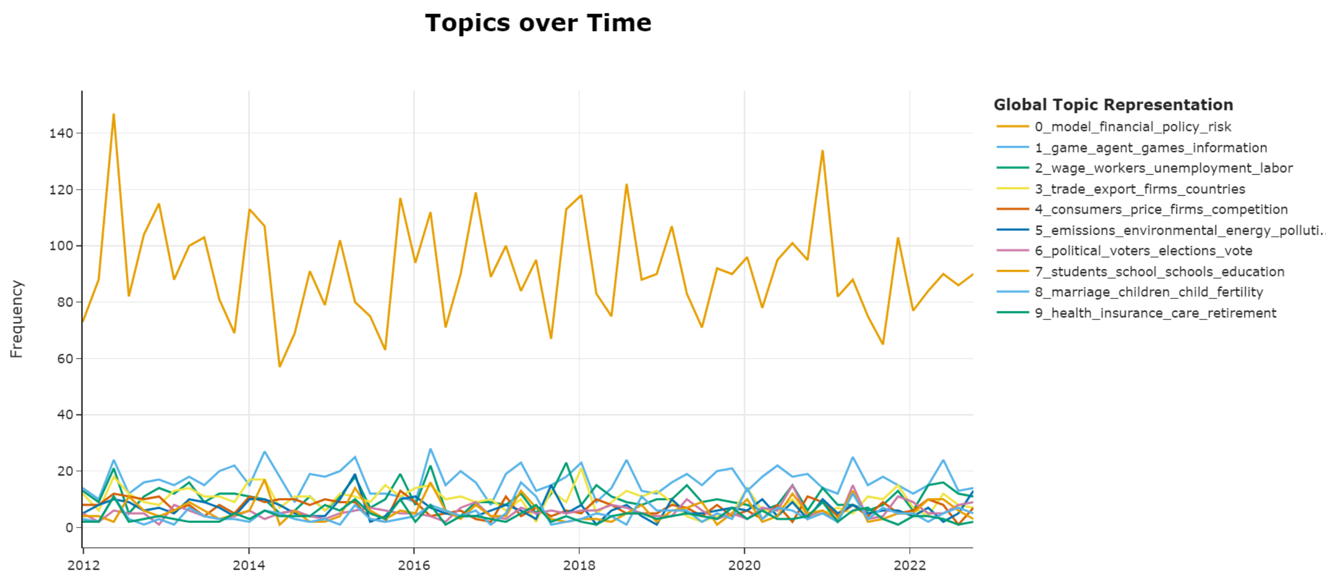


3. Dynamic Topic Modeling

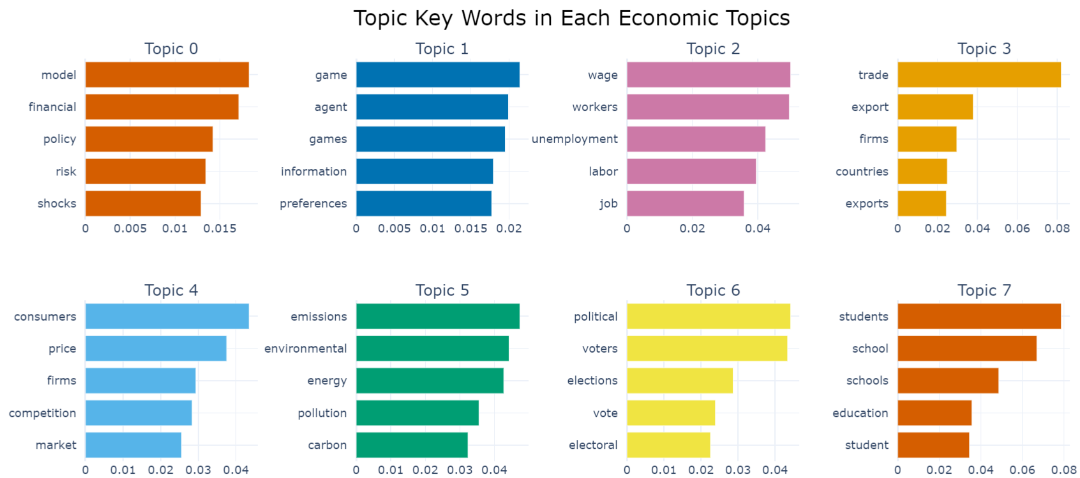
We use an emerging technique called BERTopic that uses language embedding and class-based TF-IDF for our topic modeling analysis. Compared with a traditional Gensim LDA, it has several advantages, including 1) easy to handle data (only need to clear stop words then re-concatenate them into paragraphs); 2): much less training time; 3) no need to specify number of topics to train; 4) relatively stable training results across runs; 5) comprehensive integration of visualization tools.

3.1 General Dynamic TM on Abstracts from 2012 to 2022

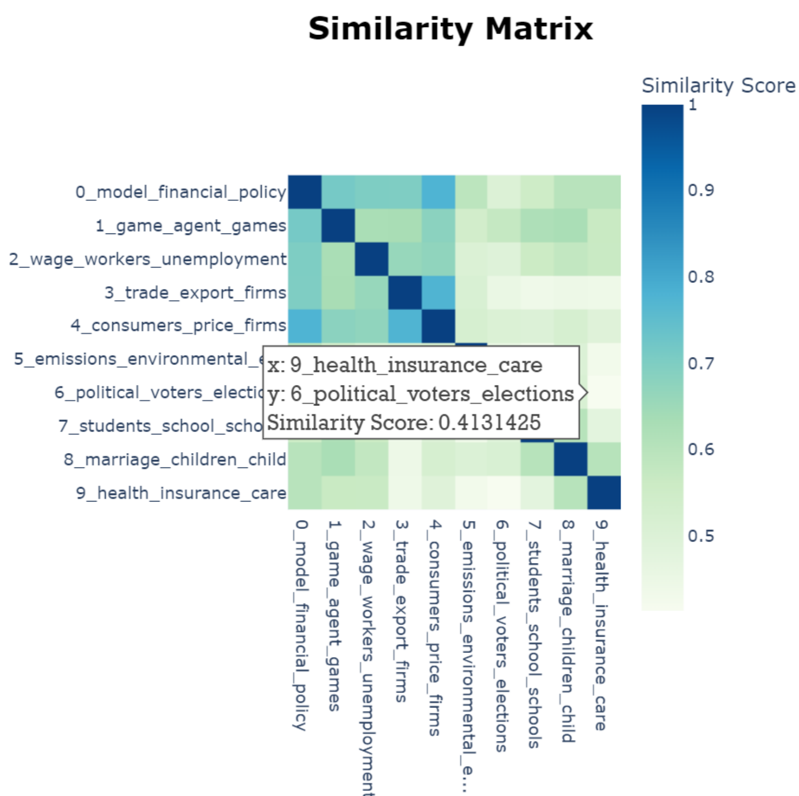
We randomly sample 1,400 articles from each year to form a dataset of 15,400 abstracts from 2012 to 2022. After several runs of training and tuning parameters, we finally used the Euclidean metric, number of components equal 5, number of neighbors equal 15, minimum number of topics equal 70, minimum distance equal 0.01 as function arguments. We obtained the training result as below:

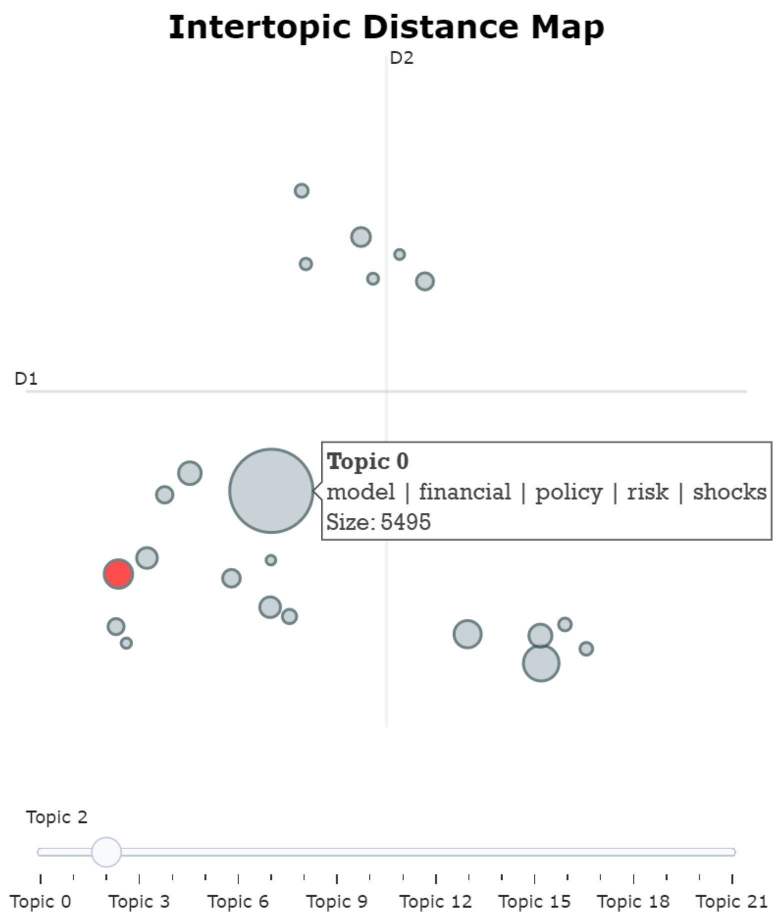


From the above, we see there is one dominating Topic 0 - model, financial, policy, risk, with fluctuations at different time points. Other topics remain at relatively low frequencies. We can examine the keywords and their corresponding weights for each topic as below:



From the training keywords, the results make much sense from an economic point of view: Topic 1 is about Game Theory, Topic 2 is about Labor Economics, Topic 3 is concerned with International Trades, etc. In order to see how similar these topics are to each other, and how our model training result is, we can refer to the similarity matrix and intertopic distance map:

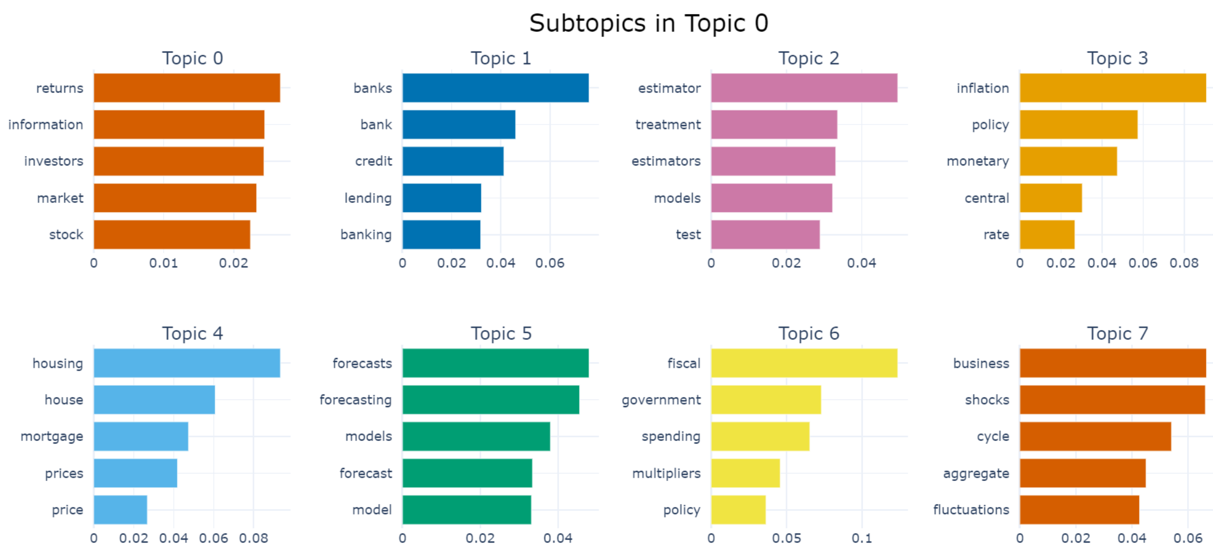




For example, Topic 6 (political, voters, elections), and Topic 9 (health, insurance, care) have a low cosine similarity score of 0.4131, which means these 2 topics are uncorrelated. For a whole picture of the distances among topics, we see on the map all topics do not overlap with each other, which is a signal for good model training.

3.2 Further Splitting of Topic 0 into Subtopics

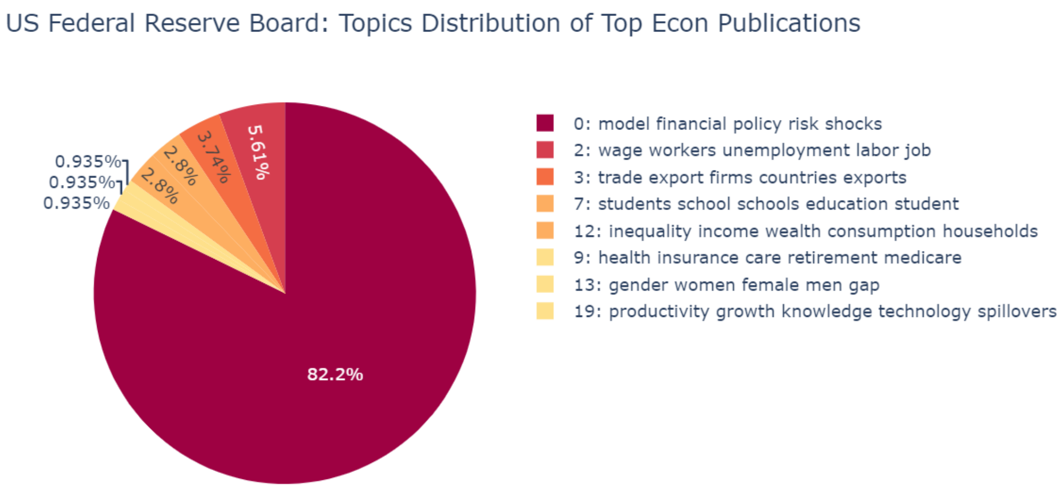
Note that the Topic 0 has a dominating weight compared with other topics, though it makes sense from an economic point of view. Further splitting of this Topic into subtopics may help us understand the details within this broad topic. We retrain the model using all the abstracts assigned with label Topic 0. The keywords we obtained are:

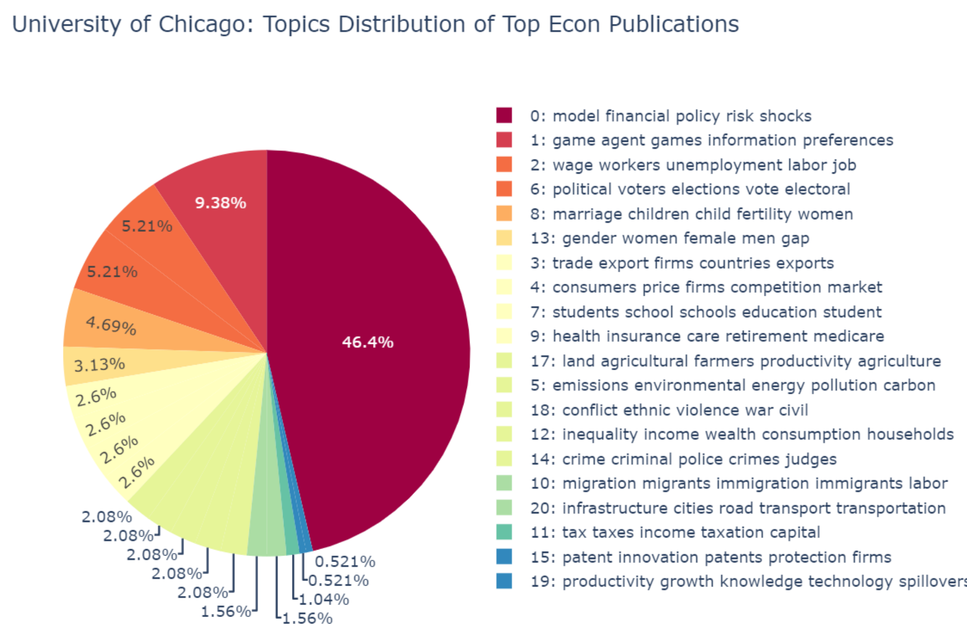


The model further classifies the documents into topics like investment, banking, estimation, inflation, etc. This enriches our knowledge about the most trending economic research topics within the broad category of "model, financial policy, shocks".

3.3 Topic Distributions of Certain Institutions

With the output of our TM, we can assign each document with the most relevant topic it belongs to. Then we can compare for different institutions, how their topic distributions compare with each other. Take the examples of Fed and UChicago:

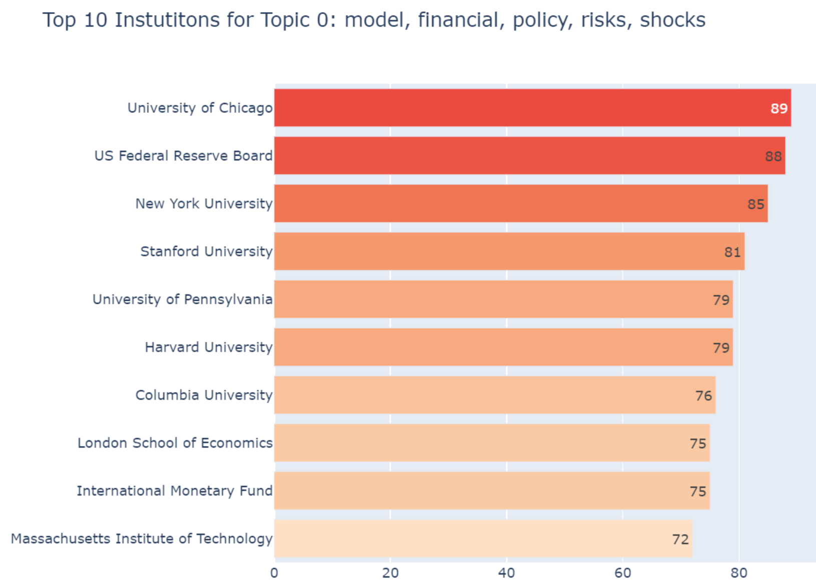


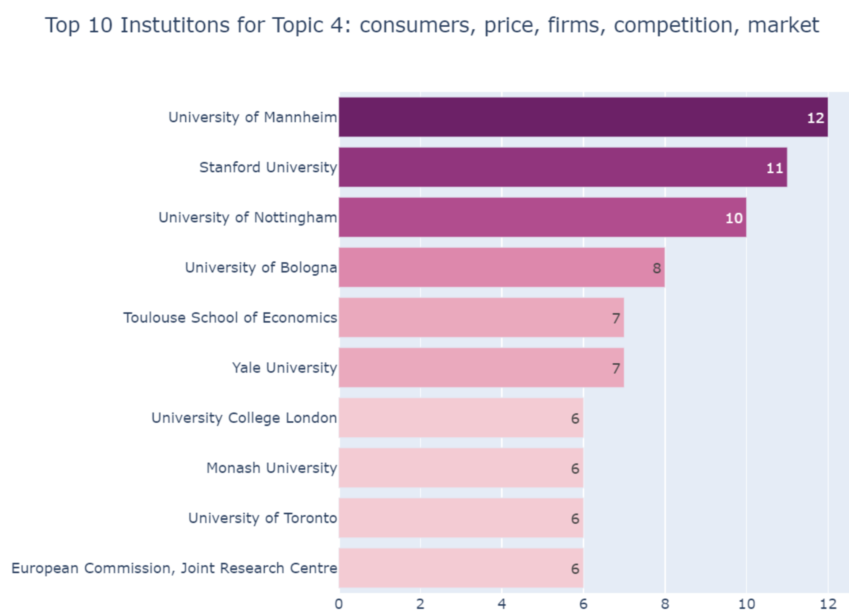


From the above pie charts, we see over 82% of the Fed's publications are about financial policies, shocks, risks, which focus on macroeconomic policies. By contrast, UChicago has a much more balanced distribution of research topics, which almost covers all of 22 topics we identify in our model. We see that although economic topics remain relatively stable across time, they do vary at the institutional level. Different institutions will have their own speciality areas of economic research.

3.4 Top Contributors under Different Economic Topics

Besides examining topic distributions for specific institutions, we can also see for each topic, what institutions contribute the most. Take the example of Topic 0 and Topic 5:

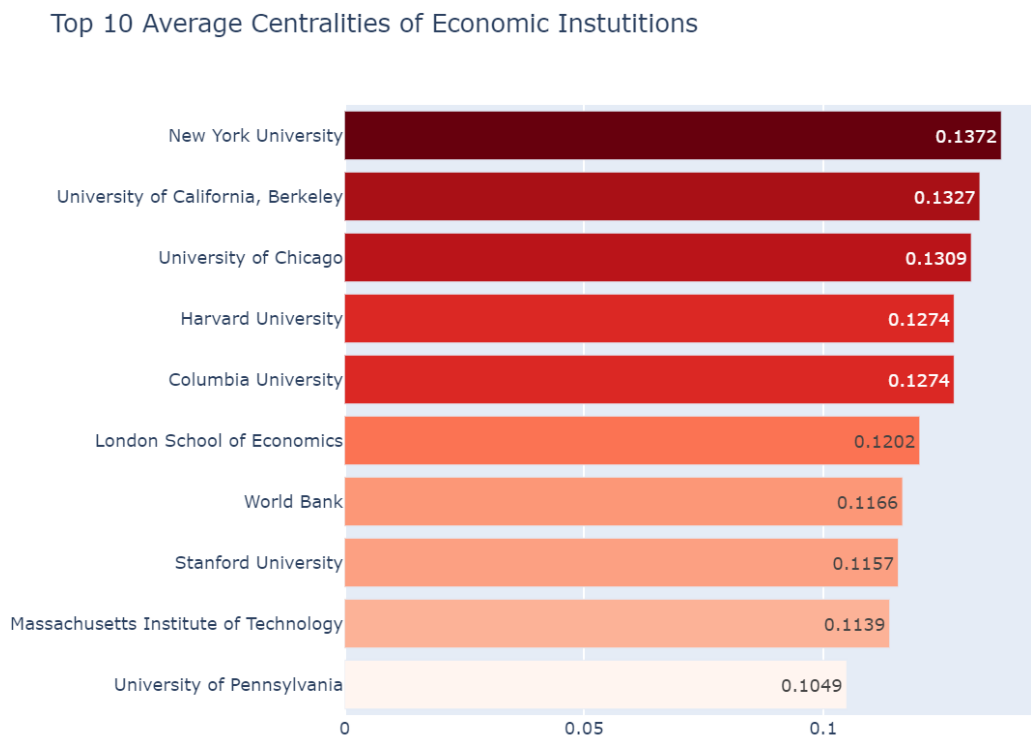




As for Topic 0, we see American institutions almost dominate the Top 10, which are all big names in economic research. However, for Topic 4, we see University of Mannheim from German tops, and also other European institutions like Nottingham, UCL, etc. It shows a completely different story from Topic 0.

4. Combining SNA and TM

In this section, we combine the results from both SNA and TM to answer a specific research question: do centralities in the collaboration network research the same topics? Recall the Top 10 centralities we get from the previous session using an average of 3 centrality measurements:



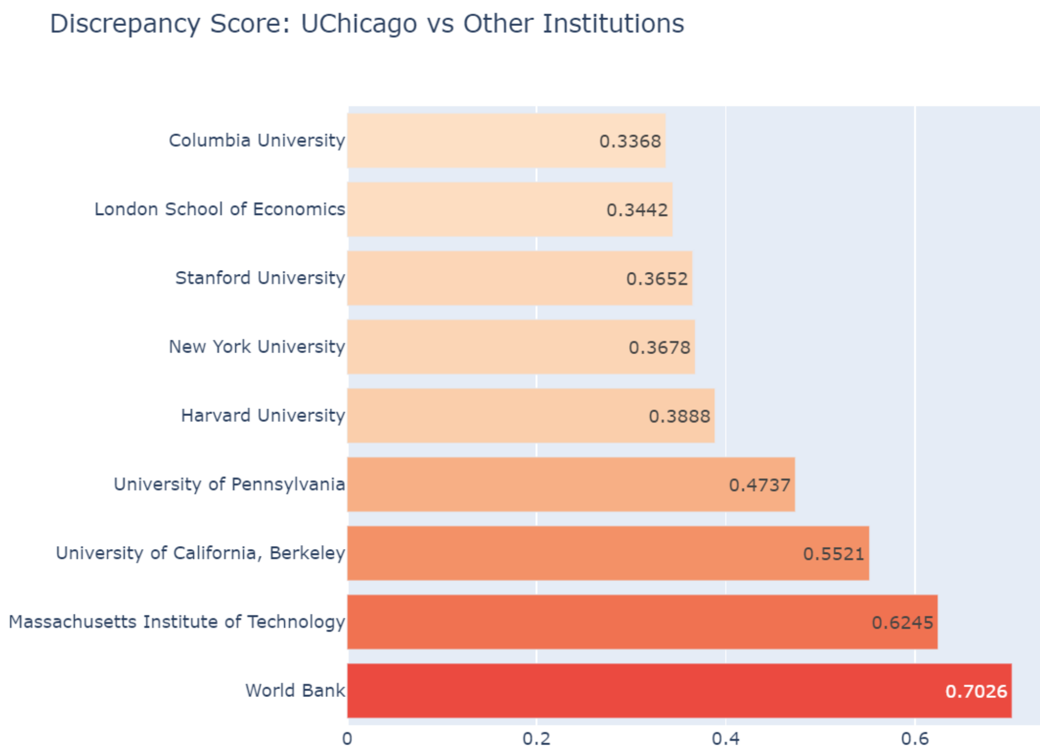
We want to see how (dis)similar these institutions' topics are by looking at a measurement of discrepancy score. We define the discrepancy score as follows:

Suppose we have 2 institutions A and B, with n attributes Xij.

- 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 topics.
- Discrepancy Score between A and B := sum of |XAj - XBj| over j

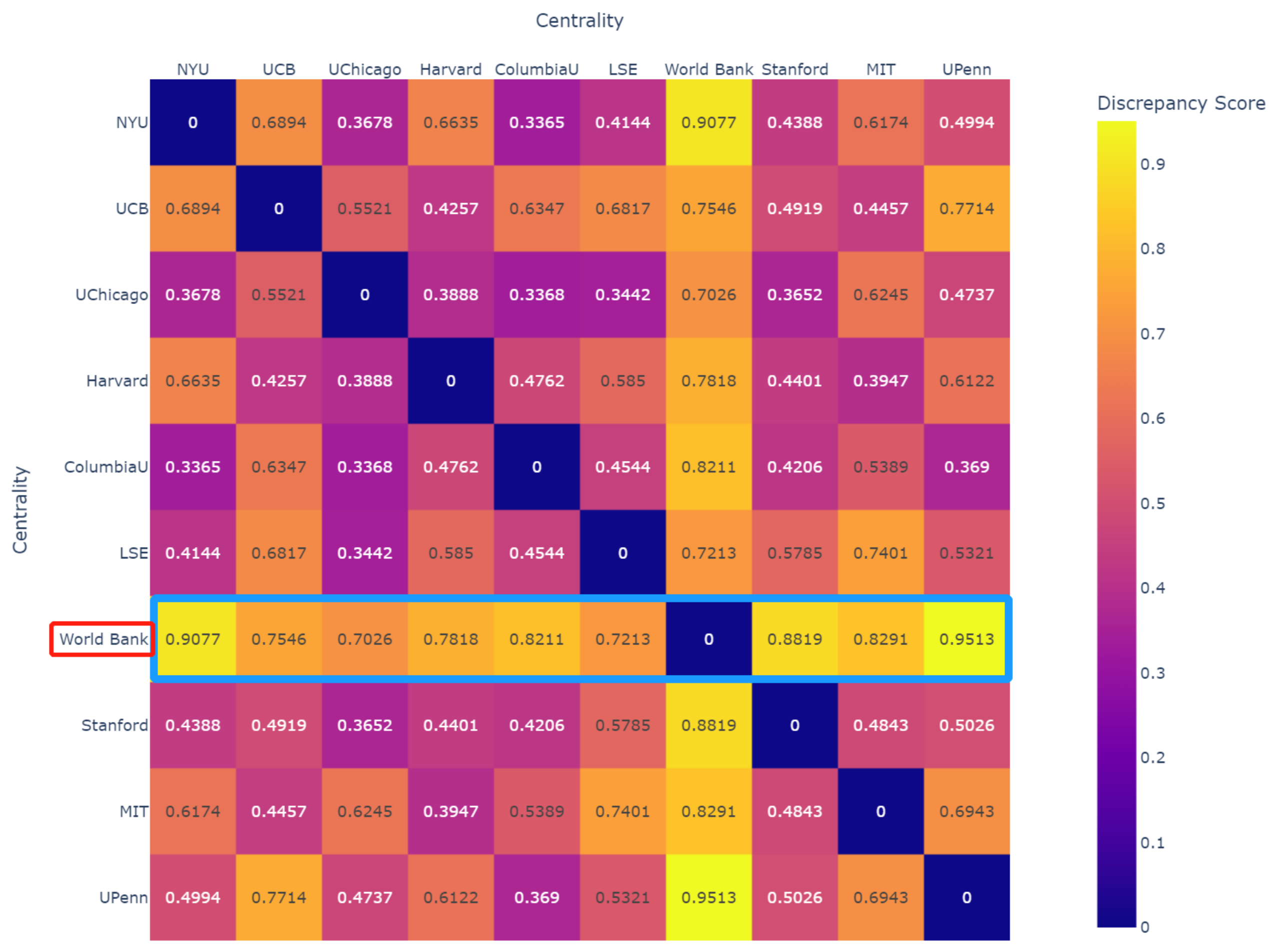
For example, A has a distribution of [Topic 0: 0.7, Topic 1: 0.3], and B has a distribution of [Topic 0: 1, Topic 1: 0]. Thus the discrepancy score is equal to |0.7-1|+|0.3-0|=0.6.

We can first examine how similar UChicago is against other centralities in the network:



We see UChicago is very similar to Columbia University and LSE in research topics, but not so similar as MIT and world bank.

In order to see every pair's discrepancy score among all 10 centralities, we have to calculate a discrepancy score matrix and visualize this matrix through a heatmap.



From the heatmap, the World Bank is clearly an outlier of this centrality community since its discrepancy scores are all very high against other university-institutions. Besides the World Bank, for the remaining centralities, their discrepancy scores are generally low which means their topics are comparable to each other.

Appendix

Calero C, van Leeuwen T, Tijssen R. Research cooperation within the bio-pharmaceutical industry: Network analyses of co-publications within and between firms[J]. Scientometrics, 2007, 71(1): 87-99.

Jones C I, Williams J C. Measuring the social return to R&D[J]. The Quarterly Journal of Economics, 1998, 113(4): 1119-1135.

Larivière V, Macaluso B, Mongeon P, et al. Vanishing industries and the rising monopoly of universities in published research[J]. PloS one, 2018, 13(8): e0202120.

Lazega E, Jourda M T, Mounier L, et al. Catching up with big fish in the big pond? Multi-level network analysis through linked design[J]. Social Networks, 2008, 30(2): 159-176.

Mansfield E. Academic research and industrial innovation[J]. Research policy, 1991, 20(1): 1-12.

Meishar-Tal H, Pieterse E. Why do academics use academic social networking sites?[J]. International Review of Research in Open and Distributed Learning, 2017, 18(1): 1-22.

Nicolaou N, Birley S. Academic networks in a trichotomous categorisation of university spinouts[J]. Journal of business venturing, 2003, 18(3): 333-359.

Ramsden P. Describing and explaining research productivity[J]. Higher education, 1994, 28(2): 207-226.

Tang J, Zhang J, Yao L, et al. Arnetminer: extraction and mining of academic social networks[C]//Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. 2008: 990-998.

Uddin S, Hossain L, Abbasi A, et al. Trend and efficiency analysis of co-authorship network[J]. scientometrics, 2012, 90(2): 687-699.

Van den Brink M, Benschop Y. Gender in academic networking: The role of gatekeepers in professorial recruitment[J]. Journal of Management Studies, 2014, 51(3): 460-492.

Zhou X, Liang W, Kevin I, et al. Academic influence aware and multidimensional network analysis for research collaboration navigation based on scholarly big data[J]. IEEE Transactions on Emerging Topics in Computing, 2018, 9(1): 246-257.