STA217 PROJECT
DEPARTMENT OF STATISTICS AND DATE SCIENCE

Project Report January 1, 2024

Analysis of Papers and Their Citations from the Top Four Statistical Journals: through S-field Pentagon Method

by

Likun Ni Fanzhe Liu Yemeng Xu

Abstract:

This project investigates the papers from four leading statistical journals with a focus on the distribution of articles, citation networks and temporal dynamics. Employing a novel S-field pentagon method, in which the papers are placed in a pentagon whose corners represent five different research domains, the analysis reveals a concentration of research in Computer Science and Mathematical Statistics. The integration of citation networks and the pentagon method offers a visualized image of inter-domain and intra-domain citations, highlighting the enduring importance of foundational mathematics, statistics and computational knowledge. Further examinations into journal-specific distributions unveil distinctive thematic focuses. Temporal analyses show persistent concentration in some domains over time. Scrutiny into citation time lags sorted by field and journal unveils intriguing temporal disparities, particularly in the Biostatistics and Medicine domain and the Biometrika journal. Overall, this project provides comprehensive insights into the evolving landscape of statistical research within the examined journals and domains.

1 Introduction

In the realm of statistical sciences, the academic exchanges are greatly enriched by the contributions emanating from the top four statistical journals. As pillars of statistical research, these journals are where groundbreaking methodologies, theoretical frameworks, and empirical applications converge. Our group chose the data concerning the top four statistical journals and the corresponding citation network between papers in an effort to better understand the field of Statistics.

1.1 The Top Four Statistical Journals

- Journal of the American Statistical Association: It includes articles on cutting-edge research, reviews, and discussions.
- Biometrika: It emphasizes mathematical rigor and innovative statistical techniques.
- Annals of Statistics: It covers a wide range of statistical topics and aims to advance the understanding of statistical principles and methods.
- Journal of the Royal Statistical Society Series B- Statistical Methodology: It features research articles, review papers, and discussions on the development and application of statistical methods.

1.2 Citation-Network

Citation-network is a directed graph that describes the citations within a collection of papers.The citation network serves as a conduit for the exchange of intellectual ideas and methodologies. Analyzing how ideas traverse across journals fosters a deeper appreciation for the interconnectedness of statistical research and the collaborative nature of academic discourse.

1.3 Our Work

We focus on the distributions and citation behaviors between papers with distinct subjects. To get a more accurate result, we came up with a “S-field pentagon” method where papers are classified into 5 categories to explore their distributions concerning publishers and publication dates as well as their intra-domain and inter-domain citation behaviors. Furthermore, we are interested in how the distribution of articles within each domain evolves

over time, allowing us to uncover emerging trends, shifts in research focus, and the pulse of innovation within specific disciplines. Additionally, our exploration extends to the temporal dynamics of citations so we analyzed the time lags between article publication and subsequent citations. Through this comprehensive exploration, we aim to contribute to a richer understanding of the intricate interplay between domains in the broad field of Statistics and the temporal evolution of research.

2 Data and Methods

2.1 Data Preparation

This section outlines the key steps involved in preparing the data for subsequent analyses.

2.1.1 Data Cleaning and Matching

To uphold the quality of our dataset, articles with missing values in critical fields, such as abstracts and publication years, were systematically identified and excluded. This ensures that our analysis is built upon a foundation of completeness, minimizing the impact of incomplete or inadequate information.

Then each paper was matched to its respective publishers, citation relations and related subjects according unique paper identifiers. The construction of a comprehensive citation network forms the backbone of our analysis. Each article was linked to its cited references and citing articles, creating a rich network that facilitates the exploration of inter-domain citations and intra-domain citations. And the related subjects are determined by tracing papers references one by one until level 1 ancestors are found, which can be interpreted as level 1 subjects. This is enabled by the citation relationships provided in the data. Those critical steps establish a direct association between each scholarly work and its source journal, setting the stage for precise analyses based on journal affiliations.

2.1.2 Classification

Among the level 1 subjects, only subjects with a minimum paper count of three are considered. In the data preparation stage, we meticulously organized primary academic subjects into five distinct domains based on their definitions for classification purposes. These five domains are: Computer Science and Data Science(C), Finance and Economics(E), Biometrics and Medicine(B), Mathematics and Statistics(M) as well as Interdisciplinary(I) Studies. Computer Science and Data Science refers to studies related to computing,

algorithms, and data analysis such as Artificial Intelligence. Finance and Economics contains research within the realm of finance and economics like Accounting and Marketing. Biometrics and Medicine involves statistical methodologies applied to biological and medical research and related subjects like neuroscience and biochemistry fall in this category. Mathematics and Statistics covers traditional mathematical disciplines along with statistical theory and methods. For instance, Mathematical Optimization. As for Interdisciplinary Studies, this includes cross-disciplinary research that bridges multiple academic fields which cannot be grouped in the above domains.

The figure below shows the amounts of data before and after the cleaning process.

	Before	After	Proportion(%)
Number of papers	6508	5092	84.0
Number of disciplines	186	139	74.7
Number of citations	23737	21638	91.2

Figure 1: The Amounts of Data Before and After the Data Cleaning

To explore potential changes in research domains over time, we further categorized the articles based on their publication years. Specifically, we grouped the publications into four intervals, each spanning five years, which are 1996 and 2000-2003, 2004-2008, 2009-2013 and 2014-2018. This temporal segmentation allows us to trace the evolution of research trends and identify any notable shifts or developments within the academic landscape. The line chart below illustrating the number of articles within each time interval reveals notable variations in paper counts across the different groups.

The classification process lays the foundation for the analysis of citation patterns within each domain, offering valuable insights into the distinct research landscapes of these academic areas.

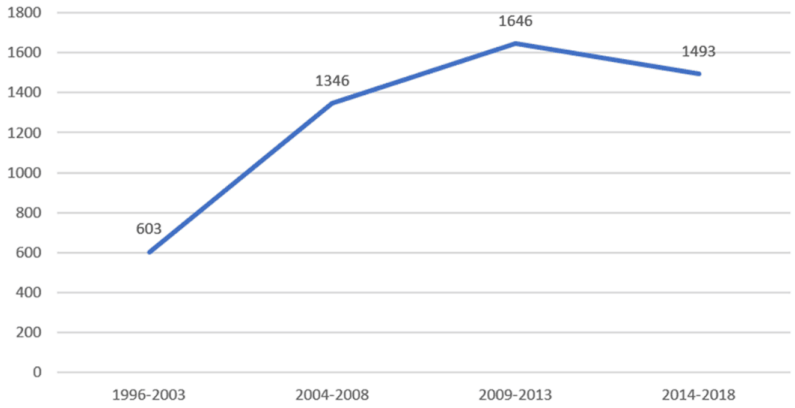

Figure 2: Paper Counts in Four Time Intervals

2.2 Method

Once the classification is done, a pie chart can be drawn to dipict the papers' distributions. As is shown in the pie chart, papers in Mathematics and Statistics as well as the Computer and Data Science take up the majority of the papers, together accounting for nearly 70%.Yet since a paper can corresponds to more than one domain, the chart is inaccurate and a new method is required to reflect the distribution more precisely.

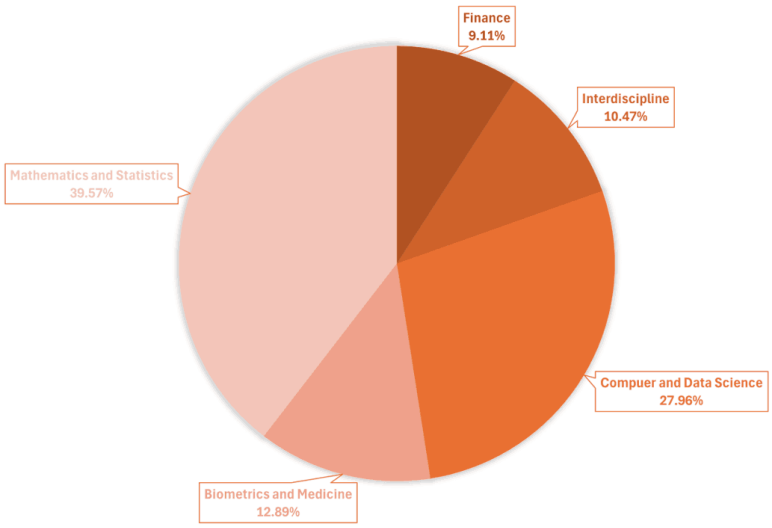

Figure 3: Paper Counts in Four Time Intervals

2.2.1 S-field Pentagon Method

The S-field pentagon method, introduced in response to the inherent complexity arising from the interdisciplinary nature of research articles, employs a geometric representation for a detailed analysis of article distribution across the five domains.

The S-field Pentagon is drawn as an equilateral pentagon centered at the origin, whose corners correspond to five research fields, represented by their initial letters, M, I, F, B, C. On a Cartesian system, each corner is a distance of 1 from the origin, with the M corner at (x, y) = (0,1). In this method, the positional coordinates of papers within the pentagon are determined by the proportional representation of level 1 subjects across the five domains. For instance, a paper whose level 1 subjects fall into a particular domain, say M, is mapped to the corresponding vertex M. And if its level 1 subjects correspond to C and M domains, the dot representing the paper should fall on segment CM. So we can map papers on the pentagon.

Furthermore, we extended our analysis by integrating this citation network into the previously introduced S-field pentagon structure. Each article positioned within the equilateral pentagon, now also carries information about its citation relationships.

3 Results

3.1 Distribution of Papers in the Five Domains and Citation

3.1.1 Distribution

In Fig 4, we positioned all papers within the pentagon to observe the overall distribution across research domains. Fig 4 right part shows that incorporates citation relationships for each paper.

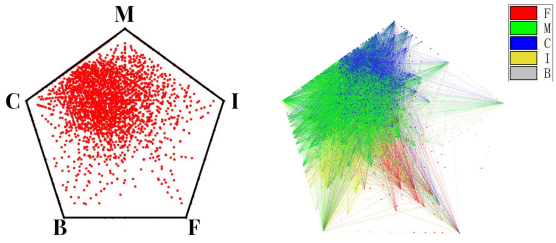

Figure 4: Paper Counts in Four Time Intervals

Notably, both the research domains and citation relationships are concentrated in the fields of Computer and Data Science and Mathematics and Statistics. The intricate and diverse intersections observed suggest that the primary focus of research in these four top journals aligns with the burgeoning field of data science, reflecting its extensive prospect.

3.1.2 Within or Between Fields

Further investigation involved categorizing articles based on their proximity to specific corners of the pentagon. This classification which a paper belongs to a specific domain only facilitated a meticulous examination of citation behaviors, emphasizing both inter-domain and intra-domain citations. A heatmap was then constructed to capture the citation relationships and proportions across different domains.

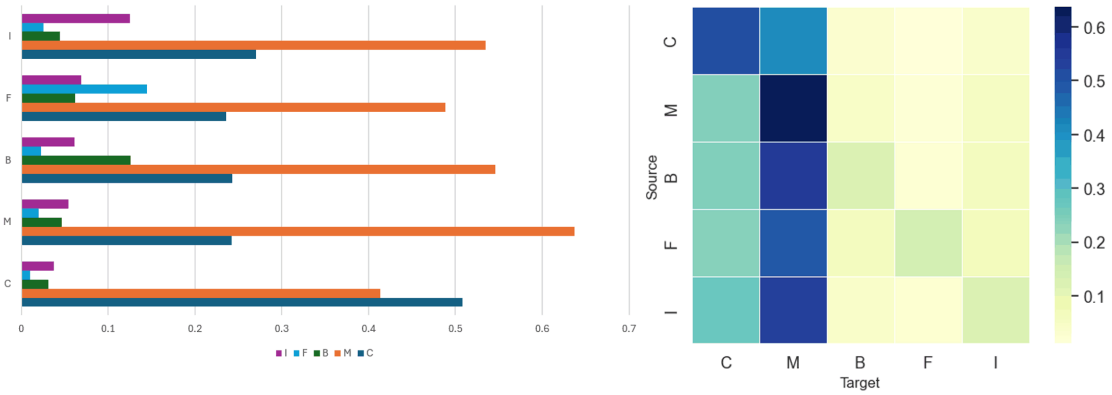

Figure 5: Paper Counts in Four Time Intervals

Regardless of the primary domain, a substantial reliance on papers related to mathematics, statistics, and computational knowledge is noteworthy. Examining intra-domain citations revealed that each domain predominantly cites articles within its own sphere.

Crucially, the analysis highlighted that, even in interdisciplinary studies, the core of citations often traced back to fundamental mathematical and statistical theories, alongside relevant knowledge from the field of computer science.

3.2 Citation between Journals

Utilizing the pentagon structure, papers were classified based on the journals they were published in, leading to the creation of Fig 6. Notably, JASA and JRSSB exhibited a concentration of publications in the fields of Computer Science, Mathematics and Statistics, whereas papers from the other two journals demonstrated a comparatively dispersed distribution across domains.

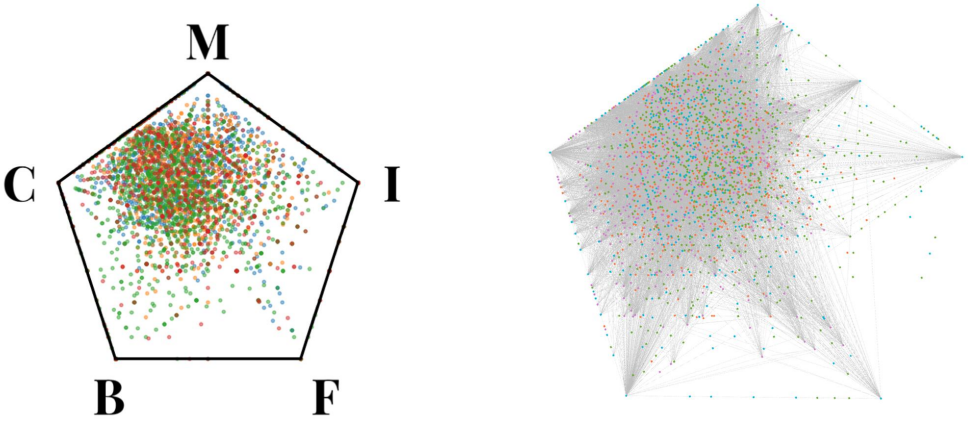

Figure 6: Paper Counts in Four Time Intervals

As is shown in Fig 7, the analysis reveals distinctive patterns – articles published in Biometrika demonstrate a lower citation proportion across the four journals, potentially due to its limited focus. While publications from AOS exhibit a broader influence, evident from a larger citation proportion.Fig 7 right part visually encapsulates the citation relationships among the four prominent journals, offering a comprehensive representation of both internal and cross-journal citations. The varying sizes of the arrows are proportional to the scales of citations. The arrows extending across journal boundaries signify citations from one journal to another while the loops on the vertexes represent inter-journal citations.

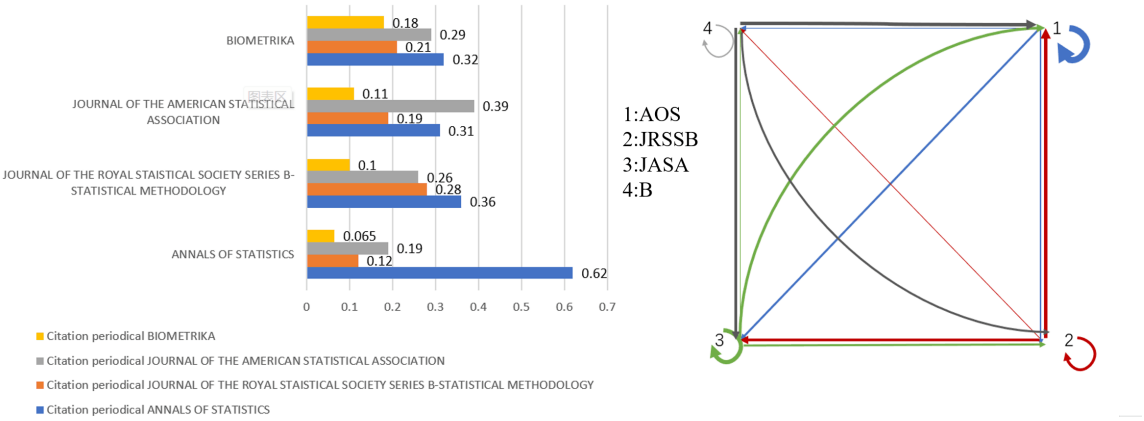

Figure 7: Paper Counts in Four Time Intervals

3.3 Distribution of Papers in Different Time Intervals

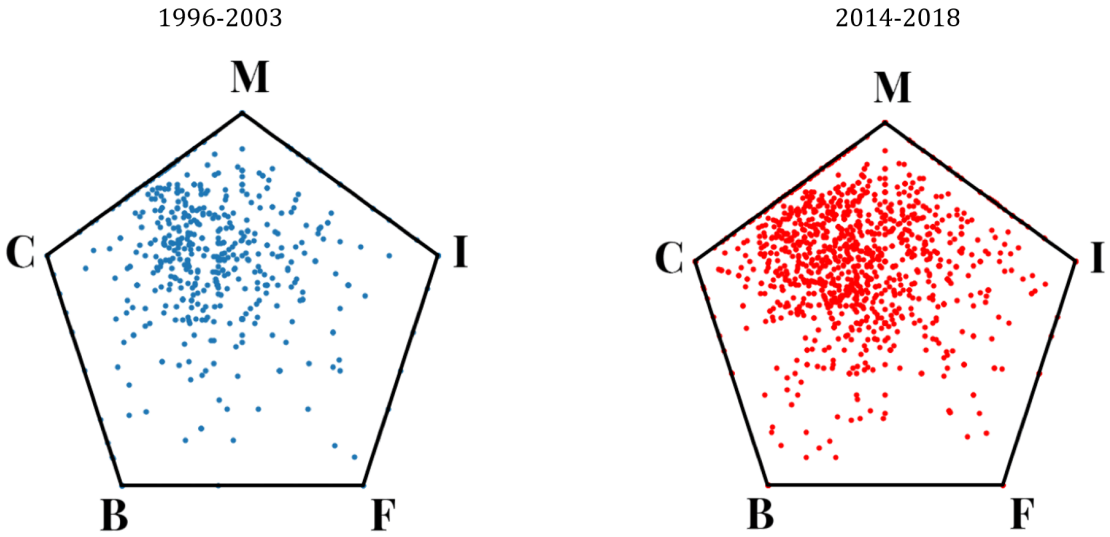

Figure 8: Paper Counts in Four Time Intervals

In examining papers from the earliest and latest time intervals, which are 1996-2003 and 2014-2018 respectively within the dataset, a consistent trend emerged, indicating a persistent focus on the domains of Computer Science and Mathematics and Statistics.

3.4 Month Lag of Papers

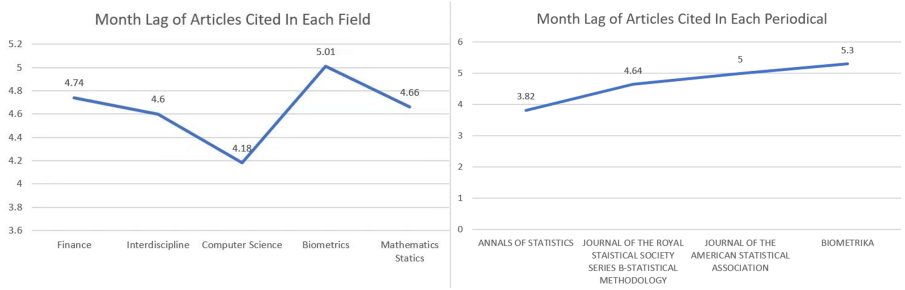

Figure 9: Paper Counts in Four Time Intervals

Figures 9 provide insights into the temporal dynamics of citation patterns at a monthly granularity, specifically examining the average time differences between the publication dates of cited articles and the articles citing them. Time lags in the Biostatistics and Medicine domain is notable, and in Biometrika as well. This may suggest a more prolonged and considered engagement with literature in this field, reflecting the meticulous nature of research in areas related to healthcare and life sciences.

4 Discussion

4.1 Limitations

The classification lacks granularity. Some subjects with a small number of papers are grouped as interdisciplinary, which may lead to the loss of certain details. When considering the position of papers within the pentagon, we only took the level 1 ancestors assigned to each article into account and did not consider information such as titles, abstracts, keywords, etc. Besides, there is a great disparity in the number of articles in each category.

4.2 Future Works

To overcome those flaws, we can continue to expand the dataset to include information from article titles, abstracts, keywords, etc. Natural Language Processing (NLP) methods can be employed for precise localization of the research areas in articles. Additionally, utilizing metrics like cosine similarity could allow for a more detailed categorization of articles.