Long-Term University Research Rankings Based on Publications in the Top Five Economics Journals

Prepared by:
Mathew Johnson
JHNMAT015
Department of Economics
University of Cape Town

Prepared for:
Prof. Co-Pierre Georg
Department of Economics
University of Cape Town

December 2024

Plagiarism Declaration

COMPULSORY DECLARATION:

1. This dissertation has been submitted to Turnitin (or equivalent similarity and originality checking software) and I confirm that my supervisor has seen my report, and any concerns revealed by such have been resolved with my supervisor.
2. I certify that I have received Ethics approval (if applicable) from the Commerce Ethics Committee.
3. This work has not been previously submitted in whole, or in part, for the award of any degree in this or any other university. It is my own work. Each significant contribution to, and quotation in, this dissertation from the work, or works of other people has been attributed, and has been cited and referenced.

Student number	JHNMAT015
Student name	Mathew Lee Johnson
Signature of Student	Signed by candidate
Date:	5 December 2024

Abstract

Commercial university ranking systems have existed since the early 2000s, with multiple studies having been conducted using various metrics to rank universities over time. However, these studies are usually performed on merely a decade's worth of data at a time. This dissertation conducts long-term university rankings from 1940 to 2010, utilising the information provided from the top five economics journals identified in a publication by Combes and Linnemer (2010), and demonstrates how these metrics have evolved.

A database of papers is created to perform this long-term ranking by downloading all the publications in the top five economics journals using a web scraper called Aaron's Kit (Chembezi, Oldnall & Van Zyl, 2021). The information from the papers is extracted from the PDFs using optical character recognition. Finally, the data is cleaned using fuzzy string matching and string replacements, with the remaining data points being filled in using Scopus and Amazon Mechanical Turk.

To produce a ranking, a weighted count of authors and affiliations is applied. These rankings are presented over time to observe changes from 1940 to 2010. The rankings are then compared to similar studies that use more complex ranking methodologies, with the results yielded being very similar. This suggests that a more complex solution may not produce better ranking results. The top-ranked universities from this study are Harvard University, the University of Chicago, and the Massachusetts Institute of Technology (MIT), in that order. It is found that author collaboration on papers has increased from an average of 1.14 authors per paper in the 1940s to 1.98 authors per paper in the 2000s.

Acknowledgements

To my supervisor, Dr. Co-Pierre Georg, thank you for continually pushing and supporting me. Without your weekly guidance, this paper would not have reached its endpoint.

To my three Fintech classmates, Si Jia Wu, Francisco Rutayebesibwa, and Sandisiwe Makhathini, who built the database with me, this project would have been impossible without your collaboration. Thank you for seeing this through to the end with me.

To my friend, William Mufford, thank you for always being there for me to bounce ideas off. More importantly, your assistance in manually checking affiliations with me was invaluable. Your advice and assistance helped get me through very difficult phases of this paper.

To my parents, Anthony and Gail Johnson and partner Olivia Bloomer, thank you for always believing in me. I could not have achieved this without your support and motivation.

Table of Contents

Table of Figures

Table of Tables

List of abbreviations

Abbreviation	Definition
AER	American Economic Review
ARWU	Academic Ranking of World Universities
CEPR	Centre for Economic Policy Research
CL-index	Combes–Linnemer Index
DOI	Digital Object Identifier
ECTA	Econometrica
H5-index	Google Scholar five-year h-index
JCR	Journal Citation Reports
JEL	Journal of Economic Literature
JIF	Journal Impact Factor
JPE	Journal of Political Economy
JSTOR	Journal Storage (digital archive)
KMS	Kalaitzidakis, Mamuneas & Stengos methodology
LD	Levenshtein Distance
MTurk	Amazon Mechanical Turk
NBER	National Bureau of Economic Research
OCR	Optical Character Recognition
PDF	Portable Document Format
QJE	Quarterly Journal of Economics
QS	Quacquarelli Symonds
RES	Review of Economic Studies
SSCI	Social Science Citation Index
THE	Times Higher Education
URL	Uniform Resource Locator

1 Introduction

University ranking systems use various metrics to compare universities and decide which one is best accordingly. Before the emergence of modern global ranking systems, earlier national and regional rankings, such as the U.S. News & World Report rankings introduced in 1983, played a prominent role in comparing universities within individual countries. Contemporary commercial international ranking systems began to appear in the early 2000s, most notably with the launch of the Academic Ranking of World Universities (ARWU) in 2003, followed by the QS World University Rankings in 2004, and the independent Times Higher Education (THE) rankings in 2010. According to the Council on Higher Education's most recent review, there are now more than 47 active global and regional university ranking systems worldwide, reflecting the rapid expansion and commercialisation of ranking practices (Council on Higher Education, 2024). Various stakeholders make use of university rankings to make decisions concerning the universities. Students use them to decide which university to attend, academics use them to decide where to teach and research, and outside investors utilise them to affiliate with the university viewed as the best. Therefore, it is in the best interests of universities to rank highly. A fall in rankings can lead to a decline in a university's reputation, which in turn can decrease the number of applications from students and academics.

There are very few studies comparing the long-term rankings of universities. A major impediment to performing these rankings is a lack of reliable historical data. While databases like Elsevier Scopus (Scopus) (Elsevier, 2024) exist to provide university publication data, their coverage for periods before the widespread adoption of digitisation remains limited. Despite recent improvements in accuracy and completeness, Scopus' historical data remains insufficient. This limitation directly impacts the reliability of rankings produced by leading publishers such as Times Higher Education (*THE*), which heavily rely on such databases.

A study by Combes and Linnemer (2010) ranked the top economic journals by utilising citation data from the Journal of Citation Reports (*JCR*). Their ranking involved a combination of the *JCR*, the *Journal of Economic Literature* (*JEL*) classification system and Google Scholar's h5-index, as detailed in section 2.

The list below is the final ranking from Combes and Linnemer (2010). The top five from their paper will be used in this paper and known as the "Top Five Economics Journals".

1. *Quarterly Journal of Economics* (*QJE*)
2. *American Economic Review* (*AER*)
3. *Journal of Political Economy* (*JPE*)
4. *Econometrica* (*ECTA*)
5. *Review Of Economic Studies* (*RES*)

This paper aims to construct a long-term ranking of the research intensity of universities globally based on academic publications. This analysis will focus on publications in the above-mentioned Top Five Economics Journals from 1940 to 2010.

Achieving this requires building a clean database of author and affiliation information, and from this, long-term university rankings are created. This process is structured around the following three objectives.

Thesis Objectives

1. Compile a database of papers published in the Top Five Economics Journals from 1940 to 2010.
2. Build a database of author and affiliation information linked to these publications.
3. Use these databases to produce:
    a. A long-term ranking of universities based on their publication output.
    b. Long-term collaboration metrics to analyse inter-university partnerships.

Thesis Outline

This paper begins with a literature review in Chapter 2, which unpacks previous university rankings based on publications, as well as commercial ranking systems and their methodologies, including their 2023 rankings. This chapter identifies the key flaws that exist in these ranking systems and explores where they source their data. Chapter 3 explains how the database of papers is built for this dissertation using a Python web scraper to collect data from the Top Five Economic Journals. Chapter 4 addresses the processing and cleaning of the collected data. It first explores the structure of the journals and explains how optical character recognition (OCR) is used to extract data from PDFs. The chapter then outlines two primary methodologies – data consolidation and fuzzy matching - used to clean the OCR output data such that it forms a useable database. Additionally, any missing data is sourced and merged to complete the database of authors and affiliations. Chapter 5 presents the results of research ranking over time, providing long-term university rankings based on publication output and illustrating trends in research productivity. Chapter 6 offers a comparative analysis of the findings in this study with other published studies on university rankings. Chapter 7 concludes the paper. Finally, Chapter 0 presents recommendations for future work, suggesting potential extensions to or improvement of the research.

2 Literature Review

This section begins by analysing the economic journals used in this paper and their impact within the field. It then goes on to explain existing university ranking systems and their methodologies. Finally, it examines the sources of data underpinning these ranking systems, highlighting their potential limits and applications.

2.1 Academic Journals

2.1.1 Journal Rankings

An academic journal is a periodical publication of academic work related to a certain field. Most academic journal publications are peer-reviewed to ensure their reliability. Generally, these journals are accessible through paywalls or institutional logins. JSTOR[1] is an example of a digital library of academic publications and journals. Many institutions, including the University of Cape Town, give students access to JSTOR.

Bush, Hamelman and Staaf (1974) created an early ranking of journals by counting the number of citations in fourteen journals from 1966 to 1970 excluding self-citations to produce a ranking. Liebowitz and Palmer (1984) built upon this idea to produce their own ranking a decade later. Their approach implemented three changes: (1) standardising the journals to compensate for size and age; (2) including more journals; (3) applying an iterative process to “impact-adjust” the number of citations received from journals. Laband and Piette (1994) further refined this work by identifying the list of journals to be included, relying on the Social Science Citation Index (SSCI) as their source.

Journal Citation Reports

Clarivate Analytics has published Journal Citation Reports (*JCR*) since 1975. The *JCR* provides a quantitative measure for ranking and comparing journals. This is based on the Journal Impact Factor (JIF), which measures the frequency with which the average article in a journal has been cited within a particular period (Mfengu, 2024). The SSCI is also published by Clarivate Analytics as a subset of the *JCR*. Table 1 provides the 2023 JIF values which have been pulled from Web of Science, a subsidiary of Clarivate (Web of Science, 2024).

Equation 1 shows the formula used by Clarivate Analytics to calculate the *JCR* (Clarivate Analytics, 1994). “Period” refers to the years over which the *JCR* is being calculated and can be accordingly adjusted to get the *JCR* for different lengths of time. Equation 1 can be adjusted by subtracting self-citations from B to calculate a revised impact factor.

Equation 1: *JCR* Formula as per Clarivate

A = total cites in period + 1 year
B = period + 1 year cites to articles published in period
C = number of articles published in period
D = B/C = impact factor for period
*Period refers to 1 or more years that the *JCR* is being calculated for

H-Index

Jorge Hirsch developed the h-index as a way to evaluate the productivity and impact of an academic. The h-index is equivalent to the largest number of publications cited at least as many times as the total number of such publications. For instance, an h-index of 20 means that an author has authored at least 20 papers that have each been cited at least 20 times (Shah & Jawaid, 2023). Google Scholar publishes journal rankings based on the h-index calculated over the past five complete years, called the h5-index (Google Scholar, 2024).

[1] *JSTOR (Journal Storage). Available at: https://www.jstor.org/*

The h-index indicates the number of well-cited papers and is not inflated by a small number of highly cited papers, which can be the case with the *JCR*. Perry and Reny (2016) do, however, outline an issue with the h-index due to certain academic fields having higher citation rates than others. For example, macroeconomists are cited 1.8 times more frequently per paper than industrial economists. Ellison (2013) initially became aware of this by analysing the variation in citation indexes across different fields and finding large disparities. While this poses a challenge when attempting to compare rankings across academic fields, it is less of a concern for this study which focuses solely on economics journals.

JEL Codes

The *JEL* classification system, developed for use within the *Journal of Economic Literature*, serves as a standard method for categorising scholarly papers within the field of economics (American Economic Association, 2024). Combes and Linnemer (2010) use *JEL* codes to normalise the contribution of each academic field. To do this, the number of citations is distributed among the *JEL* codes. For instance, if an article has three *JEL* codes and six citations, each *JEL* code is allocated two citations.

CL-Index

Combes and Linnemer (2010) developed a methodology called the CL-index which combines the *JCR*, H5-index and the *JEL* codes to minimise the individual methodologies' weaknesses. Equation 2 shows the formula applied to calculate the CL-index. A higher weight of 0.5 is applied to the *JCR* methodology as this is the traditional ranking methodology, with an even weighting of 0.25 applied to the h5-index and the $JCR_{\_Norm\_econ}$. The $JCR_{\_Norm\_econ}$ is the normalised summary index calculated based on an average score of a journal normalised by *JEL* codes. This addresses the issue initially identified by Perry and Reny (2016), where certain academic fields inherently have higher citation rates than others.

Equation 2: Calculation of the CL-Index by Combes and Linnemer

$$CL - index = 0.5 \times JCR_{\_econ} + 0.25 \times H5_{\_econ} + 0.25 \times JCR_{\_Norm\_econ}$$

Table 1 provides a summary of the top five journals ranked according to the CL-index. This paper includes data up until 2010, thus the CL-index for 2010 reflects the most recent data available in this study. The JIF for 2023 mirrors the CL-index from 2010. The editorial boards of the *AER*, *JPE*, *RES*, and *ECTA* comprises of members from multiple institutions worldwide, whereas the *QJE*'s board of editors consists solely of members from Harvard University.

Table 1: Comparison of Key Characteristics of the Top Five Economics Journals.
The table summarises publisher information, issue frequency, founding dates, and citation metrics (h5 index, JIF, CL-Index), providing context for differences in journal prominence.

Journal	Publisher	Issues per Year	Start Date	H5-index (2023)	JIF (2023)	CL-Index (2010)
***Quarterly Journal of Economics* (*QJE*)**	Oxford Academic	4	1889	108	11.1	1
***American Economic Review* (*AER*)**	American Economic Association	12	1911	153	10.5	2

Journal of Political Economy (*JPE*)	University of Chicago Press Journals	12 (2020 - present) 6 (pre-2020)	1892	81	6.9	3
***Econometrica* (*ECTA*)**	The Econometric Society	6	1933	91	6.6	4
***Review of Economic Studies* (*RESTUD*)**	Oxford Academic	6	1933	87	5.9	5

2.1.2 *Journal editors*

To fully understand how publication patterns influence long-term university rankings, it is essential to consider the role of the board of editors. Editorial boards acts as gatekeepers by determining which papers are accepted.

Table 2 gives a summary of the lead editors for each of the Top Five Economics Journals, and their institutional affiliations. JPE and QJE both have strong ties to singular institutions, while AER, RES and ECTA maintain broader and more internationally diverse boards. RES, founded at the London School of Economics, has historically been shaped by strong European and UK academic influences. AER has a wide range of university ties on its greater board of editors, particularly drawn from top U.S. universities. Current board members are from Dartmouth College, Harvard University, University of California Berkley, University of California Los Angeles and London School of Economics. ECTA also has a wide range of universities on the board these also consist mainly of top U.S. universities.

Table 2: Key Editors and Institutional Affiliations for the Top Five Economics Journals

Journal	**Editors**	**Institutional affiliations**
***Quarterly Journal of Economics* (*QJE*)**	Robert J. Barro, Lawrence F. Katz, Nathan Nunn, Andrei Shleifer, Stefanie Stantcheva	Harvard University (Oxford University Press, 2024)
***American Economic Review* (*AER*)**	Erzo F. P. Luttmer	Dartmouth College (American Economic Association, 2024)
Journal of Political Economy (*JPE*)	Esteban Rossi-Hansberg	University of Chicago (University of Chicago Press, 2024)
***Econometrica* (*ECTA*)**	Marina Halac	Yale University (Econometric Society, 2024)
***Review of Economic Studies* (*RESTUD*)**	Ruben Enikolopov	Universitat Pompeu Fabra (Review of Economic Studies, 2024)

2.1.3 *Critiques of Top Five Dominance in Economics*

A growing body of research has questioned the influence of the Top Five Economics Journals on publication practices, incentive structures, and evaluation mechanisms within the discipline. Both Heckman and Moktan (2020) and Serrano (2022) highlight concerns regarding the profession's heavy reliance on these journals as the primary indicators of research quality and academic excellence.

Heckman and Moktan (2020) describe this phenomenon as the “Tyranny of the Top Five,” demonstrating that publication in these journals plays a disproportionately influential role in determining tenure and promotion outcomes for junior economists. Using detailed career histories of tenure-track faculty hired by the top 35 U.S. economics departments between 1996 and 2010, they find that publishing in the Top Five Economics Journals is far more strongly associated with receiving tenure than publishing in other reputable outlets. Their analysis shows that this relationship persists even when holding constant factors such as citations and overall research output. They argue that this system distorts researcher incentives by encouraging safe, incremental work and by favouring scholars embedded within established editorial and institutional networks.

Serrano (2022) offers a complementary critique, arguing that the discipline’s fixation on the Top Five Economics Journals has narrowed economics’ intellectual diversity and constrained methodological innovation, a dynamic he refers to as “Top5itis.” According to Serrano, editorial concentration, the disproportionate power of referees, and structural entry barriers reinforce existing hierarchies and create an environment where publication success depends not only on scholarly merit but also on network effects and institutional path dependence. His analysis suggests that the dominance of the Five Economics Journals is not simply a reflection of research quality, but rather the outcome of long-standing institutional structures that shape what is considered valuable knowledge in the profession.

2.2 University Ranking

A wide range of methodologies has been developed to rank universities, reflecting differences in the criteria considered important for evaluating institutional performance. Surveys of ranking approaches highlight that systems may rely on bibliometric indicators, reputation scores, faculty credentials, student outcomes, or composite indices combining multiple measures (Coupé, 2003). Weighting can also be applied to these inputs based on factors such as the journal of publication or the number of authors or affiliations associated with the writing of the paper. This paper focuses on rankings based on affiliations and the number of authors.

This section investigates various publication-based ranking systems and explores some of the contemporary commercial university ranking systems, shedding light on how such rankings are calculated.

2.2.1 Publication-Based Rankings

Several methodologies have been developed to rank institutions based on their research output.

Kalaitzidakis, Mamuneas and Stengos (2003) developed a publication-based methodology (hereafter referred to as the ‘KMS methodology’) to rank institutions based on the number of publications and pages in the top thirty economic journals from 1995 to 1999. Article pages are allocated based on the author’s affiliation at the time of publication, reflecting research output rather than the author’s current affiliation. For papers with multiple coauthors, *x,* the page count is divided equally among them, with a resultant total of $\frac{1}{x}$ pages of the article assigned to each coauthor. For an author with multiple affiliations, *y,* the pages are divided among those affiliations, resulting in the affiliation being allocation $\frac{1}{y}$ of the pages corresponding to that specific author. The KMS methodology excludes research centres - such as the National Bureau of Economic Research (NBER) in the United States and the Center for Economic and Policy Research (CEPR) in the United Kingdom - as these organisations function as an umbrella entity for researchers rather than serving as their permanent institutional base.

Coupé (2003) provides an overview of fourteen ranking methodologies. The citation-based methodologies are omitted as they are not relevant to this section, leaving eleven different publication-based ranking methodologies. These rankings cover the years 1990 – 2000, with data sourced from the *EconLit* database. Of these eleven, Coupé focuses on four:

1. The KMS methodology, using page counts weighted for quality using Laband-Piette ranking of journals from the top ten journals
2. Page count weighted for quality using Laband-Piette ranking of journals from the top 71 journals.
3. Impact factor ranking, using article counts and including 258 journals.
4. Hirsch, Austin, Brooks and Moore's (HABM) ranking uses a page count weighted for page size and includes 24 journals.

From this ranking by Coupé, the most suitable methodology for this study is the first, as it includes the top ten journals, limiting the set closer to the top five journals of this study.

Dusansky and Vernon (1998) built on previous publication-based rankings to develop their own ranking methodology. First, each paper is normalised to an American Economic Review (*AER*) equivalent paper, to account for variations in character and page sizes across journals. Next, the Laband-Piette weighting is applied to each journal, as detailed in Table 3. Finally, the KMS methodology is applied to produce a final value for each paper, generating the rankings. Dusansky and Vernon applied this approach to eight journals covering the years 1990 and 1994.

Table 3: Laband-Piette Weight Applied to Each Journal for the Dusansky-Vernon Methodology

Journal	**Laband-Piette Weight**
American Economic Review **(*AER*)**	1.00
Econometrica (ECTA)	0.51
Journal of Political Economy (JPE)	0.36
Review of Economic Studies **(*RESTUD*)**	0.38
Quarterly Journal of Economics **(*QJE*)**	0.28
Journal of Economic Theory	0.23
Review of Economics and Statistics **(*RES*)**	0.24
International Economic Review **(*IER*)**	0.09

Beyond the top economics journals, Mulligan (2022) conducted a study ranking Austrian universities between 2011 and 2020. To rank the universities, a weight was applied to authors and affiliations which distributed credit proportionally based on the number of contributors and their affiliations in each paper (Mulligan, 2022). While this approach is similar to the KMS methodology, it does not factor in page count, focusing solely on the contributions of authors and affiliations.

Table 4 presents a summary of the four university ranking methodologies, detailing the journals included, the periods covered and the authors behind each approach. The KMS, Coupé and Dusansky-Vernon methodologies are all based on top economic journals, though each utilises a different subset of journals. Coupé and Dusansky-Vernon have both built their study on the KMS methodology, incorporating the Laband-Piette index to rank the journals first. Dusansky-Vernon employs the most complex approach, normalising page sizes to *AER* equivalents before applying the ranking system. Notably, all three of these studies include the Top Five Economics Journals investigated in this paper. Mulligan has conducted the most recent study with the simplest methodology. However, it does not include the top economics journals. All four of these methodologies are restricted to studies covering a single decade, emphasising the need for this paper's long-term analysis.

Table 4: Summary of Four Major Publication-Based University Ranking Methodologies.
The table compares Mulligan, KMS, Coupé, and Dusansky-Vernon based on data years, journals used, and methodological approaches.

Author, date	Methodology	Journals	Years
(Mulligan, 2022)	Weight of affiliations and authors to papers	Austrian journals	2011 - 2020
(Kalaitzidakis, Mamuneas and Stengos, 2003)	Weight of affiliations and authors to pages	Top 30 economics journals	1995 - 1999
(Coupé, 2003)	Laband-Piette index used to weight journals, then apply KMS	Top 10 economics journals	1990 - 1999
(Dusansky and Vernon, 1998)	Normalise page characters and page size to an *AER* equivalent journal, then apply Laband-Piette to weight journals, then apply KMS	Top 8 economics journals	1990 - 1994

2.2.2 *Commercial University Ranking Systems*

The foremost commercial university ranking systems have existed since the early 2000s. Before their emergence, smaller, regional rankings were done, such as the *U.S. News & World Report* ranking, which launched in 1983. Currently, there are more than 47 university ranking systems globally (Council on Higher Education, 2024). Among the most well-known are the *Times Higher Education* (*THE*), the Quacquarelli Symonds (*QS*) *World University Rankings,* and the *Academic Ranking of World Universities* (*ARWU*). Each ranking system employs distinct methodologies and metrics to calculate its ranking. As a result, the outcomes often differ, with some ranking systems favouring certain universities.

2.2.2.1 The Times Higher Education (THE)

The *THE* ranking was first published in 2010. Figure 1 illustrates the weightings applied to the 17 performance metrics used, which are grouped into five categories.

This ranking is primarily based on three main categories: Research, teaching, and citations. Research evaluates the quantity and quality of research output, teaching measures the learning experience and quality of education provided at the university, and citations assess the influence of the research produced by tracking how often a university's paper is by other institutions (Times Higher Education (THE), 2024).

Among the performance metrics, teaching reputation (15%) and research reputation (18%) are derived from surveys of staff and students, which together account for 33% of the overall ranking. This significant portion introduces a degree of subjectivity into the ranking process.

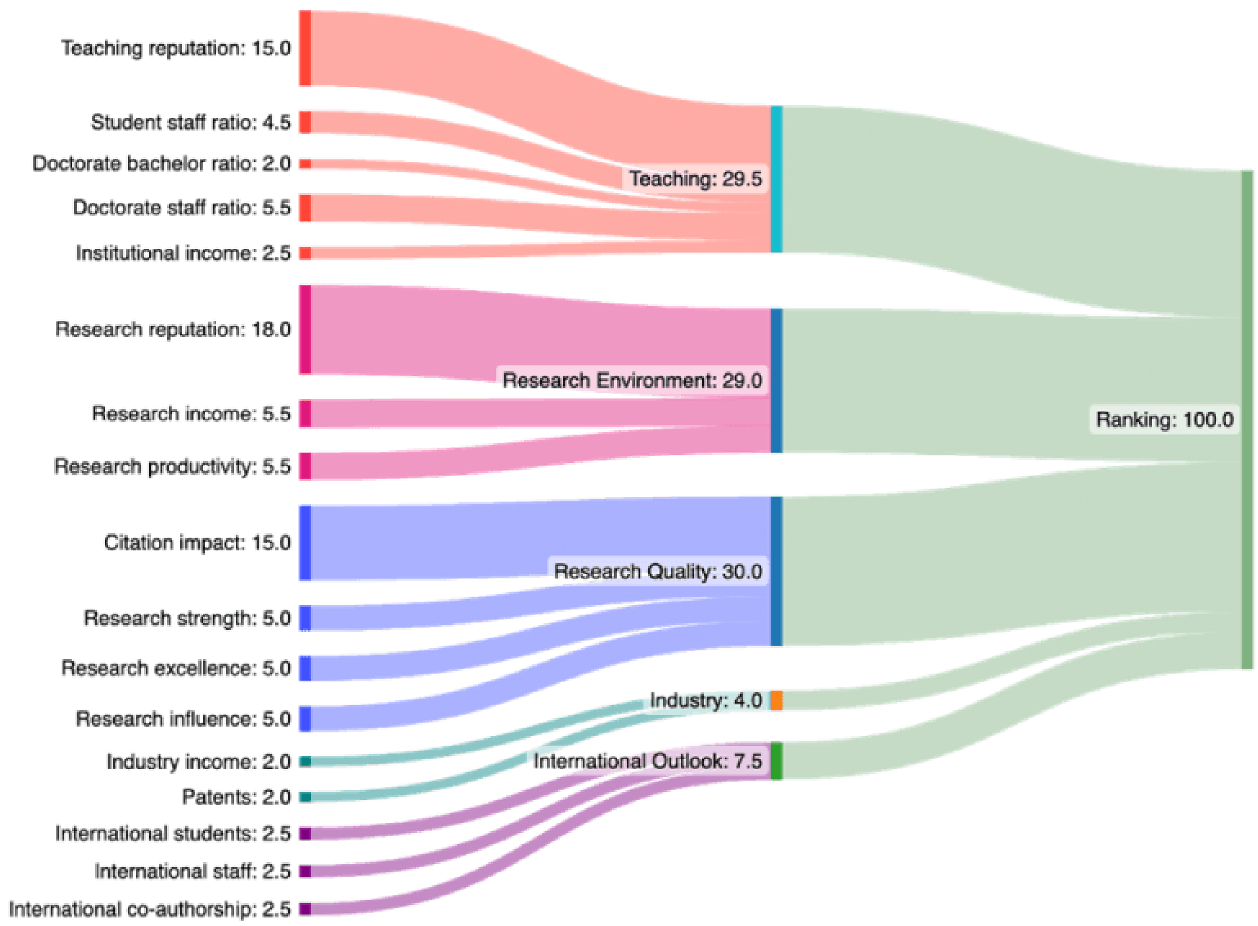


Figure 1: The Times Higher Education Ranking Methodology and Weighting Structure.
This figure displays the five categories and 17 underlying indicators used by THE to calculate university scores. The size of each segment reflects its weighting in the overall methodology, allowing the reader to understand how research, teaching, and citations contribute to final rankings.

2.2.2.2 Academic Ranking of World Universities (ARWU)

The *ARWU*, also known as the Shanghai Ranking, was established in 2003. The *ARWU* ranking methodology is divided into four criteria, as shown in Table 5, sourced from the *ARWU* website (Shanghai Ranking, 2023).

Table 5: Criteria and Weightings in the ARWU Ranking Methodology.
This table outlines the components used in ARWU, including awards, citation measures, and publication metrics, and their respective weights.

Criteria	Definition	Weight
Quality of Education	Alumni of an institution winning Nobel Prizes and Fields Medals	10%
Quality of Faculty	Staff of an institution winning Nobel Prizes and Fields Medals	20%
	Highly Cited Researchers	20%
Research Output	Papers published in Nature and Science	20%
	Papers indexed in Science Citation Index-Expanded and Social Science Citation Index	20%
Per Capita Performance	Per capita academic performance of an institution	10%

This ranking predominantly emphasises research measures, rather than teaching quality, which may not offer a comprehensive view for prospective students. The award category, based on data collected from awards websites, is particularly challenging to score highly in, resulting in the majority of universities receiving a score of 0 in this field (Fauzi et al., 2020).

2.2.2.3 Quacquarelli Symonds (QS) World University Rankings

The *QS World University Rankings*, also known as *Top Universities*, began in 2004 and has been publishing annual rankings since. The *QS World University Rankings* is divided into six criteria, as shown in Table 6, sourced from the *Top Universities* website (Quacquarelli Symonds (QS) Top Universities, 2016). It is important to note that 50% of the ranking is based on surveys sent to academics and students, introducing a high level of subjectivity into the rankings.

Table 6: Criteria, Definitions and Weight Assigned in the *QS* World University Ranking Methodology
The table details the six indicators in QS rankings and their relative contributions to the overall score.

Criteria	Definition	Weight
Academic Reputation	A survey of academics worldwide to assess which university is producing the best research in their field.	40%
Employer Reputation	A survey of employers identifying which institutions are perceived to produce the best graduates.	10%
Student-to-Faculty Ratio	The number of academic staff relative to the number of students.	20%
Citations per Faculty	A measure of the average number of citations achieved by staff based on data collected from Scopus.	20%
International Faculty Ratio	The proportion of international faculty members.	5%
International Student Ratio	The proportion of international students.	5%

2.2.3 How the Rankings Compare

Table 7 compares the rankings of the three aforementioned ranking systems for the year 2023, with the economics filter applied. The rankings show significant variation, with each system placing different universities in almost every position. The second column of the table presents the *THE* rankings, which serves as the foundation for the remainder of the table. Notably, three universities in *THE*’s top fifteen do not appear in the other two rankings, as indicated by the empty cells in the last two columns. The data for the table was sourced from the respective rankings' websites.

Table 7: Comparison of the Top 15 Ranked Universities in Economics from the *THE*, the *ARWU* and the *QS* World University Rankings
This table shows overlaps and differences across three major ranking systems, illustrating how methodology affects rank outcomes.

University name	*THE*	*ARWU*	*QS*
Massachusetts Institute of Technology	1	3	2
Stanford University	2	4	3
Harvard University	3	2	1
University of California, Berkeley	4	5	6
University of Cambridge	5	15	12
University of Oxford	6	12	9
University of Chicago	7	1	4

Tsinghua University	8		
Yale University	9	9	8
Peking University	10		
London School of Economics & Political Science	11	8	7
University of Pennsylvania	12	14	13
Northwestern University	13	11	14
Columbia University	14	7	11
National University of Singapore	15		

2.2.4 Issues with the Commercial University Ranking Systems

2.2.4.1 The Metrics Used

Subjectivity and bias inevitably arise in ranking systems when indicators cannot be measured objectively. Factors such as academic reputation, institutional prestige, or survey-based perceptions introduce judgement-based weighting, which can disproportionately favour already well-known universities (Kehm & Erkkilä, 2014). As mentioned above, 50% of the *QS World University Rankings* is based on subjective opinions collected by way of a survey sent to anonymous people, while 33% of the *THE* rankings are derived from similar input (Council on Higher Education, 2024).

2.2.4.2 Gaming the System

Despite rankers verifying the data submitted by universities, institutions are still able to manipulate their data to influence their positioning. One common way to game the system is to purchase affiliations, where a university will pay researchers from other institutions to cite them as affiliates on their papers. Another method involves including zero-hour contract lectures as academics, thus inflating the professor-to-student ratio. Additionally, universities may manipulate application numbers by including all application inquiries, making it appear as though they receive more applications than they actually do (Council on Higher Education, 2024).

The citation metric in the university ranking system can be manipulated through the use of self-citations, as demonstrated by Mussard and James (2017). For example, Valtech University, which ranked 43rd in the 2017 *THE* world ranking, scored a perfect 100 for citations, surpassing institutions like MIT and Stanford University. The citation score, accounting for 30% of the total ranking, contributed a staggering 76% to Valtech's overall score. This was largely due to the significant influence of Professor Sundarapandian Vaidyanathan, who accounted for 86% of the university's citations. Crucially, 96% of his 12933 citations were self-citations, thus contributing 74% to Valtech's overall score.

2.2.5 Transparency and Reproducibility

While university ranking systems provide general explanations of their scores methodologies and sources of data, this proves insufficient to reproduce the rankings accurately. Key limitations include the unavailability of certain data and the lack of information on how this data is processed and manipulated (Bougnol & Dula, 2015).

2.2.6 Impact of University Rankings

Universities strive to achieve high rankings due to the numerous benefits offered, often triggering the Matthew effect of accumulated advantage. A high rank can attract applications from top academics and students, increase alumni donations, strengthen corporate associations, and secure larger government budgets. Rankings are viewed as a reflection of reputation - difficult to build and maintain, but once achieved, they generate further prestige as the Matthew effect amplifies their impact (Kehm & Erkkilä, 2014).

The ranking system effectively pits universities against each other, fostering competition among universities for top positions, often at the expense of collaboration for the betterment of academia and students. Some universities, feeling pressured to maintain or improve rankings, attempt to game the system or generate media hype to boost the subjective ranking measures.

2.3 Data Source

2.3.1 Elsevier Scopus

Elsevier is a scientific publisher and data analytics company (Elsevier, 2024). The THE uses Elsevier's data to calculate certain aspects of their university rankings. The *Elsevier Scopus Reference Guide* (n.d.) highlight's Scopus' ability to provide filters for journals and sub-sections, enabling users to extract specific affiliation data.

2.3.2 EconLit

Coupé (2003) makes use of the EconLit database, a subscription-based reasource, as his source of data in analysing the rankings of economics departments and researchers. EconLit has included author and affiliation data in its database since the late 1980s.

2.4 Collaboration

Hamermesh (2013) observed an increase in the number of authors per article over the period of 1963 to 2011. Table 8 illustrates this trend, showing that in 1963, the probability of a paper having more than one author stood at 16.3%, whereas by 2011, this had risen to 79.6%.

Table 8: Distribution of the Number of Authors per Article, 1963 – 2011
This table reports descriptive statistics showing how the number of authors per paper has changed over time.

Year	Pr{>l}	Pr{3+|2+}	Pr{4+|3+}	Pr{5+|4+}	Number of articles	Number of Authors	Average Authors per Article
1963	16.3	0	0	0	86	100	1.16
1973	28.6	2.9	0	0	119	154	1.29
1983	45.6	14	0	0	125	190	1.52
1993	55.1	28	9.5	0	136	234	1.72
2003	74.1	28	21.4	0	135	269	1.99
2011	79.6	38.5	22.2	20	147	322	2.19

3 Building the Database of Papers

To investigate university rankings from 1940 to 2010, a comprehensive database of papers and associated information is required. The first step is to download all PDF documents from the five selected journals into accessible storage. This is achieved by web scraping JSTOR using a Python script. Due to the sheer volume of papers, Si Jia Wu, Francisco Rutayebesibwa, Sandisiwe Makhathini and I collaborated to build this database.

3.1 Data Landscape

Table 9 presents the number of articles in each journal and the sum of the authors in each journal. There are a total of 22795 papers authored by 33597 individuals, with an average of 1,46 authors per paper. *AER* stands out as the most significant contributor, with 9341 articles.

Table 9 : Summary Statistics of the Top Five Economics Journals
The table provides descriptive metrics such as total papers, average pages, and authorship patterns across journals.

Journal	Number of articles	Percentage contribution (%) to total	Number of authors	Average number of authors per article
AER	9341	41%	14052	1,50
JPE	3824	17%	5337	1,40
QJE	2891	13%	4280	1,48
RES	2618	11%	3776	1,44
ECTA	4121	18%	6152	1,49
Total	22795	100%	33597	1,46

3.2 Web Scraper

3.2.1 Aaron's Kit

During my master's coursework in Financial Technology at the University Of Cape Town in 2021, I participated in a course-wide project titled Aaron's Kit. The project aimed to create a decentralised repository of academic research accessible to anyone, free of charge. Due to time constraints of that year, the storage was centralised and never reached the decentralised goal. The project comprised of a backend system capable of downloading and storing papers, and a frontend user interface with which end users could interact. The backend included a JSTOR web scraper to automate the downloading of papers and populate the database. This web scraper requires login credentials via an account or university proxy server; in this case, the University of Cape Town's login proxy server was used (Chembezi, Oldnall & Van Zyl, 2021),

3.2.2 Improved Web Scraper

While the web scraper from Aaron's Kit could successfully download papers from JSTOR, it needed optimisation to be able to download thousands of papers efficiently. Two primary challenges were ReCAPTCHA verifications, and the daily limits imposed on the number of papers that can be downloaded through a university's portal. My colleague, Si Jia Wu, resolved the ReCAPTCHA issue by allowing a user to manually solve without disrupting the download loop, significantly increasing download speeds. The full implementation of this is detailed in Wu's upcoming paper (Wu, n.d.). To expedite the process, the downloading of papers was split between the four team members building the database. Despite being automated, the script required monitoring to address occasional errors, such as a webpage not loading correctly or unexpected pop-ups from JSTOR. The scraper was built in Python and run from an IPython notebook. Colleagues from universities across South Africa provided us with access to their university library logins, enabling multiple instances of the web scraper to run simultaneously. This process of downloading papers took approximately six months.

The web scraper can download any papers from JSTOR as long as their URLs are known. JSTOR maintains a consistent naming convention of papers across journals and years. From JSTOR, metadata for papers, including URLs, can be downloaded. Utilising this metadata, a list of URLs for papers that need to be downloaded can be generated. With this list of URLs, a script can be written to compare the names in the directory with the URL list to determine which papers still need to be downloaded. This feature ensures that the script is resumed where it left off should the program crash or the script be interrupted.

4 Building the Affiliations Database

This section details the process of extracting the affiliation and author data from article PDFs into a database. This multi-step process involved initial data extraction using Python, followed by a combination of Scopus, MTurk[2] and manual processing to complete the database.

4.1 Input Data

4.1.1 General Layout of Data in Different Journals

The 22795 papers obtained through web-scraping are stored in PDF format in a shared folder. The authors and their affiliations must now be extracted from the PDFs and converted into a consistent and workable format.

Understanding the layout of a paper is of paramount importance, as each journal has unique formatting and placement of author and affiliation data. Even within the same journal, changes in formatting and placement occurred over the years. This knowledge is essential to ensure that the PDF data is accurately extracted and converted into concise strings containing only the required data. If incorrect data is extracted, the affiliation and author information will not appear in the extracted strings.

Table 10 summarises the placement of affiliations across different journals. It can be seen that the *JPE*, *QJE*, and *RES* follow a similar format, listing the affiliation on a separate line either at the beginning or end of the paper. *AER* and *ECTA* are more complex, as affiliations are included in footnotes. These footnotes can change pages, vary in location and contain additional, unrelated information, complicating the extraction process.

Table 10: Comparison of Affiliation Placement Across Journals.
This table categorises journals based on where author affiliations are printed, showing the differences needed in extraction methodology.

Journal	Years	Placement in Paper	Notes	Example Article ID
AER	1940 - 1969	Affiliation on the first page after the name of the article	Gives school of study, affiliation, and address followed by acknowledgements to contributors in the same footnote	1814502
	1970 - 2010	Author under the title; affiliation in a footnote indicated by “*”, generally bottom left of first page		117315 Figure 4
JPE	1940 - 2010	Affiliation on the first page after the title of the article	Author on the line under the title followed by their affiliation on the next line	250074 Figure 2
QJE	1940 - 2010	Author listed under the title; affiliation at the end of the text before references		1882635

[2] *Amazon Mechanical Turk (MTurk). Available at: https://www.mturk.com/*

RES	1940 - 1970	Affiliation on the last page of the paper	Affiliation on the left of the page; author on the right	2967467
	1971 - 2010	Affiliation on the first page after the title of the article	Author on the line under title followed by affiliation on the next line	2297880
ECTA	1940 - 2010	Affiliation in footnotes; page and location shift from paper to paper		2951697

4.1.2 Master List of Articles

JSTOR provides a metadata list for each article it hosts, offering useful paper information, as summarised in Table 11.

Table 11: Variables Included in the JSTOR Master List.
This table lists each variable used in the database along with its definition.

JSTOR Master List Variable	Description
Stable URL	Links to the URL on JSTOR of the paper
Issue URL	Links to the URL on JSTOR for the issue that the paper is in
Content Type	Type of content contained in document: Article, MISC, Comment, Discussion, Rejoinder, Reply, Back cover, Front cover
Title	Title of the paper
Authors	Contains the author's names split by "," first, followed by "And"
Pages	Pages in the issue that the article is in
Issue	Journal issue in that year
Year	Year of publication of the article
Journal	Journal article is in

To assign each article a unique identifier, a new field called *"Article ID"* is built based on the stable URL of each paper. For example, the stable URL *https://www.jstor.org/stable/2967475* is processed to extract the section after the final forward slash ("/"), resulting in the "*Article ID*" 2967475.

Multiple content types exist depending on the pages of the issue and the content within those pages. For this paper, only publications categorized as the "*Article*" content type are considered.

4.2 PDF Data Extraction

4.2.1 Extracting Data

As mentioned in Section 3, multiple contributors compiled this database, with Si Jia Wu building the PDF scraper using the Pytesseract library, an OCR tool which extracts text from a PDF or image. The PDF scraper isolates affiliation text based on its location within each paper (Wu, n.d.). However, the scraper has some limitations. Character recognition is not always accurate, especially with scanned PDFs of older papers. Additionally, it is difficult to isolate only the affiliation text, and the scraper is thus prone to pulling surrounding text. In this regard, *AER* and *ECTA* proved to be the most difficult, as the location of the affiliations was inconsistent across papers.

4.2.2 Extraction Output

The PDF extraction process produces a multi-level dictionary containing the fields "*Article ID*", "*Authors*", "*Stable URL",* and "*Content Type*", drawn from the base master list and a dictionary which should contain affiliation data. The affiliations dictionary contains two options: "found" or "raw". When the data in the string most likely contains the affiliation, it is categorised as found. If the affiliation data cannot be isolated, where multiple potential strings could contain it, it is categorised as raw.

4.3 Extracting Affiliations

This section details the process of isolating and cleaning affiliation data from the PDF extraction output. The extraction is fully coded in Python and hosted on GitHub, where it is publicly available via the Github/Affiliations link. All the source data is available there, allowing anyone to reproduce the results presented in this paper. This includes the initial README file which explains code execution and required packages. Following this, each section also contains a README providing specific instructions on functionality and execution. IPython notebooks are used to facilitate code execution, as they facilitate markdown annotations for commentary and display outputs after code chunks, enabling checks to ensure the code is running as intended and allowing for easy variable tracking and debugging as code can be split up into smaller chunks.

There are numerous approaches to cleaning messy data and extracting information; multiple methods were tested through trial and error. The following method proved most effective when applied to the data in question.

4.3.1 Code Methodology

To understand the code, a few key elements need to be explained; these are listed below.

Data Frame:
A data frame refers to two-dimensional tabular data utilised by the Python Data Analysis Library (Pandas) (2024).

Fuzzy String matching:
Fuzzy string matching approximates string matching based on the similarity of two strings. The similarity metric is calculated using the Levenshtein Distance (LD), representing how far apart two strings are from being an exact match. The LD is given as a percentage, with 100% being an exact match, while lower percentages indicate less similarity. A useful function within fuzzy matching libraries is the partial ratio, which calculates the LD of a substring relative to a given input.

FuzzyWuzzy is a Python library initially developed in 2011 by Adam Cohen (2020) which uses the LD to calculate a Fuzzy Match. However, *FuzzyWuzzy* proves computationally slow when faced with a large set of possible matches. *RapidFuzz,* a Python library developed in 2020, is a refined version of *FuzzyWuzzy* (Bachmann, 2024). *RapidFuzz* has the same implementation as *FuzzyWuzzy* but achieves faster computation as it was written in C++.

Using the LD matched percentage, a threshold can be applied to Fuzzy matching to determine how close of a string match is needed. An 80-85% LD matched percentage is a loose match, which increases the likelihood of false positives. A higher threshold of 90-95% ensures greater accuracy by matching only highly similar strings (WinPure, 2024).

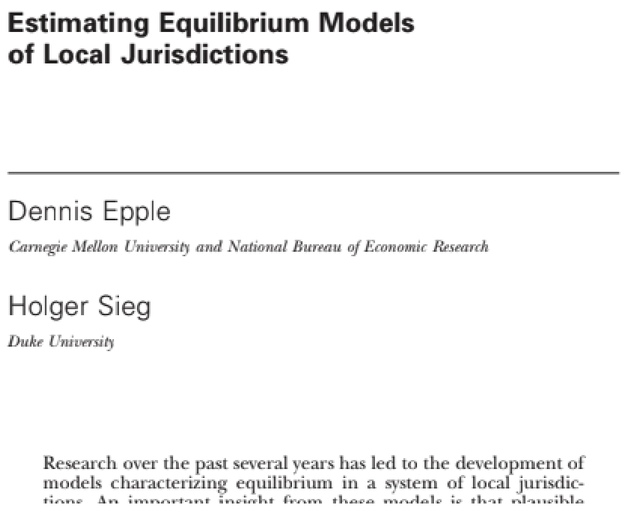
To find the optimal LD matched percentage, specific use-case testing is conducted. It is found that thresholds below 90% almost always result in false matches, while thresholds above 90% significantly reduce this occurrence. A threshold of 94% is found to be optimal, as it eliminated false matches are observed. Infrequent, minor spelling errors may still occur.

Regular Expressions:
Regular expressions are used in Python for pattern matching and string manipulation. The *re* library provides means for implementing this functionality (Python Software Foundation, 2024).

4.3.1.1 *JPE, QJE and RES*

Section 4.1.1 unpacked the varying placements of information across journals. As highlighted, *JPE*, *RES* and *QJE* are similar in structure in that their affiliations all appear on separate lines in the order of the author's names. Due to this, the same code methodology can be applied to all three.

Figure 2 provides a visual example of the structure of authors and their affiliations from *JPE*, showing the front page of a 1999 *JPE* paper with "*Article ID*" 250074. The paper follows the typical layout of *JPE* for a paper with two authors and their associated affiliations.



Estimating Equilibrium Models of Local Jurisdictions

Dennis Epple

Carnegie Mellon University and National Bureau of Economic Research

Holger Sieg

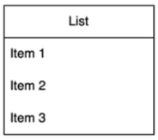
Duke University

Research over the past several years has led to the development of models characterizing equilibrium in a system of local jurisdic-

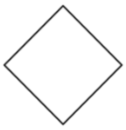
Figure 2: Example JPE Article Showing Author and Affiliation Structure.
This figure provides an extract from a Journal of Political Economy article, illustrating how author names and institutional affiliations are presented on the first page. It serves as a visual reference for how the PDF extraction process identifies affiliations.

In this example, Figure 2 shows the two authors, Dennis Epple and Holger Sieg. Dennis Epple is affiliated with Carnegie Mellon University and the National Bureau of Economic Research, while Holger Sieg is affiliated with Duke University.

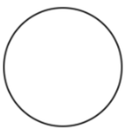
Figure 3 depicts a flow diagram explaining how the *JPE*, *QJE*, and *RES* code is constructed. Below Figure 3 is an explanation of how each stage works, with the flow diagram intended as a visual reference to track variables and steps throughout.



List blocks are output DataFrames from a code step carrying a corresponding number and a name. Within these blocks, an x represents the maximum number of lines, authors or affiliations for the paper in question.



Diamonds are decision blocks, indicating a path taken based on input data.



Circles indicate code being applied to either transform or edit the input.

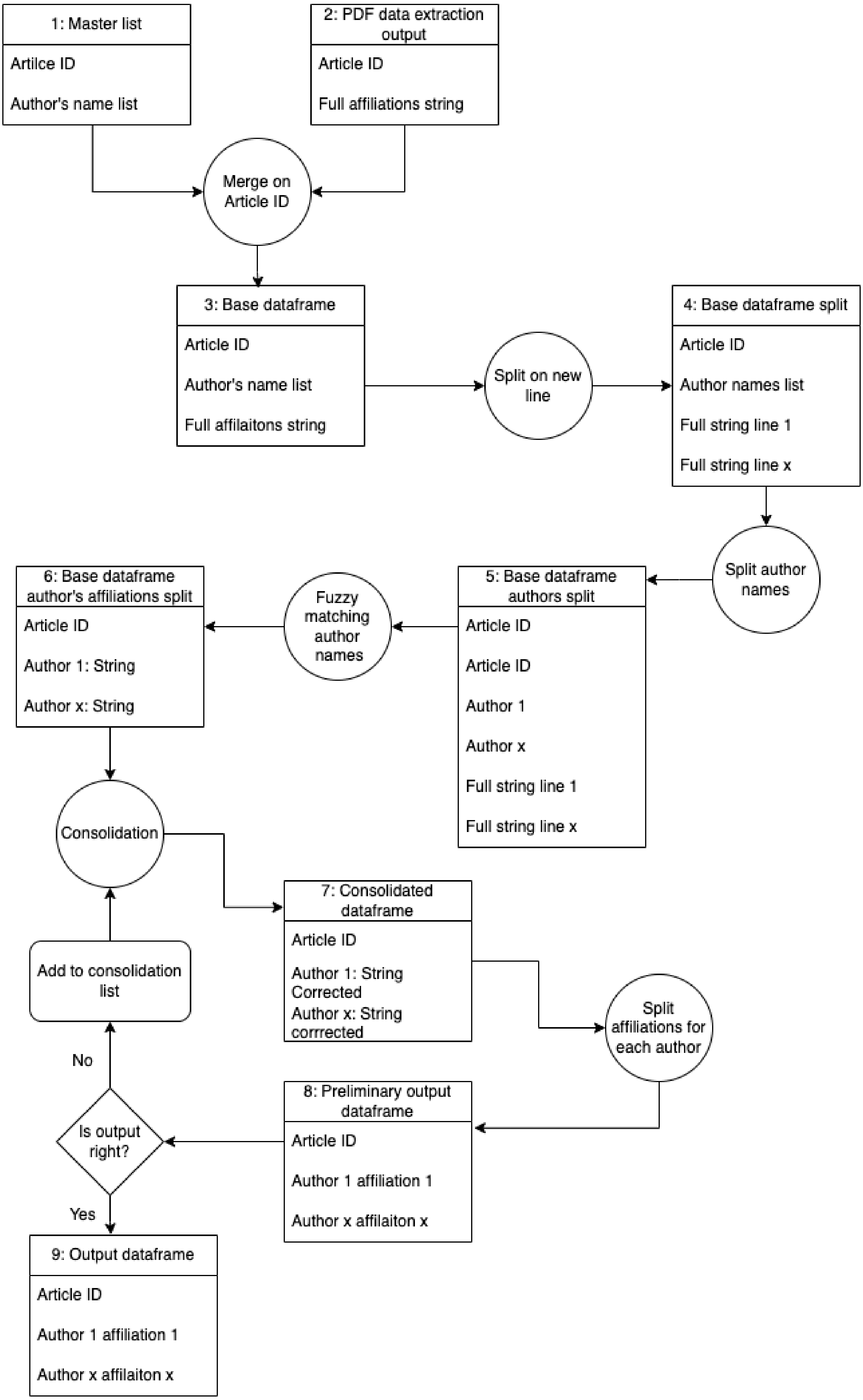


Figure 3: Flow Diagram of the Affiliation Extraction Process for JPE, QJE, and RES.
This diagram outlines the key dataframes, decision loops, and transformation steps used to extract and structure author–affiliation data from pdf files for journals with similar formatting. It provides a standalone visual summary of the code logic.

3: Base DataFrame:
The "1: Masterlist" and "2: PDF data extraction output" datasets are read into Python and converted into DataFrames. Next, these DataFrames are merged on "*Article ID*", the unique identifier for each paper. This merging process ensures that each paper is accounted for, as some papers were missing from the PDF extraction output.

4: Base DataFrame split
The full affiliations string from "3: Base DataFrame" is then split on the new line indicator for strings (\n). This separates authors and affiliations into a list format, provided the string adheres to the expected format. This is the case for the majority of the papers.

5: Base Dataframe authors split:
The "*Author names list*", downloaded from JSTOR metadata, is now parsed to extract individual authors. The list uses "And" to separate two authors, while for more than two authors, it first employs a comma (",") followed by "And". Using the Python re library, the string can be split into a list based on these two conditions. Some authors contain ", Jr." in their names such as "William G. Bostic, Jr."; this was replaced with a "Jr", removing the comma and full stop. This can then be split as above to get a list of authors.

6: Base DataFrame author's affiliations split:
Using the newly created individual author names, fuzzy string matching is applied to each full string line to search for author names. If an author's name meets or exceeds the 94% match rate threshold, that line is confirmed to contain the author's name. Following the format of the *JPE*, the next full string lines are the affiliations associated with that author until another author's name is found.

This leaves:
Author 1: String following their name
Author x: String following their name
Applying this logic, the author associated with the string is identified. This is important, as the final dataset must include both the affiliation and the author.

7: Consolidated DataFrame:
The final step is correcting the spelling of affiliations and standardising their format. This is achieved through a process I have termed "Consolidation".

To perform the consolidation and extraction, each string of affiliations for each author is fed into a nested dictionary-matching loop. This ensures consistent and accurate spelling of affiliations across journals.

The nested dictionary-matching loop operates with a list called "Corrections." This list contains two columns: the first contains the correct spelling, and the second contains the incorrect spelling. Each "Author x: String x" line is compared to the entries in "Corrections" using substring and length matching. If a match with an incorrect spelling is found, it is replaced with the correct spelling. The length of the string is important as multiple substrings within an affiliation string can be matched, but only one exact match can exist.

The "Corrections" list is divided into four subsets, listed below, to address the multiple ways a string may need to be corrected for accurate consolidation.

Comma in name:
This subset addresses cases where the affiliation string contains both a university and a city, separated by a comma. Corrections are made under two circumstances: when the city name is required in the affiliation for clarity, and when it is not. The city is required in "University of Wisconsin, Madison" which is corrected to "University of Wisconsin-Madison" to differentiate it from other universities in Wisconsin. Alternatively, when the city is redundant, as in "University of Heidelberg,

Germany,” it can be simplified to “University of Heidelberg” since there is only one university in Heidelberg.

Cities:

In some cases, the affiliation listed on the paper is the name of a city rather than a specific university. This is most common in older papers and has not been a significant issue since the 1970s. When a city is associated with a single major university, it can be safely assumed that the city name refers to that university. For example, “Oxford, April 1947” can be remapped to “University of Oxford”. However, instances like “Cambridge, Massachusetts” are more ambiguous, as they could refer to institutions such as the “Massachusetts Institute of Technology”, “Harvard University” or the “University of Cambridge.” In such cases, the entry is labelled as “Cambridge Massachusetts United States (City)” to reflect the uncertainty. Additionally, there are other cases where only the name of a city is listed, but further research into the paper reveals that the author is not affiliated with any major university and instead simply resides in this city.

Misspellings:

This subset makes up the majority of the corrections. It addresses any spelling errors and variations in university naming conventions. The spelling errors can originate from the paper itself or from the OCR process, where characters can be read in incorrectly.

Colleges:

This subset resolves discrepancies for universities with distinct colleges that still all fall under the same institution. The main universities with colleges are the “University of Oxford”, “Yale University”, the “University of Cambridge” and “Harvard University”

8: Preliminary Output DataFrame:

The next step involves splitting the affiliations for cases where an author has more than one affiliation in their “Author: String corrected”. This is initially done by splitting the string on the word “and,” followed by a split on a comma (“,”). This accounts for most instances of multiple affiliations. In cases where an “and” appears in the university’s name, this is accounted for by replacing the “and” with an ampersand (”&”) in the name during the consolidation step to prevent errors.

The output from “8: Preliminary Output DataFrame” undergoes manual verification in a decision loop to ensure accuracy. This decision loop is repeated until the output is correct. Unique affiliations from “8: Preliminary Output DataFrame” are selected from the output and sorted alphabetically to identify any misspellings or changes in naming conventions for universities. When an error is found, it is added to the “corrections” list, pairing the incorrect spelling with the corrected one. Before a name is added to this list, a Google search is conducted to verify spelling and naming conventions.

9: Output DataFrame:

Once the unique list has been fully corrected, the output is deemed satisfactory and becomes the final output.

4.3.1.2 The American Economic Review (AER)

Section 4.1.1 highlighted the differences in the journal's structure and layout. Unlike the *JPE*, *QJE*, and *RES*, where affiliations are presented on separate lines, the *AER*'s affiliations are contained in a footnote. This section details how the extraction process is adapted for the *AER*'s formatting.

Figure 4 illustrates the typical layout of the author and affiliation string in *AER* papers. It starts with the author's surname, followed by information about their affiliation. From the example in Figure 4, the affiliation information includes the department, address, affiliation, city, country, and email of the author. However, only the affiliation is needed, not this additional information. Following the first author's information, a semi-colon (";") serves as a delimiter, separating the next author's information. At the end of the affiliations is an acknowledgement section listing those who assisted with the paper. This section may also contain their affiliations, but they are not authors of the paper. This is the format which appears in papers from 2000 to 2010, although it has changed over the years. This is further discussed below.

> * Osborne: Department of Economics, 150 St. George Street, University of Toronto, Toronto, Canada M5S 3G7 (e-mail: osborne@chass.utoronto.ca; website: www.chass.utoronto.ca/~osborne/); Rosenthal: Department of Statistics, 100 St. George Street, University of Toronto, Toronto, Canada M5S 3G3 (e-mail: jeff@math.toronto.edu; website: utstat.toronto.edu/jeff/); Turner: Department of Economics, 150 St. George Street, University of Toronto, Toronto, Canada M5S 3G7 (e-mail: mturner@chass.utoronto.ca; website: www.economics.utoronto.ca/mturner/). This paper was written while Osborne was employed by McMaster University. We thank Carolyn Pitchik and two anonymous referees for very helpful comments and Robin G. Osborne for information on ancient Athens. We gratefully acknowledge financial support from the Social Sciences and Humanities Research Council of Canada (Osborne and Turner) and the Natural Sciences and Engineering Research Council of Canada (Rosenthal).

Figure 4: Example AER Article Showing Footnote-Based Affiliations.
This figure shows how the American Economic Review presents author affiliations in footnotes rather than on the first page, highlighting why a separate extraction method is required for this journal.

Despite this general layout above, the author's affiliation information may vary from paper to paper. Unlike the *JPE*, *QJE*, and *RES*, which rely on consistent formatting to split the separate affiliations into their own strings, the *AER*'s formatting varies across papers.

Figure 5 provides a flow diagram of the code used for the extraction of affiliations from *AER* papers. Similar to section 4.3.1.1 above, this serves as a visual aid for the explanation below. The shapes used in the diagram correspond to those explained in section 4.3.1.1.

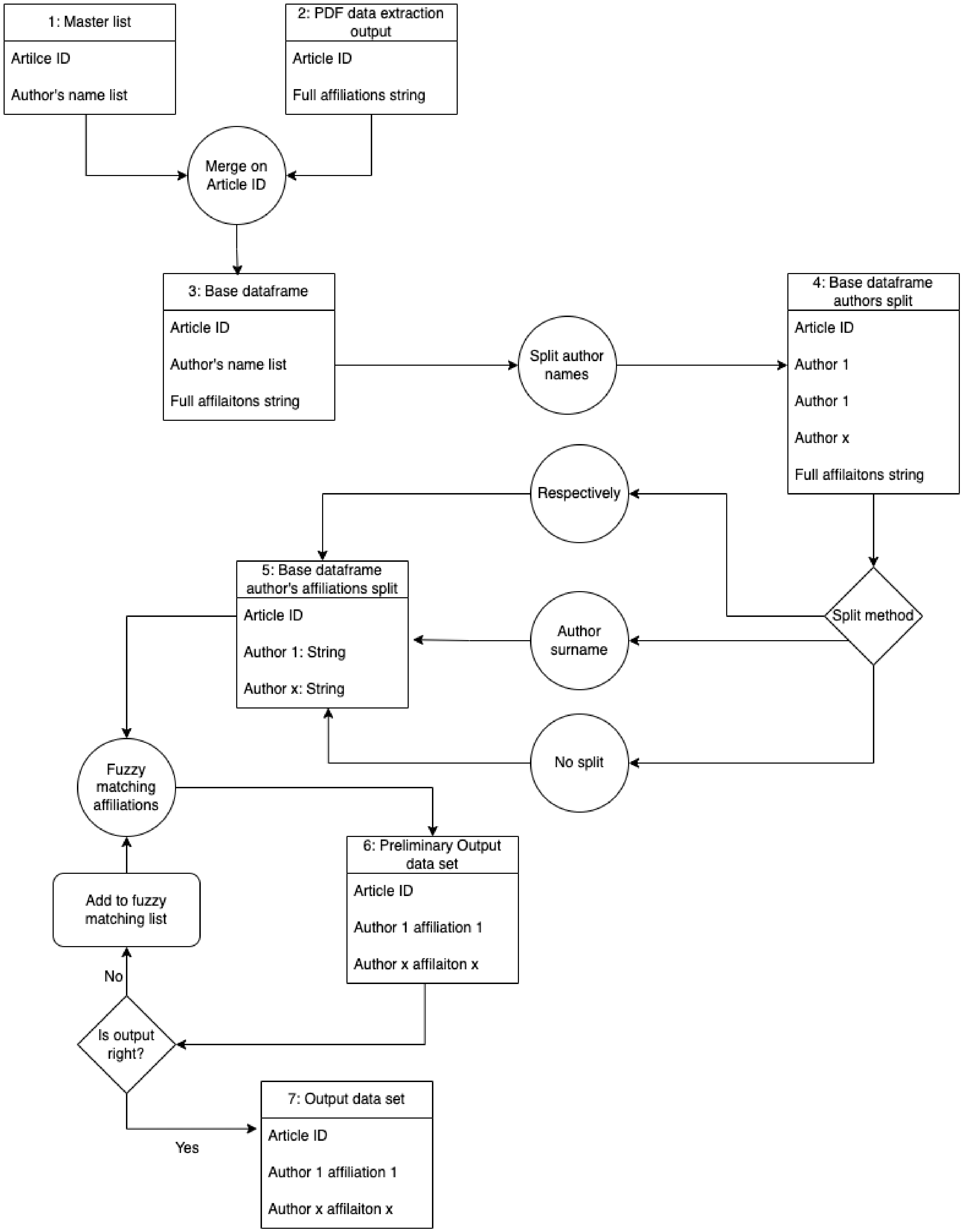


Figure 5: Flow Diagram of the Affiliation Extraction Process for the AER.
This flow diagram illustrates the steps, decision branches, and transformations unique to parsing AER's footnote-based affiliation structure. It provides a complete overview of the workflow used to generate the final dataset.

4: Base DataFrame split
The process of arriving at "4: Base DataFrame authors split" follows the same methodology employed for the *JPE*, *QJE* and *RES*.

5: Base DataFrame author's affiliations split
In cases where there is only one author, the affiliation corresponds to that author, making this step unnecessary. However, for papers with more than one author, distinct approaches are necessary due to the inconsistent formatting. Testing on different paper formats covering the full dataset finds three methods for splitting based on the input data:

1. Respectively: In this method, all of the authors' affiliations are listed consecutively, concluding with the term "respectively." The text is split by either "and" or a semi-colon (";") as delimiters, which typically appear within the sentence.

   Below is an example (1970, Article ID: 117205):

   *Department of Economics, Massachusetts Institute of Technology, 50 Memorial Drive, Cambridge, MA 02139 (e-mail: daron@mit.edu); and Department of Political Science, University of California-Berkeley, 210 Barrows Hall, Berkeley, CA 94720 (e-mail: jamesar@socrates. berkeley.edu), respectively. We are grateful to Gregory Clark and Gene Grossman for their helpful comments.

2. Author surname: This refers to cases where an author's surname is listed within the string. Fuzzy matching is used to locate an author's surname, allowing the text leading up to the next author's surname to be allocated to that specific author.

   Below is an example (1991, Article ID: 2006917)

   *Fair: Cowles Foundation For Research in Economics, Yale University, New Haven, Ct 06520, And National Bureau Of Economic Research; Dominguez: Kennedy School Of Government, Harvard University, Cambridge, Ma 02138, And National Bureau Of Economic Research. We Are Indebted To The Referees For Many Helpful Comments.

3. No split: In some cases, there is no specific differentiation between affiliations in a paper. This implies that both/all authors are from the same university; generally, when only one university is listed in the paper.

   Below is an example of this formatting (1978, Article ID: 1805279). Two authors wrote this paper, but Princeton University is the only listed affiliation. Therefore, it is known that both authors are from Princeton University.

   *Princeton University. We would like to thank Saul Hymans and Harold Shapiro for generously supplying us information concerning the Michigan Quarterly Econometric Model; Keith Carlson for gracious help with the St. Louis Model; Alan S. Blinder, George H. Borts, and a referee for useful comments; and the National Science Foundation for financial support.

Table 12 presents the number of papers processed using each split method from 1971 to 1999. There was no clear favoured method, thus conveying the inconsistent formatting of the *AER*.

Table 12: Split Methods for Multi-Author Articles in the AER (1971–1999).

The table summarises how affiliation strings were formatted for multi-author papers in the AER, informing extraction challenges.

Split Method	Author Surname	No Split	Respectively	(blank)	Grand Total
Count of Occurrence	468	627	744	8	1847

For papers published from 2000 to 2010, the author's surname is consistently listed before their affiliation, thus aligning exclusively with method 2 ("2. Authors surname") shown above. In contrast, papers from 1940 to 1969 exhibit extreme variability, with only 50% of the papers containing an affiliation string. The methodologies shown above were applied to all papers with an affiliation string. The method for getting the affiliation data for papers with no affiliation string is shown in Section 4.5.

6: Preliminary output dataset
For the *AER*, fuzzy string matching is used to extract affiliations.

For example, in Figure 4, the affiliation string for Osborne is as follows:
Osborne: Department of Economics, 150 St. George Street, University of Toronto, Toronto, Canada M5S 3G7 (e-mail: osborne@chass.utoronto.ca; website: www.chass.utoronto.ca/"osborne/);

As illustrated above, the affiliations string for the *AER* has multiple components. Fuzzy matching allows for formatting changes as only a substring of characters is searched for.

For instance, if an input of "University of Toronto" is fed into the partial ratio function with the minimum matched ratio set to 94% (consistent with the threshold used for author name matching), the output will be 100. Even if the input is changed to reflect a spelling error, such as replacing the "r" in university with an "n," the output is 95.24, still indicating a match despite a lower LD due to the error. Based on the match output, the author's affiliation is confirmed to be the University of Toronto.

For the author fuzzy matching, the author's name is known for each paper, simplifying the process. In contrast, matching affiliations requires creating a list of possible matches to compare against affiliation strings. This list must consist of correctly spelled, unique university names. For this reason, the *AER* was intentionally processed last, allowing the list of university names to be based on the correctly spelled university names from the *RES*, *JPE* and *QJE*. However, this list is not 100% exhaustive. Analogous to creating the list of incorrectly spelt university names, this is done iteratively. Rows with unmatched affiliations are manually reviewed, with their respective names being added.

A fuzzy matching loop is created to test every university in the unique list against the affiliation strings. This outputs a match ratio for each university. If this match ratio exceeds 94%, it is successfully matched.

7: Output DataFrame:
Once the unique list has been fully refined and verified, the output is deemed satisfactory and will serve as the final output.

4.3.1.3 Econometrica

MTurk is a crowdsourcing marketplace that enables individuals to outsource their processes and jobs to a distributed workforce who complete tasks virtually (Amazon Mechanical Turk, 2024). It provides an environment suitable for extracting the remaining affiliation data. However, MTurk has some shortfalls: It is relatively expensive, requires oversight to ensure that workers fill in information accurately, and necessitates manual checking and correction of the end data. Despite providing detailed task descriptions on MTurk, human error remains a concern.

Due to the variability and complexity of the *ECTA* data, MTurk was the sole source for extracting affiliation data for *this journal*. Table 14 summarises MTurk's contribution to the other journals.

4.4 Ensuring Accuracy

The code is tested for accuracy throughout the decision loop; however, this does not catch all output errors. To account for any remaining errors, the output is manually screened and checked to ensure its correctness.

To achieve this, the outputted data produced from the code in Section 4.3 is exported to Excel and labelled as the first passcode output. The first check ensures that the number of authors from the "1: Master List" matches the number of authors that have extracted affiliations. Next, the PDF data extraction output is compared to the affiliations from the code output to ensure each author's affiliations are allocated correctly. If the code output is incorrect, it is manually corrected using the information from the PDF data extraction.

4.5 Remaining Affiliation Data

After completing the Python output and accuracy checks, it is found that 11053 out of the 22790 papers have complete and accurate affiliations, leaving slightly less than half complete. The following sections detail the data-gathering sources used to complete the remaining affiliation data.

4.5.1 Scopus

Before building Python code to extract affiliations, an exercise is conducted to investigate the coverage of Scopus for the five journals.

Table 13 provides a summary of the Scopus data, including the start year, the number of articles with Scopus coverage, the number of articles matched to Scopus on Article ID, the total number of articles in each journal from the MasterList and finally the percentage matched from Scopus to the MasterList for the five journals. Only the *RES* and the *QJE* returned reasonable match rates of 55% and 60%, respectively. The considerable drop off from the Scopus coverage to the number of articles matched to Scopus on Article ID is due to the Scopus files not having a consistent, matchable and unique identifier for the articles. The Article ID defined in this paper is based on the URL of the paper on JSTOR. The Scopus data includes a column titled DOI (Digital Object Identifier), which serves as a stable URL for the *QJE* and the *RES*. This allows the DOI to be linked to Article ID to correctly assign papers. However, for the *AER*, the DOI field is either blank or contains a string of unknown characters which do not correspond to any JSTOR paper identifier. Matching could theoretically be performed using volume, issue, year, and page numbers, but any discrepancies between Scopus and JSTOR would have caused matching issues. For *ECTA*, the DOI format changed in 1999 to include a string of unknown characters, making it only usable for papers pre-1999.

Table 13: The table summarises how affiliation strings were formatted for multi-author papers in the AER, informing extraction challenges.
The table compares Scopus coverage across journals, showing start years, match counts, and match percentages.

Journal	Scopus Coverage Start Year	Number of Articles with Scopus Coverage	Number of Articles Matched to Scopus on Article ID	Total Number of Articles in Each Journal from the MasterList	Percentage Matched from Scopus to the MasterList
AER	1978	3372	0	9340	0%
ECTA	1974	721	69	4120	2%
JPE	1969	628	529	3823	14%
QJE	1940	2367	1723	2890	60%
RES	1940	2012	1441	2617	55%

Table 13 presents the number of papers Scopus data matches. The next step involves cleaning and formatting the Scopus data used. The Scopus data is formatted into a table with columns for authors and affiliations. When there is more than one author, Scopus uses a general notation system where each author is assigned a unique identifier, “--letter--" (e.g., a, b, c), which links them to their corresponding affiliations.

For example, the following Scopus affiliation data shows “Just, R.E.” is affiliated with both “a” and “b”, while “Hueth, D.L.” is affiliated only with “c”.

Authors:
Just, R.E.--a-- --b--
Hueth, D.L.--c--

Affiliations:
a--Department of Economics, Brigham Young University, United States
b--Department of Resource Economics, University of Rhode Island, United States
c--Department of Agricultural and Resource Economics, University of California-Berkeley, United States

The first step in processing is to match each author with their respective affiliation/s using the DOI by splitting it to match *“Article IDs.”* Unpacking affiliation string a--, it is structured as: “Department,” “University,” “Country.” For this paper, only the University is required. Next, a consolidation process similar to the one described in section 4.3.1.1 is applied to clean the data. This corrects the spelling errors and removes the “Department” and “Country” information from the affiliation string, leaving only the desired university name. This works correctly for most but not all cases, with some outputs needing to be manually changed after checking for errors to ensure accuracy.

Despite Scopus’ limited coverage across most journals, it proved useful for supplementing missing data in *QJE*, *RES*, and *AER*, where it helped to fill in the blanks from the code extraction.

4.5.2 Manual Extraction

Data was sent to MTurk in batches of around 2000 papers at a time. While waiting for batches to be processed, I manually extracted affiliation data for additional papers. This involves locating the relevant PDFs and copying the affiliation to the corresponding author.

4.6 Producing the Final Dataset

The final dataset is created by merging the data from the code output, MTurk, Scopus, and manual interventions.

After combing the data sets and filling in missing information, a final round of consolidation is conducted. This process begins by extracting all unique values from the combined five journals, leaving 3394 initial unique values. From this, each value is checked for spelling errors and reviewed to identify name changes over time. Numerous instances arise where a single university is listed under multiple names due to historical name changes or the use of shorthand. Another complication involves affiliations being listed in different languages, with some in English and others in the local language of the institution. Once all the unique values are identified, the consolidation process standardises the dataset to correct for these inconsistencies. Following the application of the consolidation process, 2033 unique university names remain.

The final dataset can be found on my GitHub site and accessed through the link Github/Affiliations/Final Dataset.

4.7 Unpacking the Contribution of Data-Gathering Methods

Table 14 summarises the contribution of the various methods used to extract the full set of affiliations, divided into five contributors. "Manually done" refers to papers where at least one part of the output generated was incorrect and required manual correction. "MTurk" includes those papers completed by Amazon Mechanical Turk. "Scopus" covers papers where Scopus output was used. "No affiliation" refers to those papers that were manually checked via their PDFs and confirmed to have no affiliations. Lastly, "Code" refers to those papers that were outputted correctly from Python. "Code" is the primary contributor to the output set. The metric "Percentage done through code" shows the proportion of the dataset derived from the Python code. The "True percentage done through code" reflects the code's actual accuracy when it received correct input data, achieving a 75% true accuracy rate across the journals.

Table 14: Contribution of Each Data Source to the Final Database.
This table shows the relative share of affiliations obtained from PDF extraction, Scopus, and manual processes.

Journal	Manually done	MTURK	No Affiliation	Scopus	Code	Total	Percentage Extracted Using Code	True Percentage Extracted Using Code
AER	1902	1770	165	13	5490	9340	58,8%	74,27%
ECTA		4120				4120		
JPE	674		37		3112	3823	81,4%	82,20%
QJE	895	89	29	1169	708	2890	24,5%	44,17%
RES	195	183	67	429	1743	2617	66,6%	89,94%
Total	**3666**	**6162**	**298**	**1611**	**11053**	**22790**	48,5%	75,09%

4.8 Metrics and Exclusions for Results

This section details the transformations and logic applied to the database to generate the final results. The first metric is a weighted count over time. Using the weighted count over time, a ranking is assigned to each university. Within the database, there are research institutes which are excluded from this analysis due to the focus being specifically on university rankings.

When ranking universities, it is clear that a weight needs to be applied to account for the differential impact of publications. The weighting methodologies considered, as provided in section 2.2.1., include:

1. Weighting of affiliations and authors to papers (Mulligan, 2022).
2. Weighting of affiliations and authors to pages (Kalaitzidakis, Mamuneas & Stengos, 2003).
3. First applying journal weights using the Laband-Piette index to weight journals before applying the KMS method (Coupé, 2003).
4. Normalising page characters and page size to a standard journal before applying the Laband-Piette and then KMS methods (Dusansky & Vernon, 1998).

The paper by Mulligan was conducted on Austrian Universities, whereas KMS, Coupé and Dusansky-Vernon were conducted on top economics journals. For comparability, this study adopts a methodology similar to Mulligan's, prioritising institutional contribution over journal weighting.

4.8.1 Weighted Count Over Time

To compare research intensity, a standardised metric is applied to quantify the number of papers a university has published. The simplest way to do this would be to count the number of papers a university has been affiliated with. However, this would not account for the number of authors or universities involved in each paper – an increasing number should dilute the impact attributed to any single university. Instead, a weighted count is calculated, assigning weights to the number of authors and affiliated institutions listed in a paper, as shown in Equation 3. Mulligan (2022) used a similar approach, assigning weights to each coauthor and their affiliated institutions.

This weighting is applied to each paper and then cumulatively summed over time. By using this method, the metric provides a clearer representation of a university's research intensity over time, dividing credit accordingly, per Equation 3, among the contributing parties.

Equation 3: Formula for weighting of authors and affiliations that will be applied

$$Weight_i = \frac{1}{\sum Authors} \times \frac{1}{\sum Affiliations_x}$$

Where i is the affiliation for which the weight is being calculated for, and x is the number of affiliations for a specific author.

For example, consider a paper with two authors. The first author is affiliated with Harvard University and the University of Chicago, while the second author is affiliated with the University of Cape Town. Harvard University will receive a weight of $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Similarly, the University of Chicago will receive a weight of $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$, while the University of Cape Town will receive a weight of $\frac{1}{2} \times 1 = \frac{1}{2}$.

4.8.2 Ranking

The ranking uses a weighted count over time as its base data. To assess performance over specific periods, the weighted count can be summed for each period and then reset at the beginning of the next.

4.8.3 Exclusions

The database includes universities, colleges, companies, and research institutes. Research institutes are establishments dedicated specifically to research, often collaborating with a university to publish their findings. Typically, authors affiliated with research institutes also have a primary affiliation with a university, as shown by Kalaitzidakis, Mamuneas and Stengos (2003). This paper focuses exclusively on comparing the research intensity of universities. Therefore, research institutes will be omitted from the results to prevent skewing, particularly when applying a weighting as in section 4.8.1.

Table 15 provides the top five research institutes and their respective paper counts. NBER is the most significant, with 1849 papers, followed by CEPR with 333 papers. These research institutions consistently list a university as the first affiliation mentioned in the paper. While these entities are included in the full database, they are omitted when calculating results.

Table 15: Top Five Research Institutes and Their Paper Count
The table lists research institutes and their total paper counts, illustrating why institutes are excluded from university rankings.

Research Institute	Count of Papers
National Bureau of Economic Research (NBER)	1849
Center for Economic Policy Research (CEPR)	333
Center for Operations Research and Econometrics (CORE, France)	40
Canadian Institute for Advanced Research (CIFAR)	33
National Institute of Economic and Social Research (NIESR, London)	14

5 Results

This section presents the results derived from the database created in Section 4. First, a combined analysis of all five journals is conducted, including a weighted count, ranking, and overview of trends in collaboration. This is followed by a closer examination of individual journals, comparing metrics of overall weighted counts and rankings across decades.

5.1 Top Five Economics Journals Combined

5.1.1 Results for All Journals

Figure 6 illustrates the top 50 universities ranked by a weighted count from 1940 to 2010, with stacked bars displaying contributions across decades. The top three universities from the weighted count are Harvard University, University of Chicago, and MIT, with weighted counts of 1122, 920, and 777, respectively (detailed in Table A 1 of Appendix A).

Figure 6 also reveals a few noteworthy step changes in rankings. For example, Stanford University, Princeton University and Yale University, ranked 4th to 6th with weighted paper counts of 659, 590 and 590, respectively, displaying average ranks of 5.6, 6, and 6.6. This suggests rank fluctuation throughout the decades. A similar trend is apparent for ranks 7-10 with the University of California, Berkeley, Northwestern University, the University of Pennsylvania and Columbia University.

Figure 7 provides a heatmap of the rankings for the top 50 universities. Harvard University holds the first ranking in six out of seven decades, only falling short to the University of Chicago in the 1960s.

Further insights can be gleaned from Figure B 1 in Appendix B, which provides a heatmap of the yearly ranking for the top six universities. As stated above, a step change separates these six universities from the rest, justifying the focus on this group. Harvard University has an average rank of 1.8, holding the top position in 61% of the years examined. The University of Chicago follows with an average rank of 2.66, leading in 17% of the years. The remaining universities maintain average ranks between 3.4 and 4.5, with each securing the top position in at least two years, indicating shifts in rankings over the years. Harvard University's lowest sustained ranking period occurred between 1957 and 1968, with its average ranking falling to 3.5 during this time.

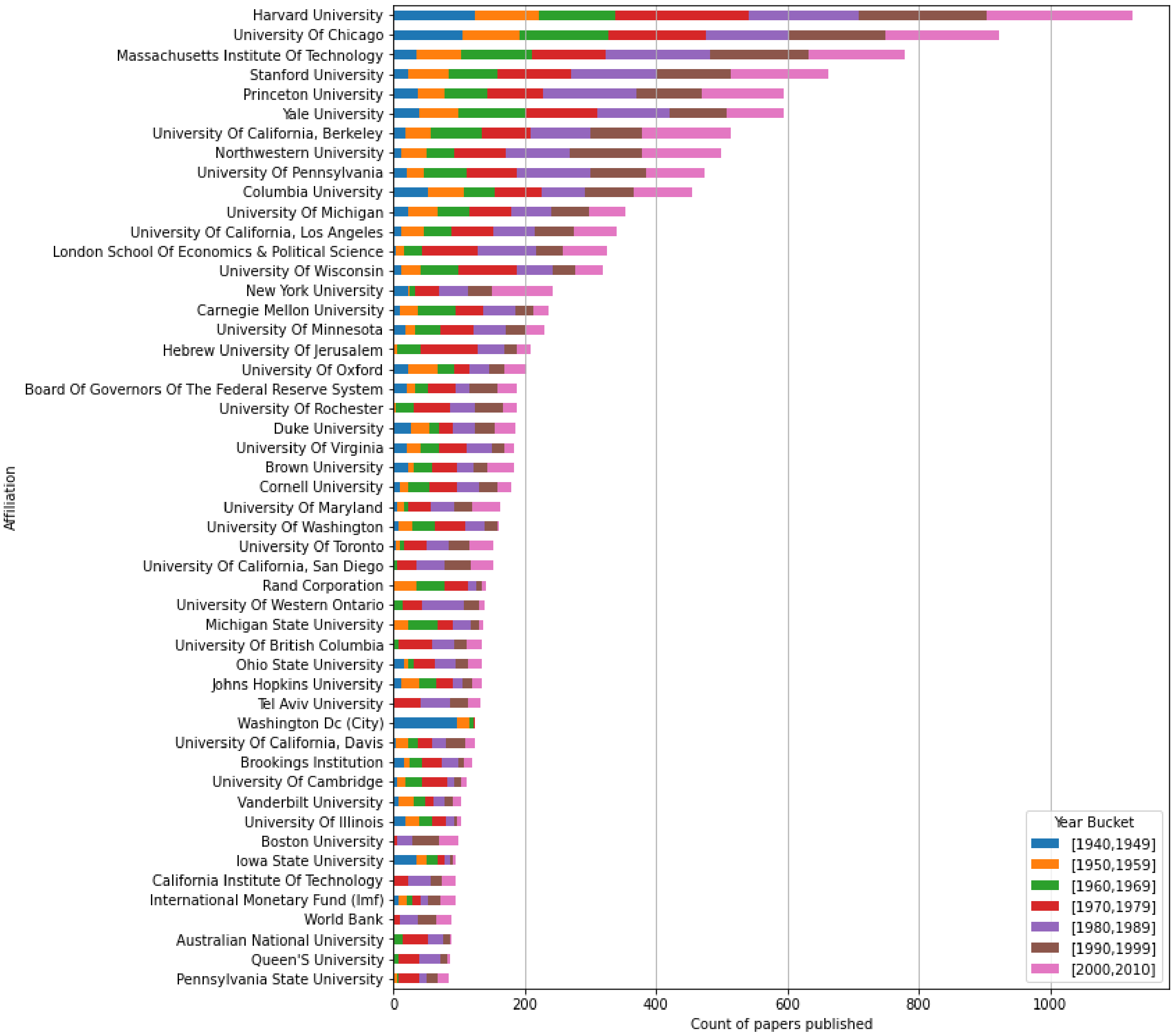


Figure 6: Top 50 Universities Ranked by Weighted Publication Count Across All Five Journals (1940–2010). Stacked bars display how each decade contributes to a university's total weighted output. This visually shows both long-term dominance and changes in publication patterns over time.

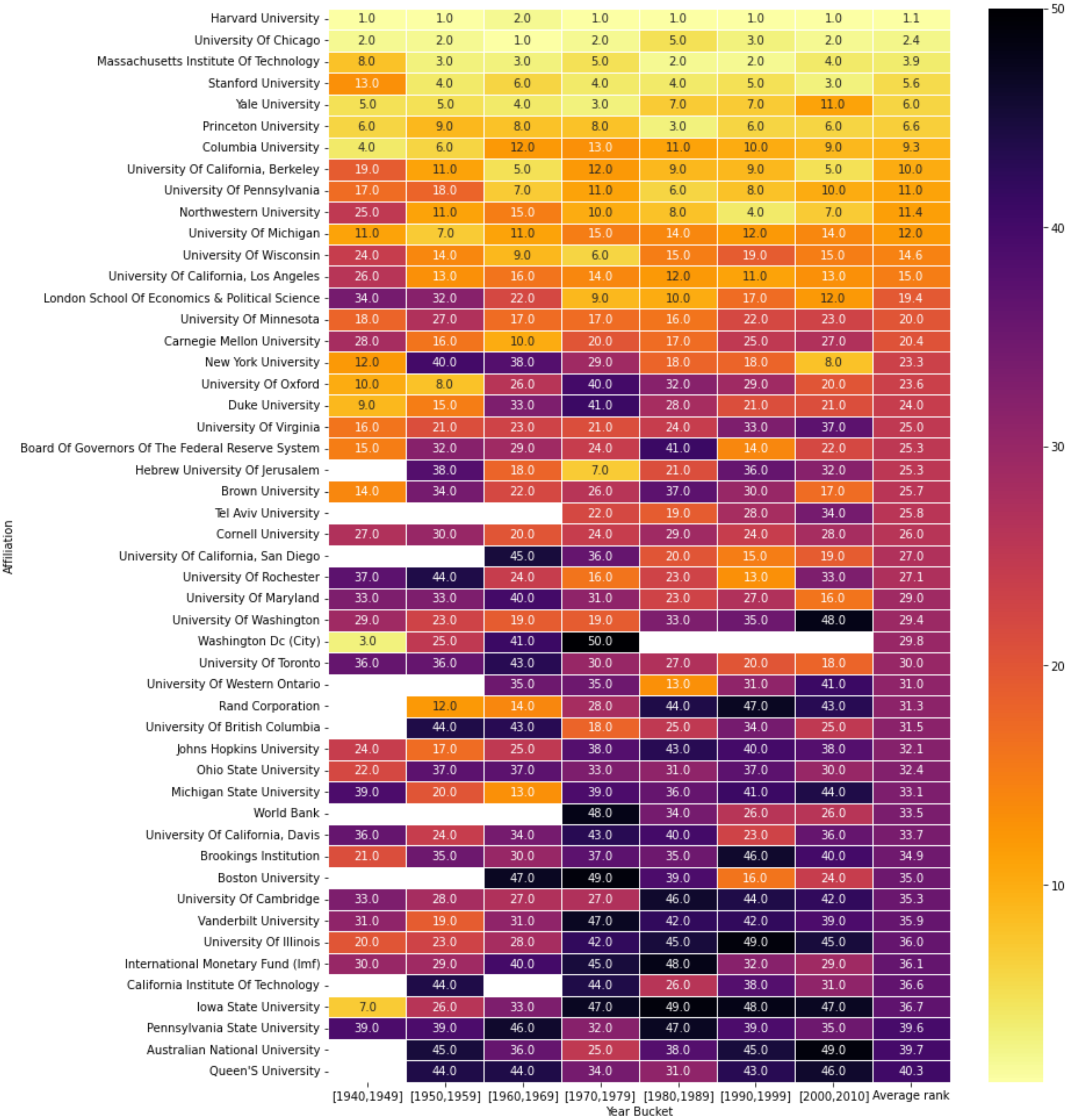


Figure 7: Heatmap of University Rankings Based on Weighted Publication Counts by Year.
This figure shows annual ranking positions for the top 50 universities, with darker colours indicating higher ranks. It enables readers to observe long-run stability and shifts in performance across decades.

5.1.2 Collaboration Metrics Over Time

This section examines how collaboration has evolved from 1940 to 2010. It begins by analysing a base metric of whether a paper involves more than one author or university, and then delves into changes in the number of authors and cross-institutional collaboration over time. Advancements in technology have streamlined collaboration by enabling seamless online communication across the globe. Technology and communication tools improvement leads to an expected increase in collaboration, both in terms of co-authorship and institutional partnerships.

5.1.2.1 Average Weight

Figure B 2 of Appendix B compares the weighted count of papers to an unweighted count, where each affiliation counts as one towards the university's total, regardless of the number of authors or universities affiliated with a publication. The rankings remain largely consistent, with the top seven universities remaining unchanged and minor shifts in those that follow. The ratio of a university's weighted count to unweighted count reveals how collaboration changes over time. Equation 4 presents the formula for calculating average weight, using the weighted count as the numerator and the unweighted count as the denominator.

Equation 4: Average Weight Assigned to a University Based on Weighted Count and Unweighted Count

$$\frac{Weighted\ count}{Unweighted\ count} = Average\ weight\ for\ University$$

A higher average weight implies fewer entities involved in writing a paper, whether it be fewer authors or affiliations.
Figure 8 displays a heatmap of the average weight for the top 10 universities across the decades from 1940 to 2010. The university names on the left represent the top ten universities. To decrease the figure size, the "…" indicates that 40 additional universities contribute to the average at the bottom. Therefore, the bottom average is calculated for the top 50 universities to provide a bigger sample size for comparison. The heatmap transitions from green on the left to red on the right, indicating an increase in the number of authors and affiliations working on a single paper over time. For example, in the 1940s, the average weight for the top 50 universities was 87%. By the year 2000, the average had declined to 43%, less than half of the value in the 1940s, reflecting a market rise in collaborative efforts over the decades.

University Name	**Weighted Count of Articles Published in Different Periods**							**Total Papers**
	[1940,1949]	[1950,1959]	[1960,1969]	[1970,1979]	[1980,1989]	[1990,1999]	[2000,2010]	
Harvard University	86%	88%	82%	74%	61%	49%	43%	60%
University of Chicago	81%	93%	74%	73%	60%	54%	44%	62%
Massachusetts Institute of Technology	88%	85%	75%	75%	65%	55%	41%	61%
Stanford University	92%	82%	75%	67%	59%	58%	48%	61%
Princeton University	93%	81%	69%	69%	61%	54%	50%	61%
Yale University	75%	87%	77%	78%	65%	62%	45%	67%
University of California, Berkeley	67%	86%	82%	65%	62%	57%	46%	60%
University of Pennsylvania	80%	83%	79%	63%	59%	48%	42%	57%
Northwestern University	86%	88%	74%	67%	62%	58%	51%	62%
Columbia University	91%	89%	77%	72%	63%	62%	42%	64%
…	…	…	…	…	…	…	…	…
Average	87%	85%	78%	69%	62%	53%	43%	62%

Figure 8: Average Weight of Publications for the Top 10 Universities by Decade.
The heatmap displays changes in average author/affiliation weight from 1940 to 2010, with redder colours indicating more collaboration. This provides insight into the increasing multi-author and multi-institution nature of research.

5.1.2.2 Number of Authors

Figure 9 compares the number of authors and the number of papers published across the decades. The line graph represents the average number of authors per paper, which has steadily increased over time. In the 1940s, the average number of authors was 1.14 per paper. By the 2000s, this had risen to 1.98, reflecting a 73% increase. This trend shows that collaboration on academic publications is increasing, with nearly two authors working on a paper on average by the 2000s.

This upward trend could be attributed in part to advances in technology, particularly the rise of online collaboration tools such as Google Docs, Overleaf, Microsoft Office 365, as well as real-time communication platforms like Zoom and Microsoft Teams. These tools have substantially reduced the cost and difficulty of collaborating across geographic boundaries, allowing researchers at different institutions to work together seamlessly. Another contributing factor is the complexity and specialised expertise required in modern econometrics research. Collaboration enables researchers to combine different skillsets to produce higher-quality work more efficiently.

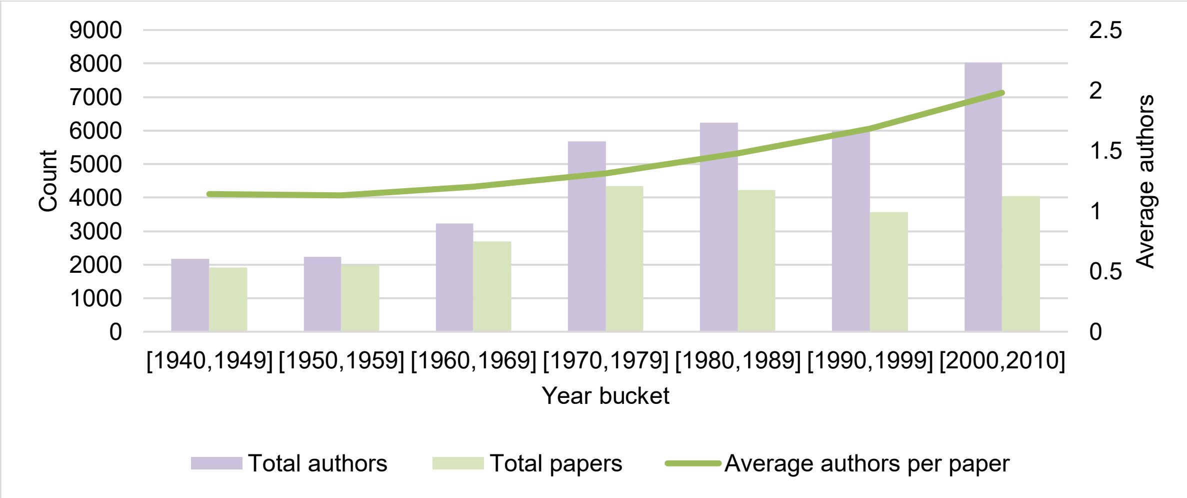


Figure 9: Total Authors, Total Papers, and Average Authors per Paper Over Time.
Bars show changes in total authors and total papers per decade, while the line tracks the rise in average authors per paper. This figure summarises the growing trend toward collaborative research.

5.1.2.3 Cross Collaboration

Figure 10 illustrates the percentage of papers with multiple authors, categorised by whether the authors are from the same university or different universities. From the 1940s to the 1960s, a higher percentage of papers were written in collaboration by authors from the same university. This trend shifts from the 1970s onwards, with a greater percentage of papers being written by a collaboration of authors from different universities. This trend of cross-university collaboration has continued to increase over time.

Cross-university collaboration has partly increased due to improvements in technology, mirroring the rise in the number of authors per paper. Digital communication tools and online collaborative platforms have significantly reduced the barriers to working with colleagues at different institutions. In addition, expanding research networks have made it easier for academics to establish partnerships across universities. As travel, communication, and international academic mobility have become more accessible, opportunities for forming and sustaining cross-institution collaborations have grown accordingly.

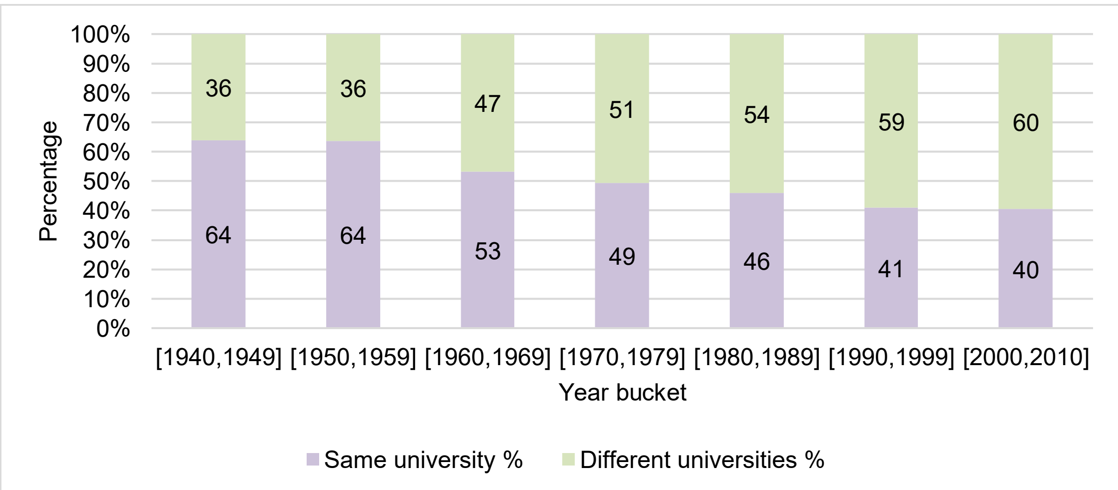


Figure 10: Share of Same University vs. Different University Co-Authorship by Decade.
The 100% stacked bars show how the proportion of cross-university collaboration has risen over time, illustrating the shift from intra-institutional to inter-institutional research partnerships.

5.2 Individual Journals Deep Dive

For each economic journal, a stacked bar chart will be presented, showcasing the top 25 universities based on their weighted count of papers calculated in this study, split into decades.

5.2.1 American Economic Review (AER)

Figure 11 illustrates the weighted count of papers published in the *AER*, grouped by year bucket. Harvard University leads in publications, followed by the University of Chicago. Table A 2 of Appendix A provides the detailed weighted paper counts used for the ranking, with Harvard University at 519 and the University of Chicago at 343. Stanford University ranks third with 308 papers, followed closely by MIT in fourth place with 305 papers.

This distribution suggests that AER publications are heavily concentrated among a small group of elite U.S. institutions, reflecting their historical editorial ties and long-standing research strength. Harvard and Chicago's dominance is consistent across decades, possibly indicating sustained institutional capacity to produce AER-quality research.

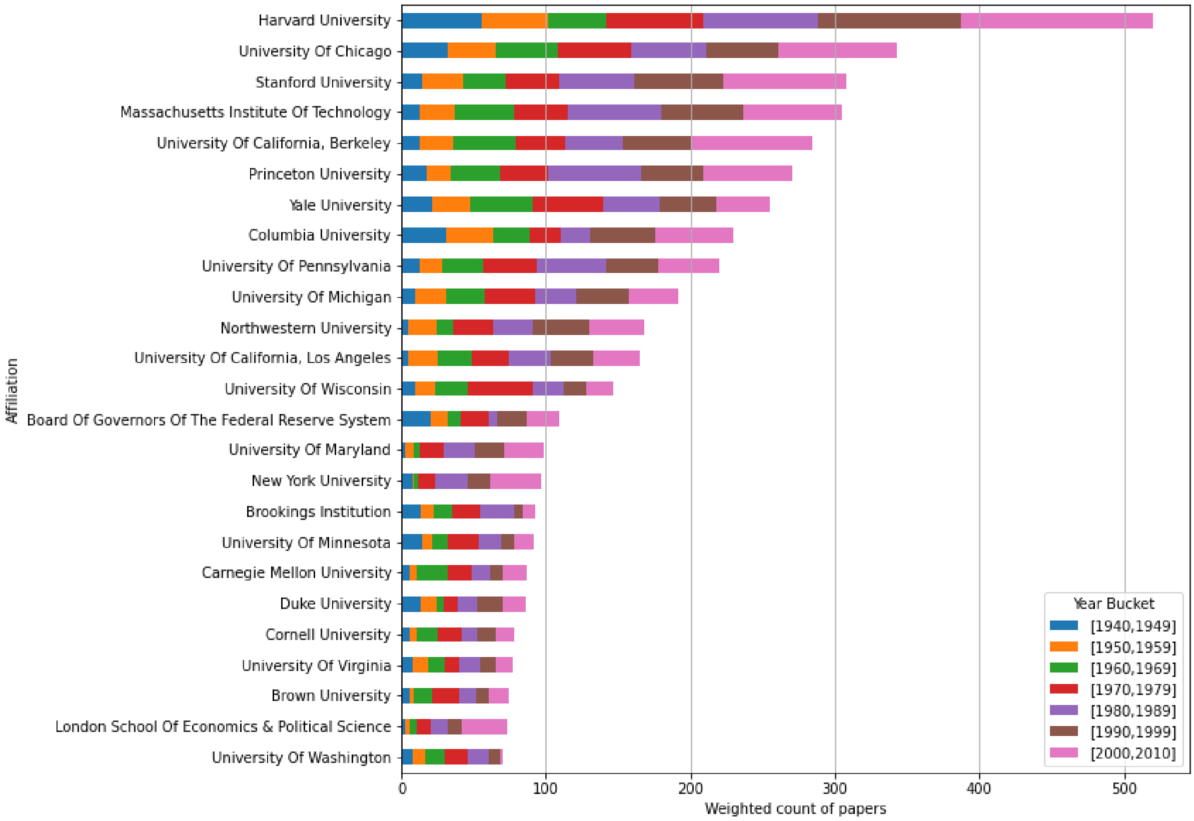


Figure 11: Top 25 Universities by Weighted Publications in the American Economic Review (AER). Stacked bars show each university's publication output by decade, enabling comparisons of long-term influence within the AER.

5.2.2 *Journal of Political Economy (JPE)*

Figure 12 presents the weighted count of papers grouped by year bucket for papers published in the *JPE*. The University of Chicago has published a significant number of papers for the *JPE*, far surpassing any other university. According to the data in Table A 3 of Appendix A, the University of Chicago has published 334 papers, followed by MIT with 123 and Harvard University with 114.

The JPE distribution shows overwhelming dominance by the University of Chicago, consistent with the journal's historic editorial base and institutional identity. Chicago's long-standing presence on the JPE editorial board appears to have reinforced its publication strength, while the large gap between Chicago and other universities highlights the strong influence of editorial networks.

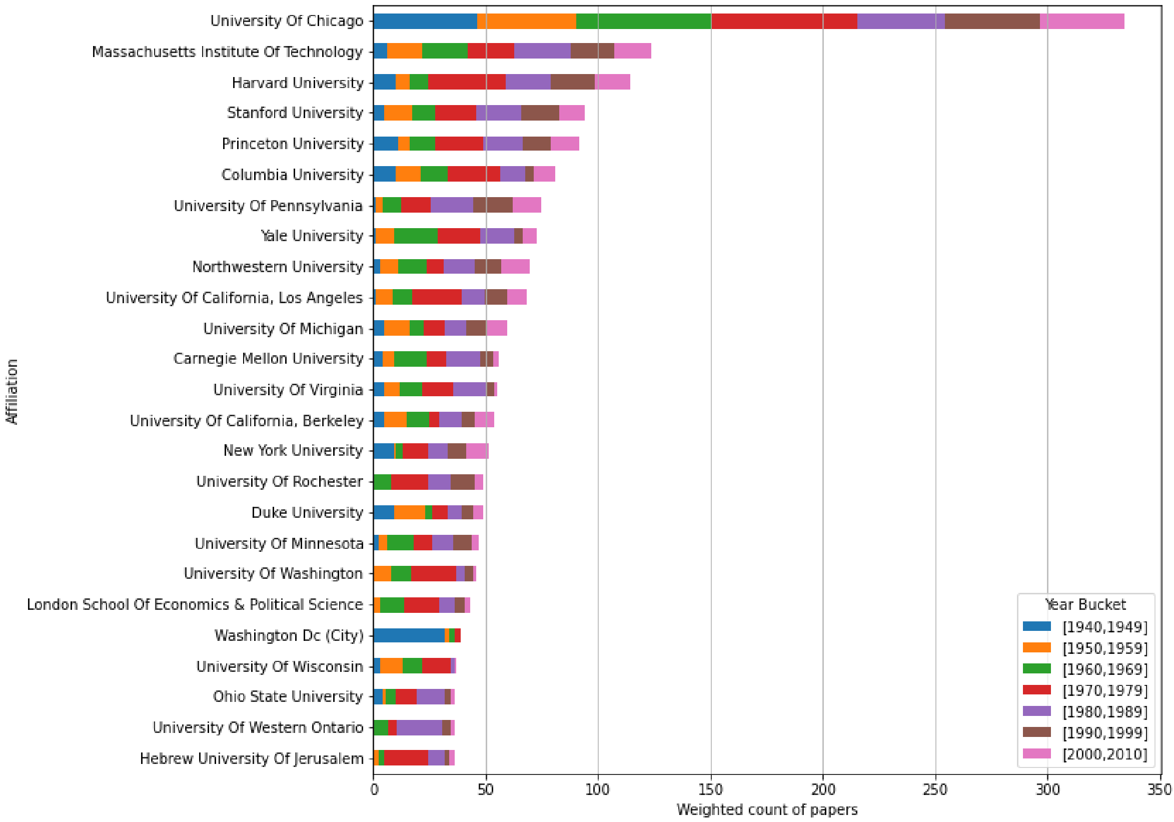


Figure 12: Top 25 Universities by Weighted Publications in the Journal of Political Economy (JPE). This figure shows the distribution of JPE publications by decade for leading universities, highlighting the journal's institutional concentration.

5.2.3 *Quarterly Journal of Economics (QJE)*

Figure 13 presents the weighted count of papers for the *QJE*, grouped by year bucket. Harvard University consistently leads the field, with a significantly higher number of publications compared to other institutions. As shown in Table A 4 of Appendix A, Harvard University has accumulated 302 weighted papers, more than double the 146 papers of MIT, the second-ranked institution.

The QJE results reveal a pronounced concentration of publications among Harvard University and MIT, reflecting not only their strong research output but also their deep historical ties to the journal's editorial leadership. The consistent dominance of these institutions suggests that proximity to QJE's editorial board has reinforced their long-standing visibility within this journal.

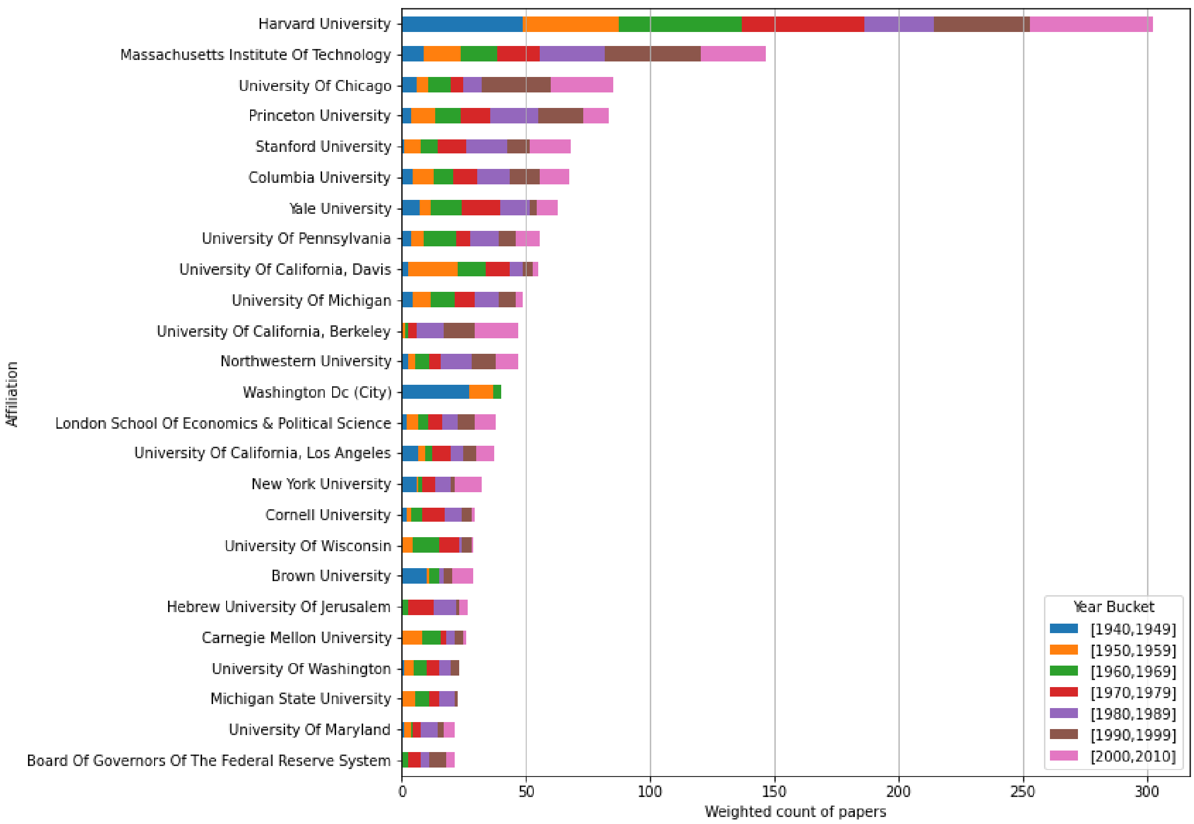


Figure 13: Top 25 Universities by Weighted Publications in the Quarterly Journal of Economics (QJE). Stacked bars reveal how QJE output is distributed across universities and time, illustrating the journal's strong ties to particular institutions.

5.2.4 *Review of Economic Studies (RES)*

Figure 14 shows the weighted count of papers published in the *RES*, grouped by year bucket. The University of Oxford emerges as the lead publisher, followed closely by the London School of Economics and Political Science. Notably, *RES* is the only journal where universities from the United Kingdom dominate the top ranks. Table A 5 provides the weighted paper counts used for rankings. The University of Oxford has published 99 papers, while the London School of Economics and Political Science has published 91 papers.

As discussed in section 2.1, Oxford academics are involved in the publication of *RES*, which may contribute to the prominent representation of the University of Oxford in this journal. However, it is important to address the anomaly in the data. Between the 1940s and the 1970s, some papers published in *RES* listed only city names as affiliations. This is evident in Figure 14, with entries such as “Cambridge, Massachusetts, United States (City)” and “London, United Kingdom (City)”. Such inconsistencies were addressed in later decades when naming conventions were revised to provide more accurate information. This period of less precise affiliation conventions can lead to an overstatement of the influence of cities like Cambridge and London, as these entries contribute to the cumulative count without properly attributing the papers to their respective institutions. This effect, though limited to earlier decades, enhances the apparent prominence of Oxford and the London School of Economics and Political Science in the *RES* publication record.

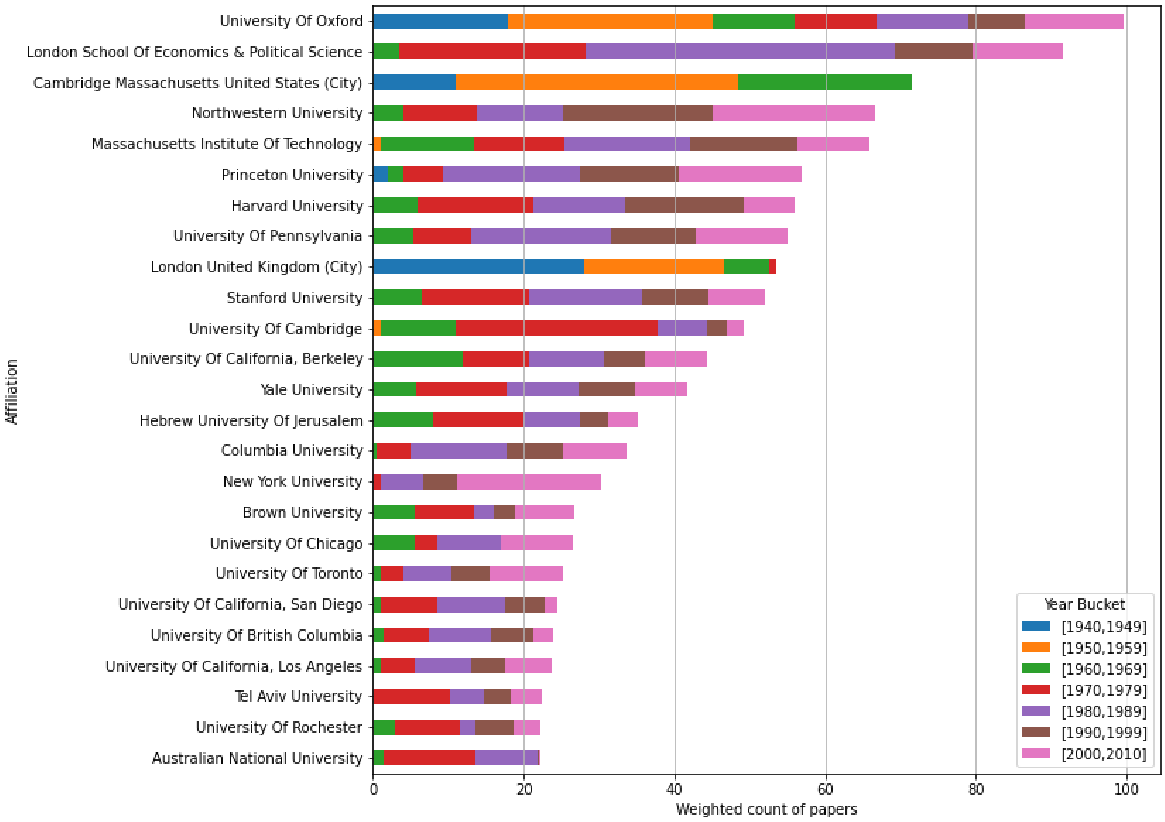


Figure 14: Top 25 Universities by Weighted Publications in the Review of Economic Studies (RES). This figure illustrates the publication patterns in RES across decades, showing a broader institutional spread than some other Top Five journals.

5.2.5 *Econometrica (ECTA)*

Figure 15 shows the weighted count of papers grouped by year bucket for *ECTA*. Yale University has published the most papers, followed by Northwestern University. As presented in Table A 6, Yale University has a weighted count of 161, while Harvard University, ranked in fifth place, has a weighted count of 133. Notably, *ECTA* exhibits the smallest range among the top five institutions, indicating a more concentrated distribution of output.

ECTA's results show greater dispersion across top universities, indicating a more diverse and international publishing landscape. The broader spread of weighted counts suggests that methodological strength in econometrics is distributed across several leading institutions rather than concentrated in a single dominant university.

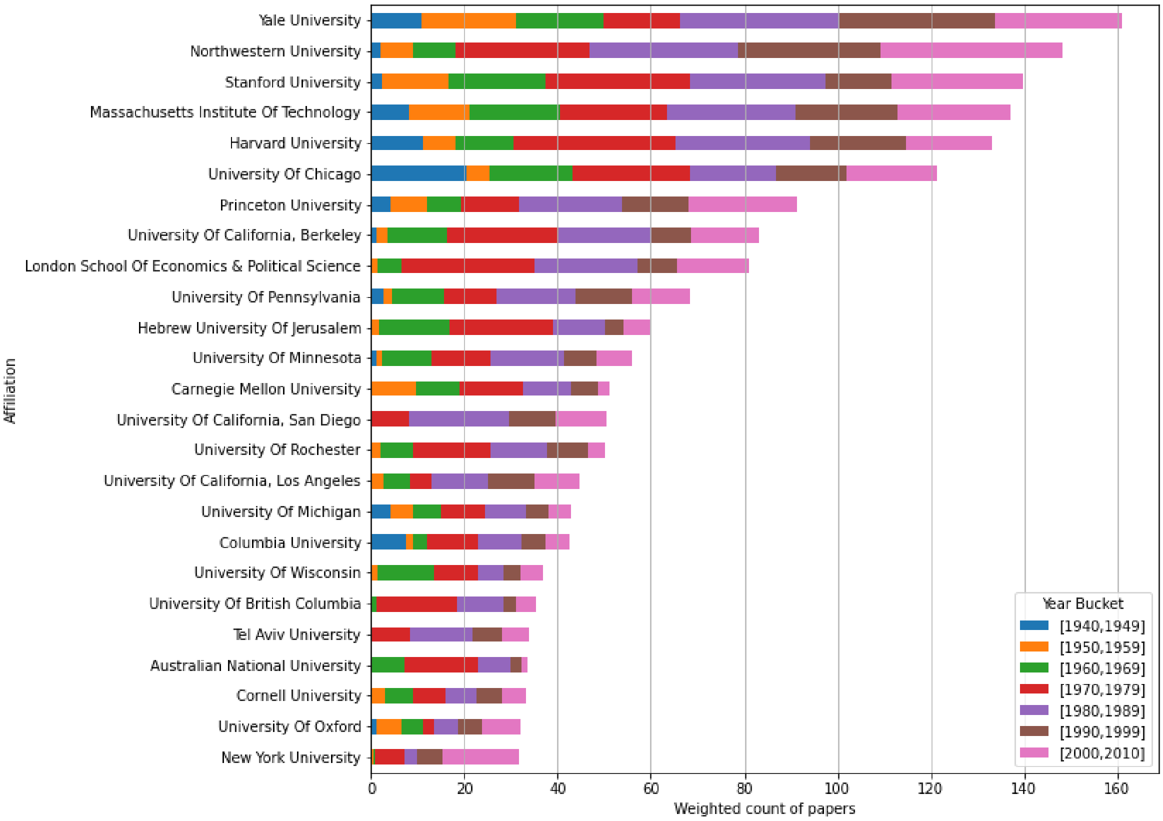


Figure 15: Top 25 Universities by Weighted Publications in Econometrica (ECTA).
This figure presents the decade-by-decade distribution of Econometrica publications across universities, demonstrating ECTA's balanced and internationally diverse output.

6 Discussion

This section compares the results of this paper to those of the comparative studies discussed in Section 2.2.1. Specifically, it compares the university rankings and collaboration metrics across the different studies.

6.1 Research Review

6.1.1 Thesis Objectives and Validation

This dissertation set out to achieve four main objectives, each of which has been fully addressed in the preceding chapters.

1. Compile a database of papers published in the Top Five Economics Journals from 1940 to 2010.

   This objective was met by downloading papers from JSTOR using the foundations set up in the Aarons kit project, from the Top Five Economics Journals. The resulting dataset, covering the period 1940–2010, is described in Section 3 and forms the foundation for all subsequent analysis.

2. Build a database of author and affiliation information linked to these publications.

   Author names and institutional affiliations were collected using a combination of automated Python matching procedures, Mturk, Scopus and manual extraction. The methodology for constructing this database is outlined in Section 4.

3. Use these databases to produce:
   a. A long-term ranking of universities based on their publication output.

      Using the database of author and affiliation information, the study generated weighted-count rankings of university research productivity over 70 years. These long-term rankings are presented and analysed in Section 5 Results.

   b. Long-term collaboration metrics to analyse inter-university partnerships.

      Co-authorship information was used to construct collaboration metrics that track cross-university partnerships over time. Trends in co-authorship and inter-institution collaboration are examined in Sections 5.1.2.2 and 5.1.2.3, demonstrating how collaborative behaviour has evolved across the period of study.

Figure 16 shows a high level summary of the data pipeline from the raw input data to the results. This further explains how the objectives have been achieved.

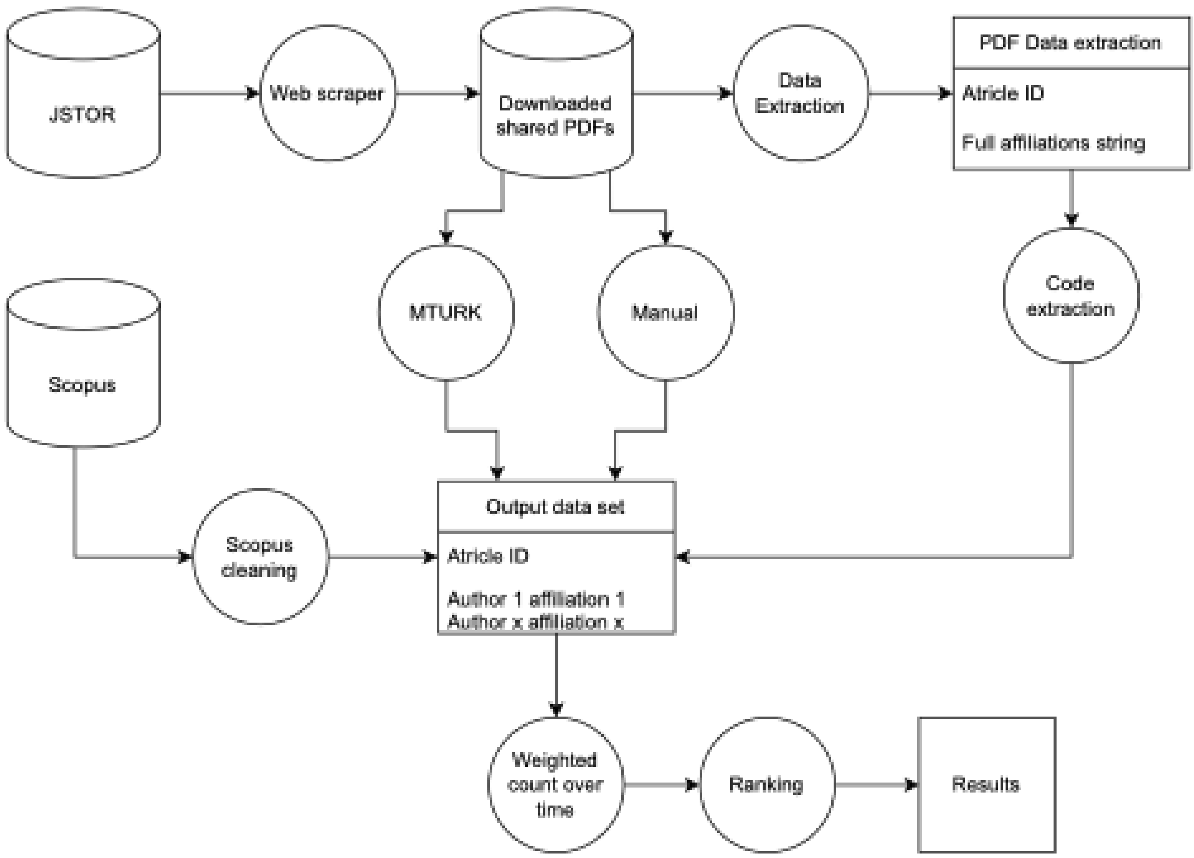


Figure 16: Flow Diagram Showing a high level summary from data input to results

6.2 Comparison of Rankings

Section 2.2.1 reviewed four major comparative studies of university rankings.
Table 4 summarises the authors, methodologies, journals and periods covered by these studies. The following compares the methodology used in this paper to the methodologies of the other studies.

Coupé (2003) applied the Laband-Piette index to assign weights to the top ten economics journals and then used the KMS methodology to rank universities for the years 1990 to 1999. To draw a meaningful comparison, Table 16 presents the rankings of the top fifteen universities based on the weighted count for the same period. A comparison of the rankings with Coupé's results reveals a high degree of similarity, with fourteen of the top fifteen universities aligning. The correlation coefficient of 0.94 indicates a near-perfect correlation between the two rankings.

Table 16: Comparison of Weighted Count and KMS rankings by Coupé, 1990 – 1999

University	Weighted Paper Count (1990-1999)	Coupé's Rankings (1990-1999)
Harvard University	1	1
Massachusetts Institute of Technology	2	3
University of Chicago	3	2
Stanford University	4	5
Northwestern University	5	4
Princeton University	6	6
Yale University	7	8
University of Pennsylvania	8	7

University of California, Berkeley	9	9
Columbia University	10	10
University of California, Los Angeles	11	11
University of Michigan	12	13
University of Rochester	13	14
London School Of Economics & Political Science	14	19
New York University	15	12

Table 17 compares the weighted count results of this study with the rankings from Dusansky and Vernon (1996) for the period 1990 to 1994. Eleven of the top fifteen universities from this paper's weighted count appear in the Dusansky-Vernon rankings. However, the correlation between the two sets is lower at 0.42, reflecting the variance in placement within the top fifteen. While Dusansky and Vernon applied a very similar methodology to Coupé, their study included a normalisation factor applied to convert pages to an *AER* equivalent. This appears to be the driving factor behind the differences. Notably, Dusansky and Vernon's study uniquely ranks Princeton University in first place.

Table 17: Comparison of This Study's Weighted Count Ranking and Dusansky-Vernon Rankings, 1990-1994

University	Weighted paper count (1990-1994)	Dusansky-Vernon ranking (1990-1994)
Harvard University	1	2
Massachusetts Institute of Technology	2	3
University of Chicago	3	17
Northwestern University	4	5
Stanford University	5	9
Princeton University	6	1
Yale University	7	7
University of California, Berkeley	8	12
University of Pennsylvania	9	3
Columbia University	10	22
University of California, Los Angeles	11	21
University of Michigan	12	23
Boston University	13	7
University of Rochester	14	12
University of California, San Diego	15	9

Table 18 shows the comparison between the weighted count and the KMS rankings for the years 1995 to 1999. Eleven of the top fifteen universities from the weighted count appear in the KMS rankings. There is a 0.85 correlation between the two sets, indicating a high degree of similarity and suggesting that the results align closely.

Table 18: Comparison of This Study's Weighted Count Ranking and KMS Rankings, 1995-1999

University	Weighted paper count (1995-1999)	KMS (1995-1999)
Harvard University	1	1
Massachusetts Institute Of Technology	2	3
University Of Chicago	3	2
Stanford University	4	8
Princeton University	5	7
Northwestern University	6	4
University Of Pennsylvania	7	5
Yale University	8	6
University Of California, Berkeley	9	9
Columbia University	10	11
University Of Michigan	11	13
University Of California, Los Angeles	12	14
University Of California, San Diego	13	12
London School Of Economics & Political Science	14	20
New York University	15	10

The comparisons presented above highlight the strong alignment between the weighted count methodology applied in this paper and other established ranking methodologies. Despite methodological differences, all three comparative studies similarly rank the top fifteen universities. While a more complex approach, such as the one employed by Dusansky and Vernon, may offer greater accuracy by incorporating additional factors such as journal impact and page numbers, the top fifteen institutions remain largely consistent. It's important to note that while these studies use different sets of journals, the application of the Laband-Piette index ensures that the top five journals will contribute most to the rankings.

6.3 Observations

6.3.1 Collaboration

Section 2.4 presented the findings from Hamermesh (2013), which showed an increase in the number of authors per paper from 1.16 in 1963 to 2.19 in 2011, based on a study of selected papers from the *AER*, *QJE*, and *RES*. This study focused on a single year of papers per decade. Section 5.1.2 of this paper extends this analysis to all papers from the Top Five Economics Journals over seventy years, revealing a similar trend. The number of authors per article increased from 1.14 in 1963 to 1.88 in 2003. While slightly lower than Hamermesh's 2003 figure of 1.99, this difference can be attributed to the inclusion of the *JPE* and *ECTA* in this study. Overall, the findings are largely consistent, with both studies confirming a trend of increasing collaboration in academic publications over time.

6.3.2 Geographic distribution

Across all the results, there is a clear geographic concentration of the top-ranked universities. The majority of the highest-performing institutions are located in the United States, with additional representation from Western Europe. This pattern aligns with the historical institutional affiliations of the editors and editorial boards of the Top Five journals, many of whom are based at leading U.S. and European universities. The geographic concentration observed in the rankings suggests that research productivity reflects longstanding inequalities in the global distribution of academic resources, editorial influence, and access to established research networks.

6.4 Limitations

A central limitation of this study is that the dataset is restricted to publications in the Top Five Economics Journals. As a result, any university that does not get published in these outlets will not appear in the rankings. The output should therefore be interpreted as a ranking of productivity within top-tier journals rather than a comprehensive global ranking of economics departments.

A second limitation arises from the use of a weighted-count approach. Although this methodology is transparent and easy to interpret, it simplifies the complexities of academic influence. Weighted counts do not account for factors such as article length, journal-specific rating variation, field differences, or citation performance. Each publication is treated as having equal impact, even though individual articles within the same journal can vary substantially in importance and scholarly contribution.

Nevertheless, the simplicity of the weighted-count approach also offers an advantage. It can serve as a useful sense-checking mechanism for more complex ranking systems. As demonstrated above, the weighted-count results closely align with those generated by more sophisticated methodologies. This makes the approach a practical tool for producing quick, preliminary insights or validating broader patterns, rather than serving as a replacement for comprehensive ranking frameworks.

Gaming the system, as noted in Section 2.2.4.2 in the context of commercial ranking systems, can also arise within the weighted-count approach. This risk stems from the equal weighting applied to all publications, regardless of their length, depth, or scholarly impact. A university could, in principle, attempt to increase its score by encouraging faculty to target shorter or less demanding papers, or by prioritising publication quantity over research quality. In more extreme cases, strategic co-authorship, such as adding additional affiliations, could artificially inflate apparent output. Mitigation comes from the fact that publication in the top 5 is extremely competitive, reducing the potential for deliberate gaming in practice.

6.5 Recommendations and Future Work

This section identifies areas for improving this study if future iterations were to occur and suggests directions for future research. Refining data extraction methods to include more recent data or incorporating additional journals would enhance the robustness of the findings. Furthermore, collaboration metrics, such as international and network-based collaborations, over time are not well documented, providing an area of research with a lot of potential.

6.5.1 Full Comparison

This paper produced a ranking similar to the methodology developed by Mulligan (2022). Incorporating alternative ranking methodologies could provide further insights into the impact of different ranking metrics. Based on the dataset produced in this study, the following approaches may be produced relatively easily:

1. Weight of affiliations and authors to papers (Mulligan, 2022)
2. Weight of affiliations and authors to pages (Kalaitzidakis, Mamuneas & Stengos, 2023).
3. First apply journal weight using the Laband-Piette index to weight journals, then apply KMS (Coupé, 2003)
4. Normalise page characters and page size to a standard journal, then apply Laband-Piette, and finally apply KMS (Dusansky & Vernon, 1998).

6.5.2 Data Extraction

The process of extracting accurate information from the OCR output proved to be a time-consuming and labour-intensive task. Despite achieving a 75% accuracy rate, manual verification was necessary for every paper to ensure accuracy. The older papers were particularly difficult due to inconsistent formatting. I recommend using a machine-learning language model to automate affiliation extraction from PDFs. This would leverage the model's ability to learn patterns and adapt to variations in layout, leading to higher accuracy, particularly with journals like *Econometrica*, where significant variability in placement and formatting proved too difficult to code a solution.

6.5.3 Extended Data Set

The dataset created for this study covers the years 1940 to 2010, Extending the data to include present-day publications would provide insights into recent trends in ranking and collaborations.

6.5.4 Collaboration Metrics

6.5.4.1 International Collaboration

Mapping universities to their respective countries would offer insights into international collaboration. For example, do universities in the United States of America primarily collaborate domestically, or do they engage in international partnerships? How has this evolved with time?

6.5.4.2 Collaboration Networks

Investigating whether certain universities exhibit a preference for collaboration with specific partners could provide valuable information for equity and inclusivity considerations within research networks. This could shed light on the extent to which top universities collaborate exclusively amongst themselves, potentially reinforcing their dominance, whilst lower-ranking universities are excluded from these collaborations.

6.5.5 Commercial ranking systems

In 2.2.2 when looking at the commercial ranking systems, it was found that these systems rely on multiple composite indicators that combine reputation surveys, citations, staff-to-student ratios, and other institutional metrics. While these cater to a wide range of factors, they can cloud the underlying publication patterns and introduce bias.

One recommendation for commercial rankings would be to incorporate a weighted top journal publication metric as a complementary indicator. This could serve as a clear quantifiable measure that is less susceptible to bias. This could also serve as a rapid benchmark against more complex ranking outputs.

7 Conclusion

This research aimed to create long-term university rankings based on publications in the Top Five Economics Journals from 1940 to 2010. To achieve this, two databases were created: a database of journal papers and a database of authors and their affiliations.

The Aaron’s Toolkit web scraper was employed to download the relevant papers. To create the database of authors and affiliations, OCR was used to extract text from PDFs based on placement, followed by the development of Python scripts for data extraction and cleaning. The Python scripts utilised two main methodologies: consolidation and fuzzy string matching. Manual checking confirmed the accuracy of the Python extraction. Additional data completion was performed by using a combination of MTurk, Scopus and manual extraction. With the authors and affiliations database refined, I applied a weighted count metric for comparison between universities, enabling the assignment of rankings based on their weighted paper counts.

The resulting database allows for long-term analysis of university research performance. By comparing the weighted count of papers across universities over time, a ranking is established. revealing Harvard University, the University of Chicago, Stanford University and the Massachusetts Institute of Technology consistently emerge as the top-performing universities.

These findings align closely with prior studies, which similarly rank these institutions highly. The rankings from this paper exhibit a strong correlation with both Coupé’s (0.94) and KMS’s (0.84) methodologies suggesting that the simplified metric of a weighted count produces results comparable to more complex ranking methodologies. While the correlation with Dusansky and Vernon's methodology is lower (0.42), the overall alignment of top-ranked institutions remains strong, indicating the robustness of the weighted count approach.

Further comparison of the top publishers from 1940 to 2010 in this study with current university ranking systems reveals a strong alignment. The *QS* and the Shanghai rankings for economics in 2023 both feature the same top four universities, while the *THE* has three out of these four. This paper explored the limitations of existing university rankings systems, including flaws in their selected metrics, their potentially unreliable data sources, and the lack of transparency in their methodologies. Despite these issues, the top-ranked universities in the contemporary university ranking systems align with the universities prominent in terms of research output as measured by weighted paper counts.

Beyond rankings, this study highlights long-term trends in collaboration. The average number of authors per paper has increased exponentially over the seven decades under review. This paper points to the growing role of technological advancements and the complexity of economics research in facilitating collaboration. Additionally, the analysis demonstrates a marked shift from intra-university to inter-university collaboration over time.

The findings also reveal journal-specific trends. The dominance of Harvard University in the *QJE*, the University of Chicago in the *JPE*, and the University of Oxford in the *RES* underscores the influence of institutional affiliation and regional focus in shaping respective journal contributions.

This paper contributes to the previous literature in two ways: first, by applying the weighted count methodology to top economics publishers; and second, by extending the ranking across an unprecedented 70-year span over the period of 1940 to 2010. The results underscore the enduring prominence of leading universities in economics research and highlight the growing importance of collaboration within the economics field.

8 Bibliography

Amazon Mechanical Turk (2024) *MTurk*. Available at: https://www.mturk.com/ (Accessed: 13 August 2024).

American Economic Association (2024) *AEA JEL codes*. Available at: https://www.aeaweb.org/econlit/jelCodes.php (Accessed: 22 October 2024).

American Economic Association (2024). *AER Editorial Board.* Available at: https://www.aeaweb.org/journals/aer/editors (Accessed: 23 November 2025).

Bachmann, M. (2024) *RapidFuzz* [Software, version 3.10.1]. Python Package Index (PyPI). Available at: https://pypi.org/project/RapidFuzz/ (Accessed: 29 September 2024).

Bougnol, M.-L. & Dula, J.H. (2015) Technical pitfalls in university rankings. *Higher education*, [Online] 69(5), 859–866.

Bush, W.C., Hamelman, P.W. & Staaf, R.J. (1974) A Quality Index for Economic Journals. *The review of economics and statistics*. [Online] 56(1), 123-125.

Chembezi, T. Oldnall, E.-R. & Van Zyl, D. (2021) Aaron's Toolkit [Web scraper, unpublished]. Cape Town.

Clarivate Analytics (1994) *The Clarivate Impact Factor*. Available at: https://clarivate.com/academia-government/essays/impact-factor/ (Accessed: 3 November 2024).

Cohen, A. (2020) *FuzzyWuzzy* [Software, version 0.18.0]. Python Package Index (PyPI). Available at: https://pypi.org/project/fuzzywuzzy/ (Accessed: 5 November 2024).

Combes, P.-P. & Linnemer, L. (2010) Inferring missing citations: A quantitative multi-criteria ranking of all journals in economics. *Working Papers halshs-00520325* [Preprint]. Available at: https://shs.hal.science/halshs-00520325v1 (Accessed: 10 October 2024).

Council on Higher Education (2024) *Briefly speaking no. 28: A critical look at the university ranking industry*. Available at: https://www.che.ac.za/publications/monitoring/brieflyspeaking-no-28-critical-look-university-ranking-industry (Accessed: 11 July 2024).

Coupé, T. (2003) Revealed performances: Worldwide rankings of economists and economics departments, 1990-2000. *Journal of the European Economic Association*. [Online] 1(6), 1309–1345.

Dusansky, R. and Vernon, C.J. (1998) Rankings of U.S. economics departments – Response from Richard Dusansky and Clayton Vernon. *The Journal of Economic Perspectives*. 12(4), 167-179.

Ellison, G. (2013) How does the market use citation data? The hirsch index in economics. *American Economic Journal: Applied Economics*. [Online] 5(3), 63–90.

Econometric Society (2024). *Econometrica – Editorial Board.* Available at: https://www.econometricsociety.org/publications/econometrica/editorial-board (Accessed: 23 November 2025).

Elsevier (2024) *Elsevier Scopus*. Available at: https://www.elsevier.com/products/scopus (Accessed: 13 August 2024).

Elsevier (n.d.) *Scopus Quick Reference Guide*. Available: https://supportcontent.elsevier.com/RightNow%20Next%20Gen/Scopus/Files/Scopus_User_Guide.pdf (Accessed: 15 July 2024).

Fauzi, M.A. et al. (2020) University rankings: A review of methodological flaws. *Issues in Educational Research*. 30(1), 79-96.

Google Scholar (2024) *Economics - Google scholar metrics*. Available at: https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=bus_economics (Accessed: 12 August 2024).

Hamermesh, D.S. (2013) Six decades of top economics publishing: Who and how? *Journal of Economic Literature.* [Online] 51(1), 162-172.

Heckman, J.J. & Moktan, S. (2020). Publishing and promotion in economics: The tyranny of the top five. *Journal of Economic Literature*, 58(2), 419–470.

Kalaitzidakis, P., Mamuneas, T.P. & Stengos, T. (2003) Rankings of academic journals and institutions in economics. *Journal of the European Economic Association.* [Online] 1(6), 1346-1366.

Kehm, B.M. & Erkkilä, T. (2014) Editorial: The ranking game. *European Journal of Education*. [Online] 49(1), 3–11.
Laband, D.N. & Piette, M.J. (1994) The relative impacts of economics journals: 1970-1990. *Journal of Economic Literature*. 32(2), 640-666.
Liebowitz, S.J. & Palmer, J.P. (1984) Assessing the Relative Impacts of Economics Journals. *Journal of Economic Literature*. 22(1), 77-88.
Mfengu, A. (2024) *Finding journal impact factors using journal citation reports*. University of Cape Town Libraries. Available at: https://libguides.lib.uct.ac.za/tracking_your_academic_footprint/journal-citation-reports (Accessed: 3 November 2024).
Mulligan, R. (2022) Publication Activity in Austrian Journals, 2011–20. *Quarterly Journal of Austrian Economics*. [Online] 25(4), 147-196.
Mussard, M. & James, A.P. (2017) Boosting the ranking of a university using self-citations. *Current Science (Bangalore).* 113(10), 1827.
Oxford University Press (2024). *Quarterly Journal of Economics – Editorial Board.* Available at: https://academic.oup.com/qje/pages/Editorial_Board (Accessed: 23 November 2025).
Perry, M. & Reny, P.J. (2016) How to count citations if you must. *American Economic Review*. [Online] 106(9), 2722–2741.
Python Data Analysis Library (Pandas) (2024) *Pandas,DataFrame* [Online documentation]. Available at: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html (Accessed: 25 September 2024).
Python Software Foundation (2024) *Regular expressions HOWTO* [Online documentation]. Available at: https://docs.python.org/3/howto/regex.html (Accessed: 29 September 2024).
Quacquarelli Symonds (*QS*) Top Universities (2016) *QS world university rankings: Methodology*. Available at: https://www.topuniversities.com/university-rankings-articles/world-university-rankings/qs-world-university-rankings-methodology (Accessed: 23 September 2024).
Review of Economic Studies (2024). *Editorial Board.* Available at: https://www.restud.com/editorial-board/ (Accessed: 23 November 2025).
Serrano, R. (2022). Top5itis revisited: Some thoughts on the notion of general interest in economics. Working Paper, Brown University.
Shah, F.A. & Jawaid, S.A. (2023) The h-index: An indicator of research and publication output. Pakistan Journal of Medical Sciences. [Online] 39(2), 315-316.
Shanghai Ranking (2023) *ShanghaiRankings' academic ranking of world universities methodology 2023*. Available at: https://www.shanghairanking.com/methodology/arwu/2023 (Accessed: 12 August 2024).
Times Higher Education (*THE*) (2024) World university rankings 2024: Methodology. Available at: https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2024-methodology (Accessed: 12 August 2024).
University of Chicago Press (2024). *Journal of Political Economy – Editorial Board.* Available at: https://www.journals.uchicago.edu/journals/jpe/board (Accessed: 23 November 2025).
Web of Science Research Team (2024) *Web of Science Journal Info.* Available at: https://wos-journal.info/ (Accessed: 12 August 2024).
WinPure (2024) The ultimate *fuzzy matching guide* [Blog, 15 November]. Available at: https://winpure.com/fuzzy-matching-guide/ (Accessed: 27 September 2024).
Wu, S.J. (in progress) Si Jia Wu Master's Dissertation [Untitled] [Unfinished master's thesis]. University Of Cape Town.

9 Appendices

A. : Data tables

Table A 1: Weighted Count of Articles Published per Decade by the Top 25 Universities in the Top Five Economic Journals

University Name	Weighted Count of Articles Published in Different Periods								Average Weighted Count of Articles Published per Decade
	1940 to 1949	1950 to 1959	1960 to 1969	1970 to 1979	1980 to 1989	1990 to 1999	2000 to 2010	1940 to 2010	
Harvard University	125	97	116	201	168	193	222	1122	160
University of Chicago	104	87	136	148	125	146	174	920	131
Massachusetts Institute of Technology	36	68	108	110	159	151	145	777	111
Stanford University	22	62	73	111	132	111	148	659	94
Princeton University	38	38	65	84	141	100	124	590	84
Yale University	40	59	99	111	109	86	86	590	84
University of California, Berkeley	18	37	78	74	91	78	133	509	73
Northwestern University	12	37	42	79	96	111	119	496	71
University of Pennsylvania	20	25	66	75	113	83	88	470	67
Columbia University	52	54	48	70	66	73	88	451	64
University of Michigan	23	44	49	62	61	57	54	350	50
University of California, Los Angeles	12	33	41	64	64	58	64	336	48
London School of Economics and Political Science	4	12	28	84	88	40	68	324	46
New York University	22	3	8	35	45	36	91	240	34
Carnegie Mellon University	9	28	55	43	47	27	23	232	33
University of Minnesota	19	14	36	51	49	29	29	227	32
Hebrew University of Jerusalem	0	5	36	86	40	18	21	206	29
University of Oxford	23	44	24	23	30	23	32	199	28
University of Rochester	1	2	27	56	37	43	20	186	27
Board Of Governors of the Federal Reserve System	21	12	19	42	20	42	29	185	26
Duke University	26	28	15	21	32	29	32	183	26
University of Wisconsin	12	29	51	52	14	14	10	182	26
Brown University	22	9	28	38	24	21	39	181	26
University of Virginia	20	21	27	43	36	19	15	181	26
Cornell University	10	12	32	42	32	27	22	177	25

Table A 2: Weighted Count of Articles Published per Decade by the Top 25 Universities in the *AER*

University Name	Weighted Count of Articles Published in Different Periods								Average Weighted Count of Articles Published per Decade
	1940 to 1949	1950 to 1959	1960 to 1969	1970 to 1979	1980 to 1989	1990 to 1999	2000 to 2010	1940 to 2010	
Harvard University	55	46	40	67	79	98	132	519	74
University of Chicago	32	33	43	50	52	49	82	343	49
Stanford University	14	29	29	36	52	61	85	308	44
Massachusetts Institute of Technology	13	23	41	37	64	56	68	305	43
University of California, Berkeley	12	23	42	34	40	46	84	284	40
Princeton University	17	16	34	33	65	42	62	270	38
Yale University	21	26	43	48	39	39	37	254	36
Columbia University	30	33	24	22	20	44	53	229	32
University of Pennsylvania	12	15	29	36	47	36	41	219	31
University of Michigan	9	21	27	34	29	35	34	191	27
Northwestern University	5	19	11	27	27	39	38	168	24
University of California, Los Angeles	5	20	23	26	29	29	32	165	23
University of Wisconsin	10	13	23	44	21	15	18	146	21
Board of Governors of the Federal Reserve System	20	11	9	18	5	20	22	109	15
University of Maryland	3	5	4	16	21	20	27	98	14
New York University	7	1	3	11	21	16	35	96	13
Brookings Institution	13	8	12	20	22	6	8	93	13
University of Minnesota	14	7	10	21	15	8	14	91	13
Duke University	13	11	5	10	13	16	16	86	12
Carnegie Mellon University	5	5	21	16	13	9	16	86	12
University of Virginia	7	11	11	9	14	10	12	77	11
Cornell University	5	4	15	16	11	12	12	77	11
Brown University	6	3	12	18	12	8	13	73	10
London School of Economics and Political Science	2	3	5	9	12	9	30	72	10
University of Washington	8	8	13	16	14	8	2	70	10

Table A 3: Weighted Count of Articles Published per Decade by the Top 25 Universities in the *JPE*

University Name	Weighted Count of Articles Published in Different Periods								Average Weighted Count of Articles Published per Decade
	1940 to 1949	1950 to 1959	1960 to 1969	1970 to 1979	1980 to 1989	1990 to 1999	2000 to 2010	1940 to 2010	
University of Chicago	46	44	60	64	39	42	37	334	47
Massachusetts Institute of Technology	6	16	19	21	24	20	16	123	17
Harvard University	10	6	8	34	20	19	15	114	16
Stanford University	5	12	10	17	20	16	11	93	13
Princeton University	11	5	11	21	17	13	12	91	13
Columbia University	10	11	11	23	11	3	9	80	11
University of Pennsylvania	1	3	8	13	18	17	12	74	10
Yale University	1	8	19	19	14	4	6	73	10
Northwestern University	3	8	12	7	13	11	12	69	9
University of California, Los Angeles	1	7	9	21	10	9	8	68	9
University of Michigan	5	11	6	9	9	9	8	59	8
University of Virginia	5	6	10	14	14	3	1	55	8
Carnegie Mellon University	4	5	14	8	15	5	2	55	8
University of California, Berkeley	5	10	10	4	10	5	8	53	7
New York University	9	1	3	11	8	8	10	51	7
University of Rochester	0	0	8	15	10	10	3	48	7
Duke University	9	14	3	7	6	4	4	48	7
University of Minnesota	2	4	12	8	9	8	3	47	7
University of Washington	0	8	9	20	3	3	1	45	6
London School of Economics and Political Science	0	3	10	15	6	4	2	43	6
Washington DC (City)	31	2	2	2	0	0	0	38	5
Ohio State University	4	1	4	9	12	2	2	36	5
Hebrew University of Jerusalem	0	2	3	19	7	1	2	36	5
University of Western Ontario	0	0	6	4	19	3	2	36	5
University of Wisconsin	3	10	9	12	1	0	0	36	5

Table A 4: Weighted Count of Articles Published per Decade by the Top 25 Universities in the *QJE*

University Name	Weighted Count of Articles Published in Different Periods								Average Weighted Count of Articles Published per Decade
	1940 to 1949	1950 to 1959	1960 to 1969	1970 to 1979	1980 to 1989	1990 to 1999	2000 to 2010	1940 to 2010	
Harvard University	49	38	49	49	28	38	49	302	43
Massachusetts Institute of Technology	9	14	15	17	25	38	26	146	21
University of Chicago	6	4	9	5	7	28	24	85	12
Princeton University	4	9	10	11	19	18	10	83	12
Columbia University	4	8	8	9	12	12	11	67	9
Stanford University	1	7	6	11	16	9	16	67	9
Yale University	7	5	12	15	11	2	8	62	9
University of Pennsylvania	4	5	12	6	11	6	9	55	8
University of California, Davis	3	19	11	9	5	3	2	54	7
University of Michigan	4	7	9	8	9	7	3	49	7
University of California, Berkeley	0	1	1	3	11	12	17	47	6
Northwestern University	2	3	5	5	12	9	8	46	6
Washington DC (City)	27	10	3	0	0	0	0	40	6
London School of Economics and Political Science	2	4	4	6	6	7	8	38	5
University of California, Los Angeles	6	3	3	7	5	5	7	37	5
New York University	6	0	2	4	6	1	11	32	4
Cornell University	2	2	4	9	6	3	1	29	4
Brown University	10	1	4	0	1	3	8	28	4
University of Wisconsin	0	4	10	8	1	3	0	28	4
Hebrew University of Jerusalem	0	0	2	10	8	1	3	26	3
Carnegie Mellon University	0	8	7	2	3	3	1	26	3
University of Washington	1	4	5	5	5	3	0	23	3
Michigan State University	0	5	5	4	6	1	0	22	3
University of Maryland	1	3	0	3	6	2	4	21	3
Board of Governors of the Federal Reserve System	0	0	3	5	3	6	3	21	3

Table A 5: Weighted Count of Articles Published per Decade by the top 25 Universities in the *RES*

University Name	Weighted Count of Articles Published in Different Periods								Average Weighted Count of Articles Published per Decade
	1940 to 1949	1950 to 1959	1960 to 1969	1970 to 1979	1980 to 1989	1990 to 1999	2000 to 2010	1940 to 2010	
University of Oxford	18	27	11	10	12	7	13	99	14
London School of Economics and Political Science	0	0	3	24	40	10	12	91	13
Cambridge, Massachusetts, United States (City)	11	37	23	0	0	0	0	71	10
Northwestern University	0	0	4	9	11	19	21	66	9
Massachusetts Institute of Technology	0	1	12	12	16	14	9	65	9
Princeton University	2	0	2	5	18	13	16	56	8
University of Pennsylvania	0	0	5	7	18	11	12	55	8
Harvard University	0	0	6	15	12	15	6	55	8
London United Kingdom (City)	28	18	6	1	0	0	0	53	8
Stanford University	0	0	6	14	15	8	7	52	7
University of Cambridge	0	1	10	26	6	2	2	49	7
University of California, Berkeley	0	0	12	8	10	5	8	44	6
Yale University	0	0	5	12	9	7	7	41	6
Hebrew University of Jerusalem	0	0	8	12	7	3	3	35	5
Columbia University	0	0	0	4	12	7	8	33	4
New York University	0	0	0	1	5	4	19	30	4
Brown University	0	0	5	8	2	2	7	26	3
University of Chicago	0	0	5	3	8	0	9	26	4
University of Toronto	0	0	1	3	6	5	9	25	3
University of California, San Diego	0	0	1	7	8	5	1	24	3
University of British Columbia	0	0	1	6	8	5	2	23	3
University of California, Los Angeles	0	0	1	4	7	4	6	23	3
Tel Aviv University	0	0	0	10	4	3	4	22	3
University of Rochester	0	0	3	8	2	5	3	22	3
Australian National University	0	0	1	12	8	0	0	22	3

Table A 6: Weighted Count of Articles Published per Decade by the Top 25 Universities in *ECTA*

University Name	Weighted Count of Articles Published in Different Periods								Average Weighted Count of Articles Published per Decade
	1940 to 1949	1950 to 1959	1960 to 1969	1970 to 1979	1980 to 1989	1990 to 1999	2000 to 2010	1940 to 2010	
Yale University	10	20	18	16	34	33	27	161	23
Northwestern University	2	7	9	28	31	30	38	148	21
Stanford University	2	14	21	30	28	14	28	139	20
Massachusetts Institute of Technology	8	13	19	23	27	21	24	137	19
Harvard University	11	7	12	34	28	20	18	133	19
University of Chicago	20	4	18	25	18	15	19	121	17
Princeton University	4	8	7	12	22	14	23	91	13
University of California, Berkeley	1	2	12	23	20	8	14	83	11
London School of Economics and Political Science	0	1	5	28	22	8	15	81	11
University of Pennsylvania	2	2	11	11	17	11	12	68	9
Hebrew University of Jerusalem	0	1	15	22	11	4	5	59	8
University of Minnesota	1	1	10	12	15	6	7	55	7
University of Rochester	0	2	7	16	12	8	3	50	7
Carnegie Mellon University	0	9	9	13	10	5	2	50	7
University of California, San Diego	0	0	0	8	21	9	11	50	7
University of California, Los Angeles	0	2	5	4	12	9	9	44	6
Columbia University	7	1	3	10	9	4	5	42	6
University of Michigan	4	5	6	9	8	4	5	42	6
University of Wisconsin	0	1	12	9	5	3	4	36	5
University of British Columbia	0	0	1	17	9	2	4	35	5
Australian National University	0	0	7	16	7	2	1	33	5
Cornell University	0	2	6	6	6	5	5	33	4
Tel Aviv University	0	0	0	8	13	6	5	33	5
University of Oxford	1	5	4	2	5	5	8	31	4
New York University	0	0	0	6	3	5	16	31	4

B. : Additional figures

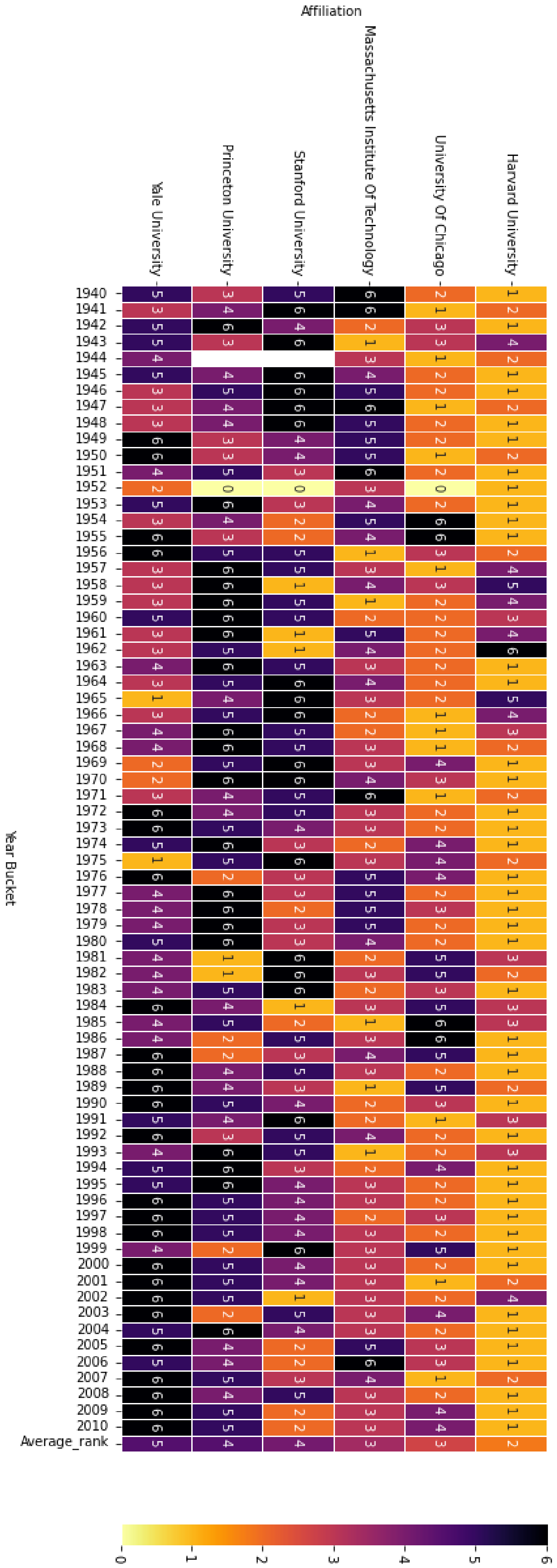


Figure B 1: Heatmap of Yearly Rankings for the Top Six Universities Across All Journals. Heatmap of Yearly Rankings for the Top Six Universities Across All Journals.

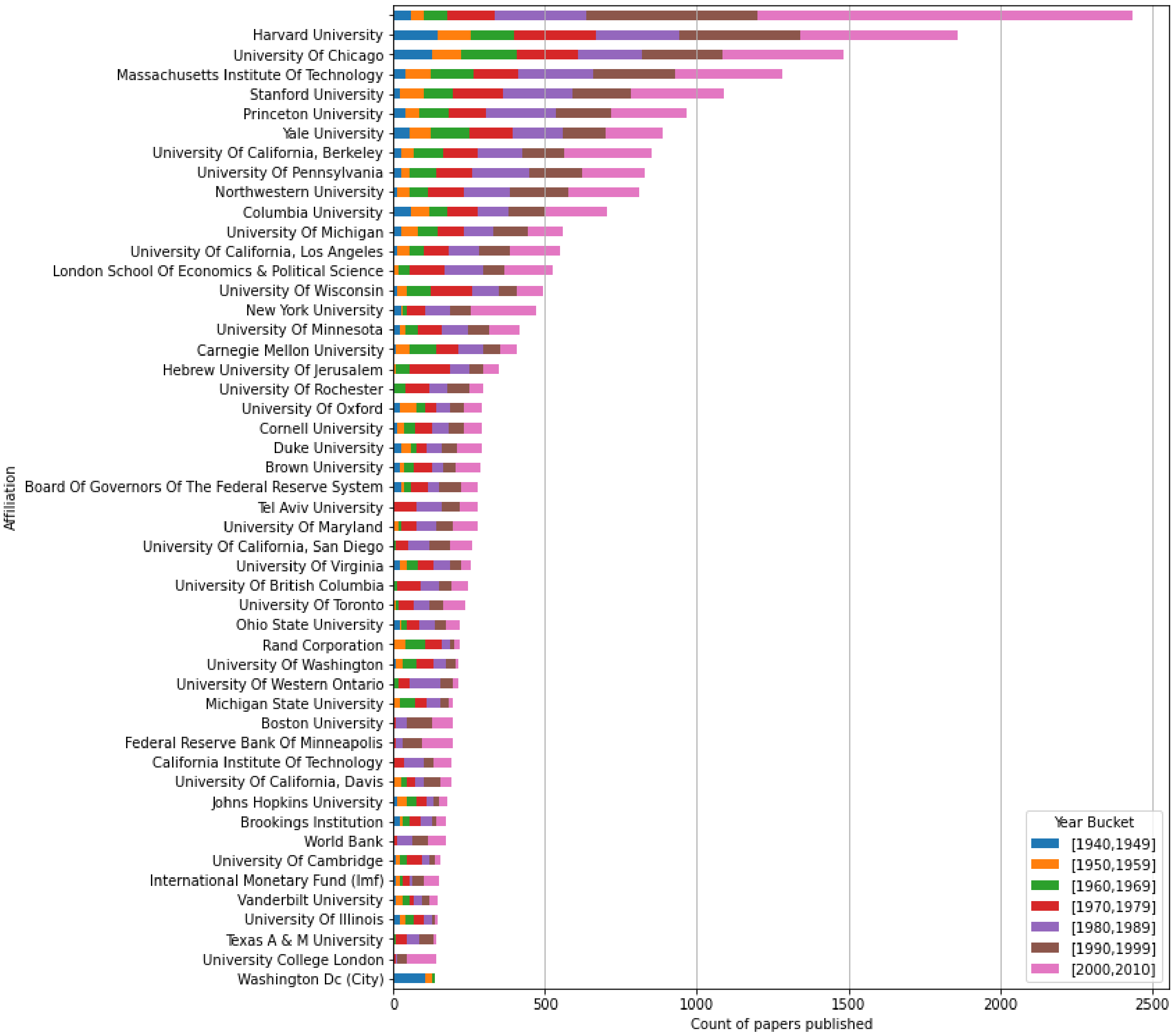


Figure B 2: Total Paper Counts for the Top 50 Universities Across All Journals.
This stacked bar chart shows raw, unweighted publication counts by decade, complementing the weighted analysis and allowing comparison of absolute output.