Project 2: Web Scraping, Text Processing, API Integration, and Graph Analysis

Donald Moratz*, Shanze Fatima Rauf†, Daniel Wang‡and Ruiwen Yang§

March 2025

1 Project Summary

This project establishes an extension of the repository of research outputs leveraging Federal Statistical Research Data Centers (FSRDC). Extant sources provide an account of these outputs, but high levels of missingness makes it difficult to track the extent and impact of FSRDC-related work. By integrating web scraping tools, public APIs (specifically OpenAlex), and advanced text-processing steps, we systematically locate and confirm whether a publication uses FSRDC microdata through the presence of restricted-data acknowledgments, RDC location references, and disclosure-review statements. Our approach employs both exact and fuzzy name matching to handle the inherent variability of author and project naming conventions, thereby minimizing duplicate entries.

A major contribution of this effort is a combined dataset of 1,444 unique publications, with 1,391 drawn from the OpenAlex API and 53 collected via targeted web scraping of RePEc Census Working Papers. Throughout data collection, we pay particular attention to screening out invalid or incomplete entries, ensuring that only outputs meeting the FSRDC-use criteria remain. This rigorous curation helps shape a comprehensive record of the diverse scholarly outputs (articles, working papers, dissertations, etc.) generated under the FSRDC umbrella. In turn, the resulting dataset lays a firm groundwork for advanced analysis and serves as a valuable resource for stakeholders interested in the scope and evolution of FSRDC-supported research.

Network analysis (Section 5) demonstrates that the newly identified research outputs form a robust web of scholarly interconnections. An undirected graph of these publications, constructed based on shared authors, topics, or locations, reveals certain "hub" or "bridge" papers with notably high betweenness centrality — an indication that they link otherwise separate clusters of research. Moreover, the largest connected component highlights how outputs related to demographic data and household structures cluster around major RDC sites, with the Boston RDC comprising a particularly sizable node set. These findings underscore the potential for a few influential works to shape networks of collaboration and topic development across FSRDC projects. We also implement discrete event simulations to explore editorial workflows, highlighting potential bottlenecks in the peer-review process and the efficacy gains achievable by allocating additional editorial resources. We find that the likely highest bottleneck point for review is the initial editor process and recommend

*dmoratz@sas.upenn.edu
†sfrauf@sas.upenn.edu
‡wang50@seas.upenn.edu
§ruiweny@seas.upenn.edu

that journals expand their editorial teams to increase the speed at which papers are distributed to independent reviewers.

The analysis in Section 6 extends from descriptive statistics into regression modeling, relying on BERT-based topic classification to categorize each publication. We note that the existing dataset misses papers throughout the timeseries, with the bulk of the newly found papers identified between 2008 and 2024. We are less successful at identifying new working papers than previous efforts, but we find significant numbers of journal article publications that are new. In our analysis, we find evidence that while papers on demographic data and household structure have been popular since 2005, there was an uptick later in the time series in papers focused on both Growth and Rural Infastructure. Our regression analysis also produces interesting results. Notably, we observe that time since publication significantly interacts with certain topics—such as efficiency, productivity, and economic growth—to drive citation counts, although the overall explanatory power of these models leaves ample room for external factors like journal prestige and author networks.

This report is organized as follows. Section 2 provides details on the web scraping methods and name normalization strategies. Section 3 describes the construction of the unified dataset, including how various sources (e.g., GitHub's All Metadata CSV, RePEc) are merged and de-duplicated. Section 4 covers data processing and entity resolution, describing the exact and fuzzy matching logic used to verify each publication's eligibility. Section 5 shifts focus to graph construction, network metrics, and community detection, while Section 6 delves into broader empirical findings—from topic-based citation patterns to editorial workflow simulations.

2 Web Scraping

To identify potential research outputs related to the Federal Statistical Research Data Centers (FSRDC), we developed a Python-based web scraping pipeline that systematically extracts metadata from external publication repositories, using GitHub metadata as the seed reference.

We began by leveraging the `All_Metadata.csv` file from GitHub, which contains a list of principal investigators (PIs) and researchers currently or previously involved in FSRDC projects. According to project requirements, only outputs with at least one co-author matching a researcher or PI from this file were retained. To ensure accurate matching, we normalized names using regular expressions and performed string comparisons on cleaned author fields.

We scraped abstracts, authors, and publication metadata from the RePEc Census Working Papers Series (`https://ideas.repec.org/s/cen/wpaper.html`) using the `requests` and `BeautifulSoup` libraries. For each paper, we applied rigorous cleaning procedures including deduplication, whitespace removal, and exclusion of malformed entries (e.g., "Registered:" placeholders in author lists). HTTP error handling, response validation, and rate limiting via `time.sleep()` were implemented to ensure reliability and compliance with external site policies.

To determine whether a research output was likely derived from restricted FSRDC microdata, we designed a multi-layered keyword matching system based on the project's selection criteria. Specifically, we parsed abstracts and identified the presence of the following:

- **Acknowledgments:** Mentions of the Census Bureau, FSRDC, or specific RDCs (e.g., Michigan RDC, Texas RDC);
- **Data Description:** References to restricted microdata sources such as "IRS", "BEA", or "confidential microdata";
- **Disclosure Review:** Statements confirming confidentiality or disclosure review (e.g., "disclosure review" or "reviewed for disclosure");

- **RDC Mentions:** Named RDC locations using a curated list of U.S. centers;
- **Dataset Mentions:** Identifiable references to known restricted datasets (e.g., *American Community Survey*, *Annual Survey of Manufactures*, *Census of Construction Industries*, etc.) as listed in the project's metadata file.

Outputs satisfying at least one of these criteria and containing an eligible co-author were retained for further analysis.

To enrich the data, we also queried the OpenAlex API (`https://openalex.org/`) using the paper title to retrieve metadata such as DOI, citation count, publication type, and topical concepts.

In summary, our scraping pipeline successfully balances precision in researcher validation, robust pattern matching based on project logic, and responsible data collection practices. The resulting dataset serves as a high-confidence foundation for subsequent analysis steps in the project. As a result of the filtering and enrichment process, the final dataset contains 64 research outputs.

3 Data Construction

The data used in this project came from this spreadsheet provided as part of the University of Pennsylvania course CIT5900_002. This data includes 1,735 Research Outputs related to FSRDC projects, including the Output Title, the Output Bibliography, the Year, and the supporting Agency. In order to conduct this research, we began by collecting the relevant data via scraping the Github link provided for this project. We utilized two spreadsheets from the data, "researchers" and "datasets". We merge these two datasets to form a single dataset that is the combination of researchers who have worked on FSRDC work and the datasets that they accessed as part of their projects. To account for data idiosyncrasies, we lengthen the data to ensure all individuals listed as PIs are also listed as researchers, along with the correct dataset terms. We then extend the data further by taking each researcher with the location they worked at (allowing multiple locations where appropriate) and added the name of the location in conjunction with the string "RDC" to the dataset terms. Finally, we did the same for each researcher and the term "FSRDC" and "Census Bureau". This combined dataset gave us just over 48,200 unique researcher name plus dataset terms to search. We supplemented the data provided in the aforementioned spreadsheet with data from OpenAlex. As part of this project, we explored several different APIs, including CORE API, BASE API, NSF's PAR and ORCID. We ultimately settled on OpenAlex as the main target of our API data collection effort. OpenAlex is an open, comprehensive, and free catalog of the global research system, and importantly, the search tools provided by the site allowed us the opportunity to conduct testing on our API queries by generating the exact API query using their inbuilt search feature and checking it against the API query generated by our script. This allowed us to conduct pre-testing before running our script, since execution time takes roughly 10 hours.

OpenAlex provides detailed data on scholarly articles, authors, institutions, venues, and research topics. We used OpenAlex's free API to search the combination of keywords and dataset terms. We counted a paper as an FSRDC output if the full text contains both one of our researchers as well as one of the dataset terms (including the mention of the RDC, FSRDC or Census Bureau). We created a script that searches for each author to find their unique author ID and then searches all of the works by that author for the dataset terms listed in our generated csv. To cut down on the number of queries, we were able to combine dataset terms to conduct multiple searchers at once- within a limit. To ensure accurate searching, we discovered that limiting the combined length of the string of the dataset terms allowed smooth execution of the code. As a result, we were able to reduce our roughly 48,000 unique combinations to just over 15,000 API searches. For each

search, we found title, dois, abstracts, keywords, other authors, author affiliations, publication year and date, number of citations, the publishing journal (where appropriate), the type of output (i.e. article, book chapter, etc) and a record of the dataset term that was found in the text. This final step allows us to confirm that the found paper indeed belongs in the set according to the criteria discussed in section 2. Accounting for deduplification of title, OpenAlex had 1,661 unique outputs. We output the results of our search into a csv called *openalex_researcher_datasets_matches.csv*. This csv includes 16 columns, including those above, it also has the researcher, the dataset terms used in the search that returned the paper and the RDC associated with that research (based on the RDC recorded in the original *AllMetadata.xlsx*. This dataset, in conjunction with the results of our 64 web scraped papers serve as the basis of our analysis.

4 Data Processing and Entity Resolution

4.1 Discrepancies between the dataset and webscraping results

Retrieved data compared to 2024 dataset:

- 2024 dataset has information about both the projects and the outputs; however, the date we retrieved from the Open Alex API and web scraping only contains information regarding the research output without information about the project.
- In the 2024 dataset, RDC information is directly available, but this piece of information is not available directly through the Open Alex API and web scraping.
- In the 2024 dataset, it only has one column regarding Project PI, which contains only one single person, but Open Alex API and web scraping does not have this information available; it has researcher (one person) and authors (more than one person) information instead. And the researcher may appear in the authors, but it is not guaranteed. In terms of the name format, they have different formats even for the same person, such as missing middle name, middle initial vs full middle name, first name first, or first name last.
- In the 2024 dataset, there is no abstract or dataset information, but they could be collected using the Open Alex API or web scraping
- In the 2024 dataset, it only has the publication year for research outputs. However, in the Open Alex API or web scraping, it can give more detailed publication dates instead of just the year
- There are different types of output used between the two sources:
  - 2024 dataset has the following 11 types of outputs: book chapter, blog, Ph.D. dissertation, dataset, article, mimeo, Master's thesis, report, software, CES Technical Note, working paper
  - API data has the following 12 types of outputs: article, preprint, report, dataset, review, book-chapter, other, book, dissertation, editorial, paratext, letter
  - Web scraping data has the following 7 types of outputs: article, preprint, review, report, book-chapter, book, paratext
  - Among these types, they only share 3 categories in common

– Affiliations and author ID are only available when extracting the data from the Open Alex API, not through web scraping. In general, the OpenAlex API could give us more information than web scraping since it's highly structured and already processed the data before putting the information into the API

4.2 Matching Methodology

For each retrieved research outputs dataset (API/ webscraping):

1. Filter by years first: 2024 research output dataset only has outputs published between 2000 and 2024, any research outputs that were published outside this year range are unique records

2. Filter with fuzzy matching: First of all, two research outputs (new research output from our retrieved data and old research output from the 2024 dataset) are considered as the same research output if they satisfy the following conditions:

   - Two research outputs are published in the same year
   - Old research output has the PI listed as a researcher or one of the authors listed
   - Two research outputs have the same title

The rationale for this definition is due to the following reasons:
Even though two research outputs may share a very similar title, but if they are published in different years or by different teams, they are very likely to be two different outputs instead of being considered as one research output.

In reality, it is easy to identify if two research outputs were published in the same year using exact comparison since there is no ambiguity in comparing two numbers. However, it is hard to tell if two outputs have the same titles or PIs using exact comparison, since minor change in titles prevent exact matching. Similarly, the same person represented by a single PI name could show up wth a slightly different name in the newly generated dataset (for example, one dataset could list a full middle name, while the other just lists a single middle initial). To account for this difficulty, we incorporate fuzzy matching in both our author name and title comparison.

Fuzzy matching pseudocode:

```
for each research output ro in the new dataset:
    if ro is published outside the year range existing in the 2024 dataset:
        Mark ro as a unique record
    if ro is published within the year range existing in the 2024 dataset:
        Only consider research outputs published in the same year in the 2024 dataset
        for each research output ro2024 in the filtered 2024 dataset:
            if PI is listed in ro as researcher or author (exact matching):
                Mark this research output with the PI flag  as True
            if PI is not listed in ro as researcher or author ( fuzzy matching):
                for each researcher and author name listed in ro:
                    Calculate the similarity score between PI and the name
                get the maximum similarity score
                if the maximum similarity score >= threshold:
                    Mark PI flag as True
```

```
            if the maximum similarity score < threshold:
                Mark PI flag as False
        if the PI flag is False:
            Continue with the next research output in the 2024 dataset
        if the PI flag is True:
            Get the similarity score between the titles of ro and Ro2024
    Get the maximum similarity score for ro in all candidates in the 2024 dataset
    if the maximum similarity score >= threshold:
        Save ro's best matched record in the 2024 dataset:
            Save the index, title, and the similarity score
    if the maximum similarity score < threshold:
        Mark ro as a unique record
```

4.3 Choice of Threshold

We decide to use threshold of 90 for both PI name and title fuzzy matching. We chose this threshold to allow for matches between two titles or names that are fairly similar without significant difference. This is especially important for names since they are usually much shorter than titles; choosing a high threshold makes sure that we don't overmatch the records that are not overly similar. We also tested with a lower threshold for robustness, and found no significant difference in matches using either 80 or 85.

4.4 Uniqueness and FSRDC

Uniqueness was ensured by our fuzzy matching methodology, and since the way we query the OpenAlex API – we searched via valid dataset provided – the research outputs we found are indeed FSRDC Research Outputs. Proof of this was shown in column matched_dataset_terms, which tells us which dataset the research output is using. Therefore, we ensure that each of our research outputs is a unique and valid FSRDC research output. For web scraping data, we can tell whether a research output is a valid output through columns 'matched_dataset_terms', 'mention_acknowledgment', 'mention_restricted_data', 'mention_disclosure_review', 'mention_rdc' since all these columns tell us whether each of the criteria is met or not for a research output considered to be a valid FSRDC research output. Each of the research outputs meet at least one of our established criteria in our web scraping results.

4.5 Data Structures

In this step, we decided to use a set to store all authors listed for each research output, since membership check is much faster than using fuzzy matching to calculate the similarity score. Furthermore, to filter the unique records during the fuzzy matching, we stored the indices of all unique research outputs in a set since order does not matter and set-based membership checking is computationally efficient. This set is used to mark the unique research output in the entire results. For research outputs that have a match, we store the index, title, and the similarity score of the best-matched research output in the 2024 dataset using a dictionary. Since dictionaries store key-value pairs, this enables us to store the mapping between two entities. In this case, dictionaries allow us to save the best-matched research output in the 2024 dataset for each new research output since it builds the connection between the new research output we have and the original research output in the 2024 dataset.

4.6 Result

We have 1,661 research outputs derived from API queries and 64 research outputs from web scraping in total. The table below shows the summary of research outputs for each of the categories either derived from API queries or web scraping

Category	Number of API Retrieved Research Outputs	Number of Webscraping Research Outputs
Unique	1391	53
Non-Unique	270	11
Total	1661	64

Table 1: Summary of Matching Results

There are four files generated at this step:

- unique_research_outputs.csv: All unique research outputs identified for API retrieved data
- matched_research_outputs.csv: All non-unique research outputs identified with information of the corresponding best-matching research output in the 2024 dataset, including index, title, and the similarity score for API retrieved data
- unique_output_webscraping.csv: All unique research outputs identified for webscraping data
- matched_outputs_webscraping.csv: All non-unique research outputs identified with information of the corresponding best-matching research output in the 2024 dataset, including index, title, and the similarity score for webscraping data

5 Graph Construction and Analysis

5.1 General Network Statistics for API dataset

Table 2: Network Statistics Summary

Property	Value
Connected components	1
Network size	841 outputs with 159,054 connections
Average clustering coefficient	0.050
Detected communities	4

By Degree Centrality (most connections)

- 0.849: Migration and Dispersal of Hispanic and Asian Groups: An Analysis of the 2006-2008 Multiyear American Community Survey
- 0.837: Complex Survey Questions and the Impact of Enumeration Procedures: Census/American Community Survey Disability Questions
- 0.827: Trends in Minority Commuting Behavior: What Does the American Community Survey Tell Us?

- 0.814: Public Use Microdata Area Fragmentation: Research and Policy Implications of Polygon Discontiguity
- 0.814: Recessions, Housing Market Disruptions, and the Mobility of Workers

By Betweenness Centrality (bridge outputs)

- 0.004: Using estimates of undocumented immigrants to study the immigration-crime relationship
- 0.004: Trends in Minority Commuting Behavior: What Does the American Community Survey Tell Us?
- 0.004: The Uptick in Income Segregation: Real Trend or Random Sampling Variance
- 0.004: Complex Survey Questions and the Impact of Enumeration Procedures: Census/American Community Survey Disability Questions
- 0.003: Commuter-Adjusted Population Estimates: ACS 2006-10

5.2 General Network Analysis for API dataset

In this part of our project, we constructed and analyzed a research output network to gain deeper insights into the structure of research collaborations and thematic overlaps among outputs associated with the Federal Statistical Research Data Centers (FSRDC). This network-based analysis was built upon a carefully curated dataset of unique research outputs derived from web scraping, API integration, and metadata cleaning, as outlined in the earlier stages of the project. The aim was to uncover latent relationships among outputs based on shared metadata features—such as authors, datasets used, affiliations, topics, and geographical locations—and to identify key outputs and communities that play a central role in the FSRDC research landscape.

To begin, we used NetworkX to construct an undirected graph where each node represents a research output (uniquely identified by DOI or title) and edges denote the presence of shared attributes. Specifically, an edge between two outputs was added if they shared any authors, datasets, topics, affiliations, or locations, with edge weights reflecting the total number of shared features. This method allowed us to capture both strong and weak relationships between outputs, providing a nuanced view of how research outputs are interconnected. The graph we constructed contained 841 nodes and an impressive 159,054 edges, suggesting a dense network with substantial interconnectedness, particularly among outputs that share common authorship or research topics. We removed duplicate outputs with the same DOI or title to ensure that each node in the graph represents a unique research output, and that's why we have 841 nodes instead of 1392 number of rows in the unique research outputs dataset. This prevents artificial inflation of the network, avoids misleading connections and centrality scores, and ensures that community detection reflects true research clusters rather than data redundancies.

We then computed several network metrics to quantify the graph's structural properties and identify influential nodes. The average clustering coefficient was 0.050, indicating that although the network is highly connected, local clusters (i.e., triadic closures among outputs) are relatively sparse. This is consistent with a network where many outputs are connected via shared authors or datasets, but not necessarily forming tightly knit communities of three or more closely linked nodes.

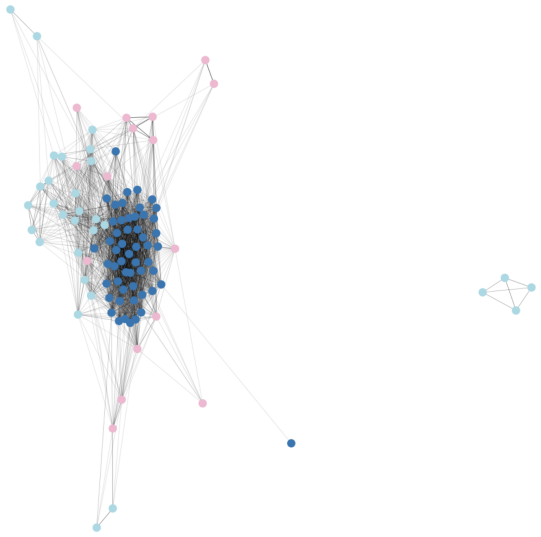


Figure 1: Research Outputs Network (Nodes colored by community)

Centrality analysis highlighted the most prominent outputs in the network. In terms of degree centrality—which captures the number of direct connections—a few outputs stand out as key hubs. For instance, "Migration and Dispersal of Hispanic and Asian Groups" and "Complex Survey Questions and the Impact of Enumeration Procedures" exhibit exceptionally high centrality scores (0.849 and 0.837, respectively), meaning they are directly connected to a large portion of the network. These outputs likely serve as foundational or integrative works that bring together multiple authors and datasets. Meanwhile, betweenness centrality—a measure of how often a node appears on the shortest paths between others—revealed that some outputs serve as critical bridges in the network. The same output on enumeration procedures also ranks highly in this metric, indicating it not only connects to many other outputs but also plays a strategic role in linking disparate clusters within the graph.

5.3 Graph Visualization and Analysis

The visualized graph shown in Figure 1 represents only a subset of the full research output network. Since the original graph contains 841 nodes—too large for clear visualization—we displayed a sample of nodes to make the structure and communities more interpretable while preserving the network's general characteristics.

To further explore the network's structure, we applied the Louvain community detection algorithm, which revealed five distinct communities within the graph. These communities, visualized in the graph image using different node colors, likely correspond to groups of research outputs that revolve around similar topics or are co-authored by recurring sets of researchers. Interestingly, one of the communities appears as a small, tightly connected cluster that is relatively isolated from the main component—possibly indicating a niche subfield or a group of collaborators who are not well integrated with the rest of the FSRDC research ecosystem.

The graph shown in Figure 2 is the result of setting max_nodes=10 in the visualize_graph function. By restricting the visualization to a small number of nodes, we are able to clearly display

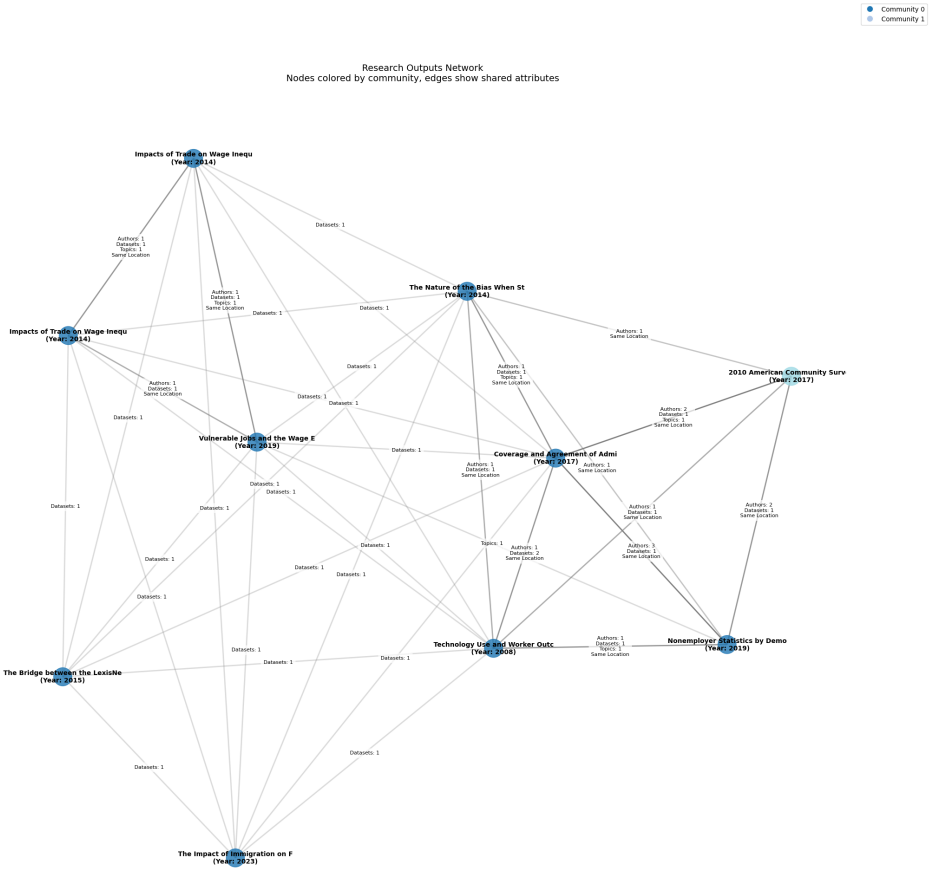


Figure 2: Research Outputs Network with 10 Nodes

each research output's title and its connections to others. The edges are labeled with the number of shared attributes—such as common authors, datasets, topics, affiliations, or location—providing an interpretable view of how these outputs are related. This kind of focused visualization is especially helpful when analyzing a specific subset of interest, allowing us to examine detailed relationships without being overwhelmed by the complexity of the full network.

5.4 Supplementary Network from Web-Scraped Dataset

To supplement our primary analysis, we constructed an additional research output network using a the dataset collected via web scraping. This dataset includes 53 unique research outputs and was processed using the same graph construction methodology as outlined earlier—linking outputs via shared metadata such as authors, topics, and datasets.

The resulting graph contains 1 connected component, with 53 nodes and 806 edges. The average clustering coefficient is 0.129, notably higher than the previous network (0.050), suggesting denser local clusters and potentially tighter thematic overlap or collaboration patterns. Community detection via the Louvain algorithm revealed 3 distinct communities.

Top Outputs by Degree Centrality (most connections):

- 0.846: The Effect of Oil News Shocks on Job Creation and Destruction
- 0.846: The Impact of Industrial Opt-Out from Utility Sponsored Energy Efficiency Programs
- 0.846: Race and Mobility in U.S. Marriage Markets: Quantifying the Role of Segregation
- 0.846: Long-Run Adult Socio-economic Outcomes from In Utero Airborne Lead Exposure
- 0.827: Revisions to the LEHD Establishment Imputation Procedure and Applications to Administrative Job Frame

Top Outputs by Betweenness Centrality (bridge outputs):

- 0.101: The Underserved Have Less Access to Employer-Sponsored Telemedicine Coverage
- 0.067: The Changing Nature of Pollution, Income, and Environmental Inequality in the United States
- 0.049: The Impact of Industrial Opt-Out from Utility Sponsored Energy Efficiency Programs
- 0.041: The Long-Term Effects of Income for At-Risk Infants: Evidence from Supplemental Security Income
- 0.038: Methodology on Creating the U.S. Linked Retail Health Clinic (LiRHC) Database

Although smaller in size, this supplementary network provides a focused view of research activity concentrated around areas such as environmental inequality, energy efficiency, and public health. The higher clustering coefficient reflects tighter intra-group connections, likely stemming from more specialized or collaborative research patterns. These findings highlight the value of integrating scraped datasets to enrich network analysis with recent and thematically cohesive research clusters.

5.5 Discrete Event Simulation

Assumptions:

- Assume editors are needed in the initial review, and each editor can only review one paper at a time, and there are 2 editors available, each research paper needs one editor for the initial review stage
- Assume the initial review time follows the exponential distribution with an average of 5 days, and there is a 30% chance that the research paper will be rejected by the editor during the initial review
- Assume peer reviewers are needed in the initial review, and each peer reviewer can only review one paper at a time, and there are 5 peer reviewers available, each research paper needs one peer reviewer for the peer review stage
- Assume the peer review time follows the exponential distribution with an average of 7 days, and there is a 30% chance that the research paper will be rejected by the peer reviewer during the peer review
- Assume the revision process needs an author to make changes to it and one peer reviewer to check for revisions for one research paper, and we have unlimited authors to make revisions (Authors can always start making revisions right away), and we still have 5 peer reviewers and each peer reviewer could only check revision/ peer review for one paper at a time
- Assume that a research paper could go through revision at most three times, and there is a 50% chance that the paper needs one round of revision independently
- Assume the author revision time follows the exponential distribution with an average of 7 days, and the peer reviewer revision checking time follows the exponential distribution with an average of 5 days

What-if Scenario: More Available Editors

Category	Average Time with 2 editors	Average Time with 3 editors
Editor Wait Time	7.57 days	0.89 days
Peer Reviewer Wait Time	2.22 days	0.28 days
Revision Peer Reviewer Wait Time	2.05 days	0.07 days
Initial Review Reject Time	11.01 days	5.46 days
Peer Review Reject Time	22.88 days	12.62 days
Published Time	31.43 days	20.68 days

Table 3: Average time comparison between different numbers of available editors

Based on Table 3, we can tell that the editors needed in the initial review are our bottleneck since that gives us the longest wait time compared to other wait times, which makes the initial review reject time, peer review reject time, and published time longer. After getting one more editor, all wait times decrease significantly, which further reduces the time needed to make a decision to reject or publish papers.

6 Analysis

We use unique datasets identified from the scraped from RePEc Census Working Papers Series and OpenAlex API to conduct three additional analyses. Firstly, we use **BERT**, a popular natural language processing model to classify research outputs into five topic categories. Secondly, we use these classifications to analyze how citation count changes as a function of time and topic classification. Lastly, we create an additional network graph by linking outputs that have both the RDC and topic in common to identify clusters of activity.

In order to perform our analysis, we identify and remove duplicates from the API and web scraped data based on whether the listed research output has the same title, publication date,output venue (e.g. journal name etc), abstract, keywords, authors, project RDC and type of publication (e.g. journal article, working paper etc.). It is important to note that there are no research outputs in our web scraped data that are duplicates of research outputs obtained from the API. Hence, we combine both the API dataset with web scraped dataset. Additionally, as the Federal Statistical Research Data Centers (FSRDC) program started in 1989, we remove any output that is published before that.[1] We also identify duplicates in the FSRDC-verified Projects 2024 , which is accessible via: **FSRDC-verified Projects** based on the title, type of publication, authors, project RDC, and output venue. Table 4 shows the size of the datasets after removing duplicates.

Dataset	No. of Research Outputs	After Removal of Duplicates	Used for Analysis
API	1391	1105	Yes
Web scraped	53	53	Yes
FSRDC-verified Projects 2024	1735	1734	No

Table 4: Details of Datasets Used

As a first step we compare the count of research outputs per year for our datasets with the cont of research outputs per year in the FSRDC-verified Projects 2024 dataset. Figure 3 plots this yearly count. In comparison to FSRDC-verified Projects 2024, we are able to collect research outputs starting from 1989 till 2025. We are able to collect more research ouputs for 2009 and 2008. For example, in 2009 we are able to collect 40 while FSRDC-verified Projects 2024 has 22. The highest count of research outputs in our dataset is 2018 with 97 outputs. Additionally we also compare the type of research outputs across our dataset and FSRDC-verified Projects 2024 dataset (Figure 4). We create 7 categories for research output type: (1) 'Journal Article Publication' if the output has been published in a journal, (2) 'Working Paper' if the output is classified as a working paper in FSRDC-verified Projects dataset or a report in our combined dataset (these reports were NBER working papers), (3) 'Graduate Research Output' if the research output is a dissertation or master's thesis, (4) 'Book' if the output is a book or book chapter, (5) 'Other Publication' if the output is a mimeo, blog, CES Technical Note, review in a journal, and letter in a journal, (6) 'Dataset' if the output is listed as a dataset, and 'Software' if the output is listed as software. Our combined dataset across the years has 59.74 percent more journal article publications than the FSRDC-verified Projects dataset while FSRDC-verified Projects dataset has 34.53 percent more working papers. This is logically consistent as the method used focused on scraping RePEc Census Working Papers Series and used OpenAlex API. Both of these resources have more published papers than other ways to identify outputs. Our dataset is also able to capture research outputs from RDCs such as Kansas City, Florida and CMU, which are not present in FSRDC-verified Projects dataset (Appendix Figure 1). There is yearly variation in research outputs across RDCs with the main

[1] We have once case- with output year is 1978

outlier being UCLA with 29 outputs in 2008 (Appendix Figure 2).

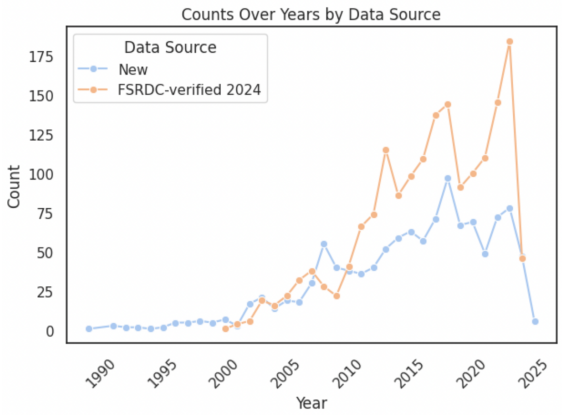


Figure 3: Research Outputs Across the Years

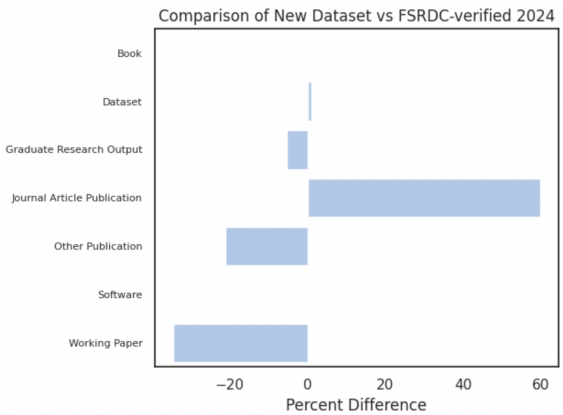


Figure 4: Difference in Output Types

Our main analyses focus on our combined dataset. In order to classify each research output into a topic, we use the abstracts for each research output. There are 21 instances where there are no abstracts in our dataset. We replace these 21 instances with the research output titles. Before we run BERT we lower the case, remove stop words and punctuation from our abstracts.[2] We classify each output into five topics. Table 5 shows the top five words and weights for each topic category. 680 research outputs are classified in 'Rural Infrastructure and Employment Data', 246 for 'Demographic Data and Household Structures', 186 in 'Growth, Firm and Productivity', 39 in 'Efficiency, Productivity and Economic Growth', and 3 in 'Human Capital, Labor Force and Productivity'. We estimate how these topics vary over time. Figure 5 shows the variation in topics over time. The y axis plots the percentage of research outputs classified in a particular topic category and x axis plots the time. The time we used is our analysis is the publication year extracted from the publication date available in our dataset.[3] There is variation in distribution of topics over time. Although since 1996, most of our research outputs are about demographic data and household structures, we see a decline in research outputs about efficiency, productivity and economic growth and uptick in growth, firm and productivity research outputs particularly in 2024. This uptick corresponds to a decline in outputs classified as being about demographic data and household structures.

We also look at how the topic classification are associated with the cite count. There is variation across topic types and cite count (Appendix Figure 3). The key outliers are papers about efficiency, productivity and economic growth published in 1995. This is also a function of the fact that almost all of our research outputs in that year are about efficiency, productivity and economic growth (See Figure 5). To estimate the relationship between cite count and topics, we create a new variable 'days since publication', which is a difference between date of publication and scraping date i.e. 30 March 2025. Since we anticipate that there will be variation in cite count across categories of topics we estimate a simple linear regression interacting the topics with days since publication. Table 6 details the model summary. The base category is 'Demographic Data and Household Structures' research outputs. The key estimates are the interaction terms that estimate the difference in effect across topic categories. We see that as time since publication increases research outputs about efficiency, productivity and economic growth have more citations than research outputs about

[2]We do not need to tokenize as BERT does it

[3]There are 5 instances where publication date was missing. It was replaced by Year column part of our dataset after verifying that it corresponded to year of publication.

Figure 5: Topics Over time

demographic data and household structures. This effect is statistically significant at 1 percent level of significance. We find a similar effect for topics about rural infrastructure and employment. It is important to note these are all associations and the model only explains 7 percent of the variation in cite count.

Word	Topic	Weight
data	Rural Infrastructure and Employment Data	0.0362
census	Rural Infrastructure and Employment Data	0.0332
us	Rural Infrastructure and Employment Data	0.0300
employment	Rural Infrastructure and Employment Data	0.02445
broadband	Rural Infrastructure and Employment Data	0.02070
data	Demographic Data and Household Structures	0.04674
census	Demographic Data and Household Structures	0.03366
survey	Demographic Data and Household Structures	0.02679
us	Demographic Data and Household Structures	0.021852
american	Demographic Data and Household Structures	0.01996
firms	Growth, Firm and Productivity	0.064775
us	Growth, Firm and Productivity	0.04240
data	Growth, Firm and Productivity	0.04109
business	Growth, Firm and Productivity	0.0385
firm	Growth, Firm and Productivity	0.03616
efficiency	Efficiency, Productivity and Economic Growth	0.08487
plants	Efficiency, Productivity and Economic Growth	0.0744
firms	Efficiency, Productivity and Economic Growth	0.06960
vc	Efficiency, Productivity and Economic Growth	0.04754
productivity	Efficiency, Productivity and Economic Growth	0.04508
human	Human Capital, Labor Force and Productivity	0.29951
capital	Human Capital, Labor Force and Productivity	0.27370
value	Human Capital, Labor Force and Productivity	0.2552
measures	Human Capital, Labor Force and Productivity	0.21746
market	Human Capital, Labor Force and Productivity	0.17133

Table 5: Top 5 Words, Weights and Topics

Table 6: OLS Regression Results

Variable	**Coef.**	**Std. Err.**	**t**	**P>|t|**	**[0.025, 0.975]**
Intercept	16.6382	35.461	0.469	0.639	[-52.939, 86.215]
T.Efficiency, Productivity and Economic Growth	-1015.9755	184.749	-5.499	0.000	[-1378.460, -653.491]
T.Growth, Firm and Productivity	65.8482	63.347	1.039	0.299	[-58.440, 190.137]
T.Human Capital, Labor Force and Productivity	452.3133	2414.283	0.187	0.851	[-4284.616, 5189.242]
T.Rural Infrastructure & Employment Data	-144.1353	76.621	-1.881	0.060	[-294.470, 6.199]
days since publication	0.0050	0.008	0.615	0.539	[-0.011, 0.021]
days since publication × T.Efficiency, Productivity and Economic Growth	0.2543	0.033	7.667	0.000	[0.189, 0.319]
days since publication × T.Growth, Firm and Productivity	-0.0034	0.014	-0.243	0.808	[-0.031, 0.024]
days since publication × T.Human Capital, Labor Force and Productivity	-0.0589	0.307	-0.191	0.848	[-0.662, 0.544]
days since publication × T.Rural Infrastructure & Employment Data	0.0608	0.018	3.379	0.001	[0.026, 0.096]

Model Summary	
R-squared	0.078
Adj. R-squared	0.071
F-statistic	10.74
Prob (F-statistic)	4.01e-16
Log-Likelihood	-8780.3
AIC	17580
BIC	17630
No. Observations	1152
Df Residuals	1142
Df Model	9
Durbin-Watson	1.784

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large (7.64e+05), which might indicate strong multicollinearity or other numerical problems.

Lastly, in order to see the relationship between research topics and research locations, we create a undirected network graph creating edges only in cases when research outputs have the same RDC **and** topic. Figure 6 displays the results. For the purpose of this analysis our dataset is converted into an object "ResearchOutput" with each instance of the class representing a row of the dataset and values corresponding to the title, year, project RDC, topics, keywords and authors. This ensures that related attributes for each research output are together making data management easier and efficient. Using ResearchOutput class also gives us the added benefit of using functionality of inheritance and polymorphism, and enables easier refactoring. By default, in Python, two objects of a class are only considered equivalent if both variables point to the exact same object in memory. This means there are instances where research outputs with the same title, year, project RDC, topics, keywords and authors will be considered as *different* objects. To overcome this, we incorporate an equivalence method within the class definition to make sure all research outputs with the same values for title, year, project RDC, topics, keywords and authors are considered equivalent. Based on this definition, each node in our graph represents a unique research output. There are 918 nodes and 13635 edges in our graph. We identify and plot the number of connected components based on this graph. Figure 6 shows the number of connected components along with the number of research outputs in each. There are 51 connected components with the largest corresponding to Boston with research outputs relating to demographic data and household structures (N=686). The other research outputs are significantly smaller and range from 27 to 1. This tells us a lot of our data has research from Boston RDC about demographic data and household structures across the years.

7 Conclusion

his project underscores how methodical data collection and rigorous text analysis can significantly improve our ability to trace the scholarly footprints of FSRDC-based studies. By developing a pipeline that merges web scraping, API retrieval, and advanced pattern matching, we compiled 1,444 newly confirmed research outputs, thereby extending existing datasets and offering a richer foundation for future investigations. Our network analysis, in particular, reveals clusters of high-impact works and highlights how a handful of "bridge" publications shape the connectivity of the broader research ecosystem. In addition, our regression results emphasize the relevance of publication timing and topical focus in determining citation patterns—though many nuances, such as journal prestige and co-author networks, remain to be fully explored.

Despite these advances, our findings also suggest that a substantial volume of yet-undiscovered FSRDC research remains hidden across other repositories and databases. As we have demonstrated, venturing beyond one or two primary sources (e.g., RePEc, OpenAlex) can uncover dozens of new publications that meet official FSRDC criteria. A logical extension of this work involves integrating other large-scale scholarly APIs, such as CORE or BASE, or targeting specialized subject-matter databases that might capture narrower fields of inquiry. Continued expansion of the search universe would not only yield more comprehensive insights into the FSRDC landscape, but also further refine our understanding of how restricted data contributes to advancing empirical research.

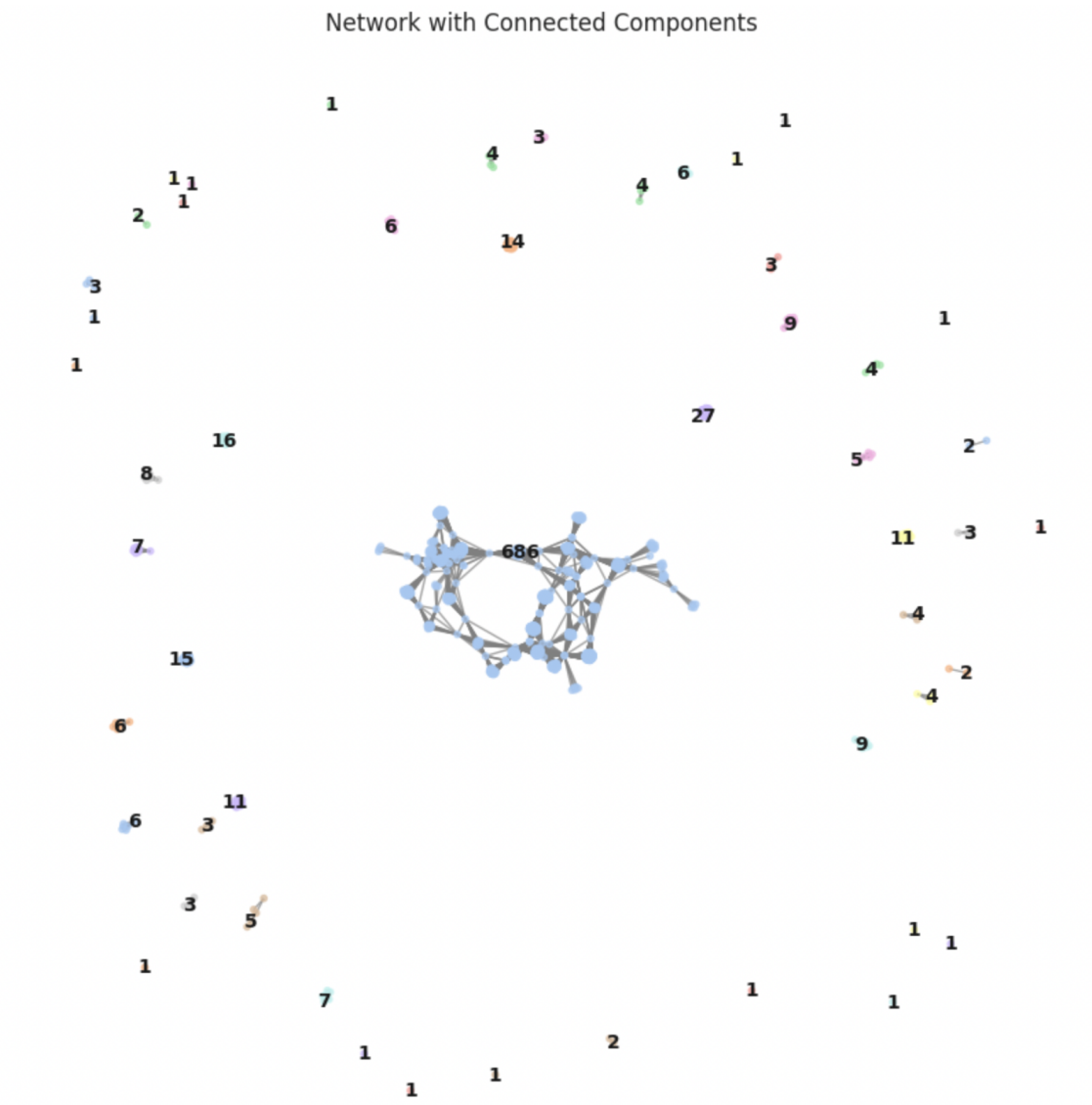


Figure 6: This figure shows that many of our papers are highly connected, with the largest connected component featuring 686 papers. Nevertheless, there are many small clusters of papers, ranging from 1 to 27 that exist outside of this largest cluster.

8 Appendix

8.1 Running the Project

The project can be run via the *main.py* file. This file will execute all python subfiles for the project except for Part 5. **Warning:** Part 5 must be executed on Google Colab. Upload the file *visualization.ipynb* along with the data files *unique_outputs_webscraping.csv* and *unique_research_outputs.csv* to run Part 5. **Warning:** the file *api_integration.py* will only execute if the final two lines in the file are uncommented. They are currently commented out because execution takes roughly 10 hours to run.

8.2 Data Construction

Github Scraping Pseudocode

```
# Pseudo code for "github_scraper.py"

# 1. Import necessary packages
import pandas as pd

# 2. Load Excel files from a GitHub URL (ProjectsAllMetadata.xlsx)
try:
    all_metadata = read_excel_from_github()
    abstracts = read_excel_sheet("Abstracts")
    datasets = read_excel_sheet("Datasets")
    researchers = read_excel_sheet("Researchers")
except FileNotFoundError:
    print_error_message()
except OtherException:
    handle_exception()

# 3. Merge datasets and researchers
merged_data = merge_on_title_and_PI(datasets, researchers)

# 4. Rename columns for consistency
rename_columns(merged_data)

# 5. Ensure PIs appear as researchers
unique_combinations = find_unique_rows(merged_data, ["title", "location", "dataset", "pi"])
new_rows_for_missing_PIs = []
for combo in unique_combinations:
    filtered_subset = filter_rows(merged_data, combo)
    if PI_not_in_researchers(filtered_subset):
        create_new_row_for_PI()

append_new_rows(merged_data, new_rows_for_missing_PIs)

# 6. Generate extra dataset terms (e.g., "FSRDC", "Census Bureau")
unique_combinations_again = find_unique_rows(
```

```
    merged_data, ["title", "location", "pi", "researcher"])
placeholder_rows = []
for combo in unique_combinations_again:
    build_row_with_location_RDC()
    build_row_for_FSRDC()
    build_row_for_Census_Bureau()

final_dataset = append_new_rows(merged_data, placeholder_rows)

# 7. Write final results to CSV
save_to_csv(final_dataset, "dataset_data.csv")
```

OpenAlex API Pseudocode

```
# Pseudo code for "abstract_search.py"

# 1. Import packages
import pandas, requests, time

# 2. Define a helper to reconstruct abstract from inverted index
def reconstruct_abstract(inverted_index):
    # Find max index, create placeholder list of words
    # Fill positions with corresponding words
    # Return joined string

# 3. Define a function to fetch data by DOI
def fetch_openalex_data_by_doi(doi):
    # Clean DOI and call the OpenAlex API
    # Parse JSON and return first result if any

# 4. Define a function to fetch data by title
def fetch_openalex_data_by_title(title):
    # Send request to OpenAlex, handle errors
    # Return first matching record or None

# 5. Process returned OpenAlex object for abstract and keywords
def process_openalex_data(work):
    # Reconstruct abstract from 'abstract_inverted_index'
    # Extract concepts to form comma-separated keywords

# 6. Attempt to get abstract & keywords using DOI first, then title
def get_abstract_and_keywords(row):
    # If valid DOI, fetch data
    # Fallback to title if DOI fails
    # Return abstract, keywords or "No work found"

# 7. Read CSV, iterate rows, fetch abstract/keywords, save results
def process_csv(input_csv, output_csv):
```

```
    # Load file
    # For each row, call get_abstract_and_keywords
    # Store results in new columns
    # Write updated CSV

if __name__ == "__main__":
    process_csv("cleaned_biblio.csv", "cleaned_abstracts.csv")
```

8.3 Schema For Part 3 and Part5

Table Schema of Part 3 matching results files

- unique_research_outputs.csv: Same schema as part 2: API retrieved research outputs file
- matched_research_outputs.csv: Same schema as part 2: API retrieved research outputs file with three additional columns:
  - best_match_idx: The index of the best matched research output in the 2024 dataset for the record we retrieved from API queries
  - best_match_title: The title of the best matched research output in the 2024 dataset for the record we retrieved from API queries
  - best_match_score: The similarity score of the best matched research output (in terms of titles) in the 2024 dataset for the record we retrieved from API queries
- unique_outputs_webscraping.csv: Same schema as part 1: Web scraped research outputs file
- matched_outputs_webscraping.csv: Same schema as part 1: Web scraped research outputs file with three additional columns:
  - best_match_idx: The index of the best matched research output in the 2024 dataset for the record we retrieved from web scraping
  - best_match_title: The title of the best matched research output in the 2024 dataset for the record we retrieved from web scraping
  - best_match_score: The similarity score of the best matched research output (in terms of titles) in the 2024 dataset for the record we retrieved from web scraping

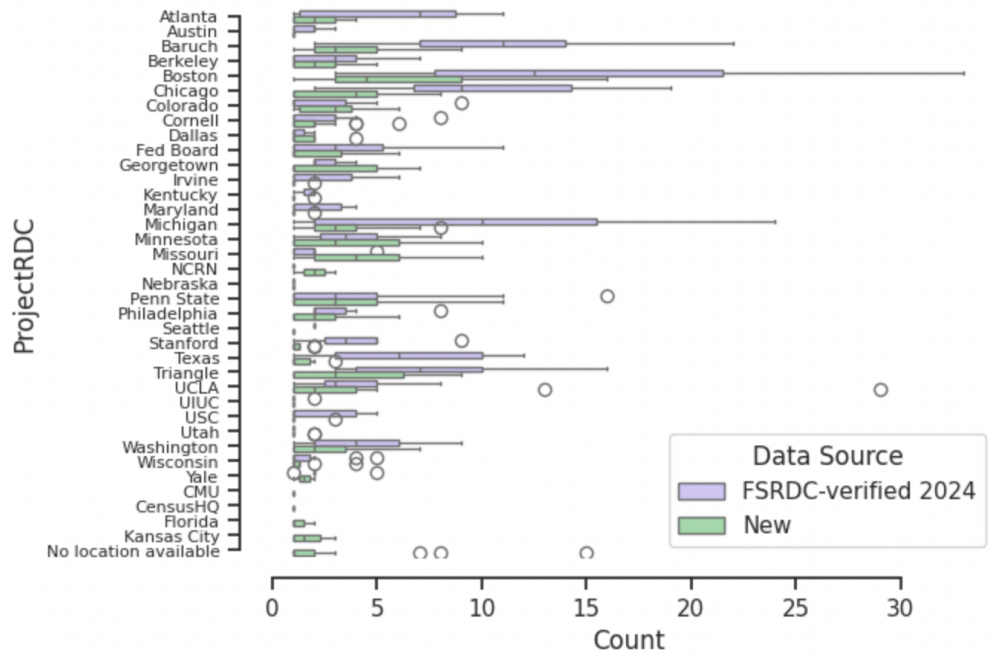
Table Schema of Analysis

- **Datasets Used:**
  - unique_research_outputs.csv
  - unique_outputs_webscraping.csv
  - FRSDC-verified 2024
- **Subset Dataset created:** For unique_research_outputs.csv and unique_outputs_webscraping.csv we only keep selected columns and rename them as follows:
  - ”title”: title of publication **Renamed to:OutputTitle**
  - ”year”: scraped year (can be different from publication year)
  - ”publication_date” : publication date **Renamed to:PublicationDate**

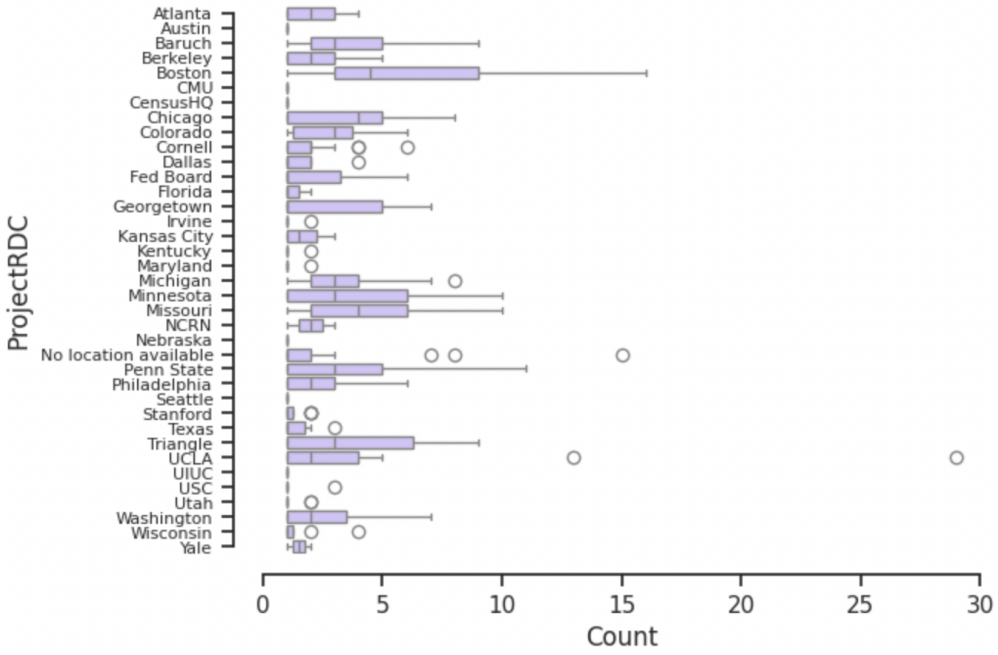
- ”authors”: set of authors **Renamed to:ProjectPI**
- ”abstract”: abstract if given **Renamed to:Abstract**
- ”topics”: keywords **Renamed to:Keywords**
- ”type_crossref”: type of output i.e. journal atricle, book etc
- ”location”: RDC **Renamed to:ProjectRDC**
- ”cited_by_count”: citation count **Renamed to:CiteCount**
- We split the set of authors and store each author name in a column : Author_1 .... Author_48 (The largest number of authors is 48)
- We create OutputYear by extracting year from PublicationDate
- PublicationDate_clean: PublicationDate stored as date time object
- OutputType_Clean: If output type is: 'article': 'Journal Article Publication', 'preprint': 'Journal Article Publication', 'report': 'Working Paper', 'dataset': 'Dataset', 'book-chapter': 'Book', 'book': 'Book', 'dissertation': 'Graduate Research Output', 'other': 'Other Publication', 'paratext': 'Other Publication', 'review': 'Other Publication', 'editorial': 'Other Publication', 'letter': 'Other Publication'.
- Data Source: Given value ”New” for all outputs we scraped

- **For descriptive:** Concatenated dataset used with only”ProjectRDC”, ”OutputTitle”,”OutputType_Clean”,” Source” being kept.
  - For the FRSDC-Verified dataset OutputType_Clean is created using the following: 'JA': 'Journal Article Publication', 'SW': 'Software', 'WP': 'Working Paper', 'DS': 'Dataset', 'BC': 'Book', 'DI': 'Graduate Research Output', 'MT': 'Graduate Research Output', 'RE': 'Other Publication', 'BG': 'Other Publication', 'MI': 'Other Publication', 'TN': 'Other Publication'
  - **For Main Analysis:** Only combined ‘New’ dataset with subset columns used

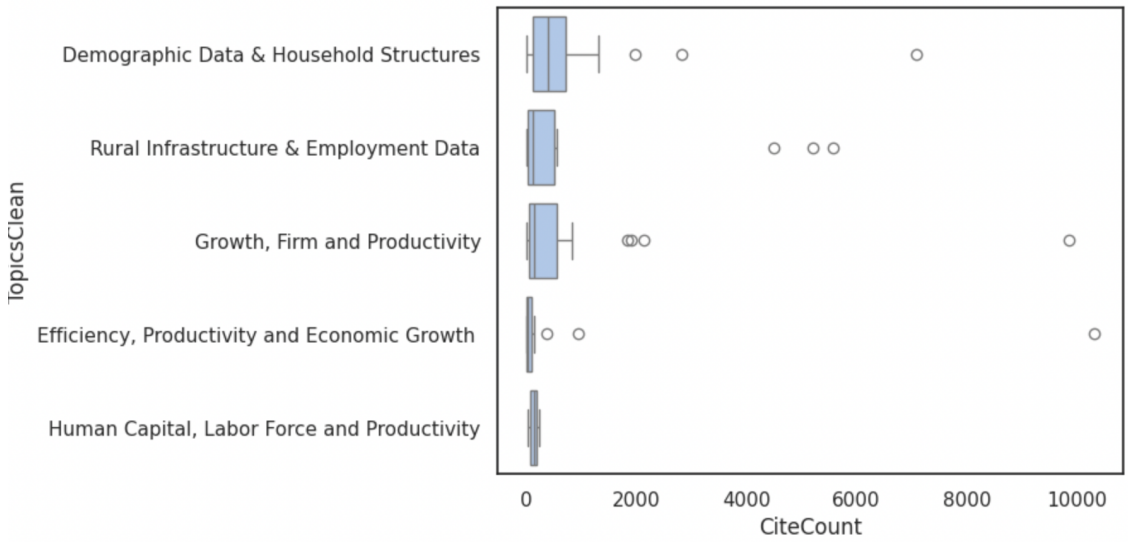
8.4 Analysis



Appendix Figure 1: Distribution of Research Outputs



Appendix Figure 2: Distribution of Research Outputs: New Dataset



Appendix Figure 3: Topics and Cite Counts