Citation Network Analysis of the Big-4 Statistics Journals

La Li, Fei Wang, Zhengnan Xu

December 18, 2025

1 Introduction

Citation networks provide a useful way to study how research influence is distributed within an academic field. In statistics, citations reflect not only the popularity of papers but also how ideas and methods build on one another over time. Analyzing citation relationships as a network allows us to move beyond simple citation counts and examine the structural roles that papers and journals play in the development of the field.

In this project, we analyze the citation network formed by papers published in four major statistics journals: Journal of the American Statistical Association (JASA), Biometrika, Journal of the Royal Statistical Society: Series B (JRSS-B), and Annals of Statistics. These journals represent the core of statistical theory and methodology. By focusing only on citations among papers within these journals, we construct a closed citation network that captures influence within the central statistics literature.

Our analysis proceeds in three steps. First, we identify the most cited papers in the network using internal citation counts (in-degree) as a descriptive baseline. Second, we apply the HITS algorithm to distinguish papers that are structurally influential (authorities) from those that primarily cite many important works (hubs), allowing us to construct a network-based "must-read" list. Finally, we aggregate citations at the journal level and examine how the four journals cite each other, revealing systematic citation patterns and asymmetries.

Overall, this project demonstrates how network-based methods can provide additional insight into scholarly influence beyond raw citation counts, and how citation structure differs across papers and journals within the core statistics literature.

2 Data and Scope

2.1 Dataset Overview

This study uses citation data from papers published in four major statistics journals: Journal of the American Statistical Association (JASA), Biometrika, Journal of the Royal Statistical Society: Series B (JRSS-B), and Annals of Statistics. The dataset includes paper-level information and citation links between papers within these journals.

Only citations among papers in these four journals are considered. Therefore, all analyses measure influence within the Big-4 statistics literature, rather than overall citation impact across all academic fields.

The analysis is based on two core tables: a paper-level metadata table (`paper_info`) and a directed citation edge list (`paper_citations`). Each citation is represented as a binary directed edge from the citing paper (`source`) to the cited paper (`target`).

Statistic	Value
Original papers	5,746
Final papers (Big-4)	5,738
Total citations	23,737
Average citations per paper	4.14
JASA papers	2,017
Annals papers	1,733
Biometrika papers	1,238
JRSS-B papers	750

Table 1: Summary statistics of the Big-4 citation network dataset.

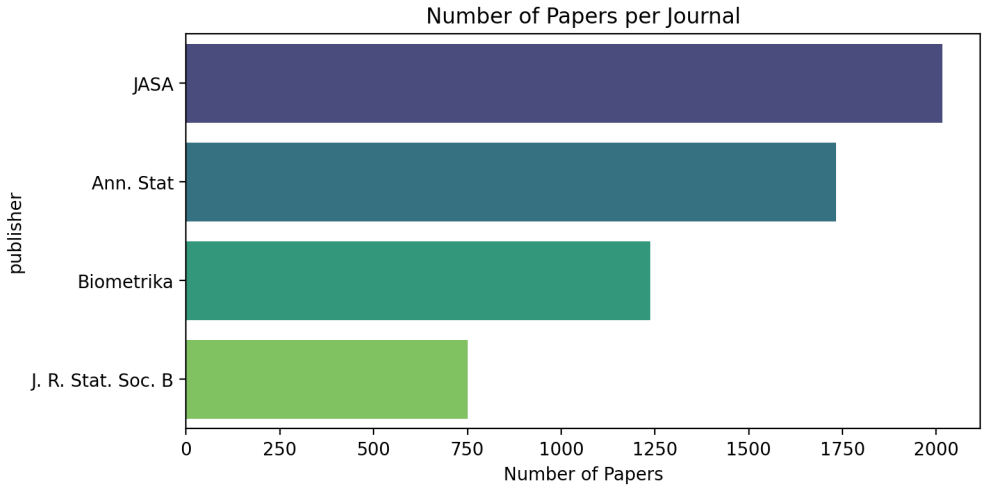
The original dataset contains 5,746 papers, of which 8 papers not published in the Big-4 journals were removed. The final dataset consists of 5,738 papers and 23,737 citation links within the Big-4 citation network. On average, each paper receives 4.14 citations within the dataset. The distribution of papers across journals is unbalanced, with JASA contributing the largest number of papers and JRSS-B the smallest.

2.2 Data Quality and Coverage

3 Exploratory Data Analysis

3.1 Dataset Composition

Bar chart showing the number of papers per journal. This visualization provides basic context on publication volume across the four journals and highlights potential size imbalance that may affect later analyses.

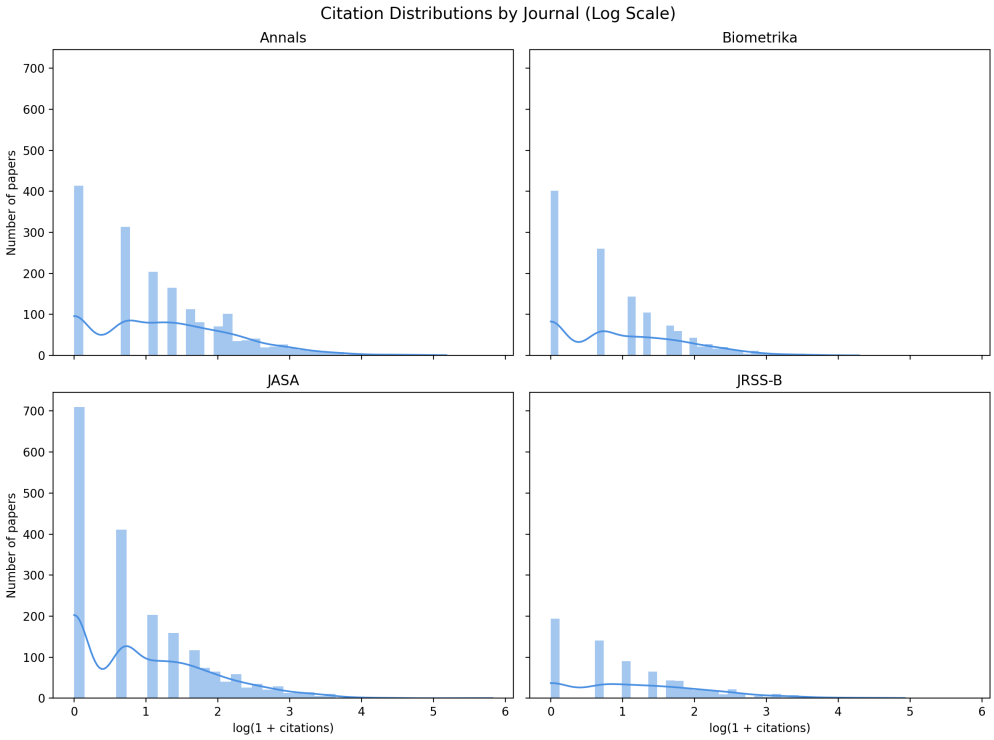


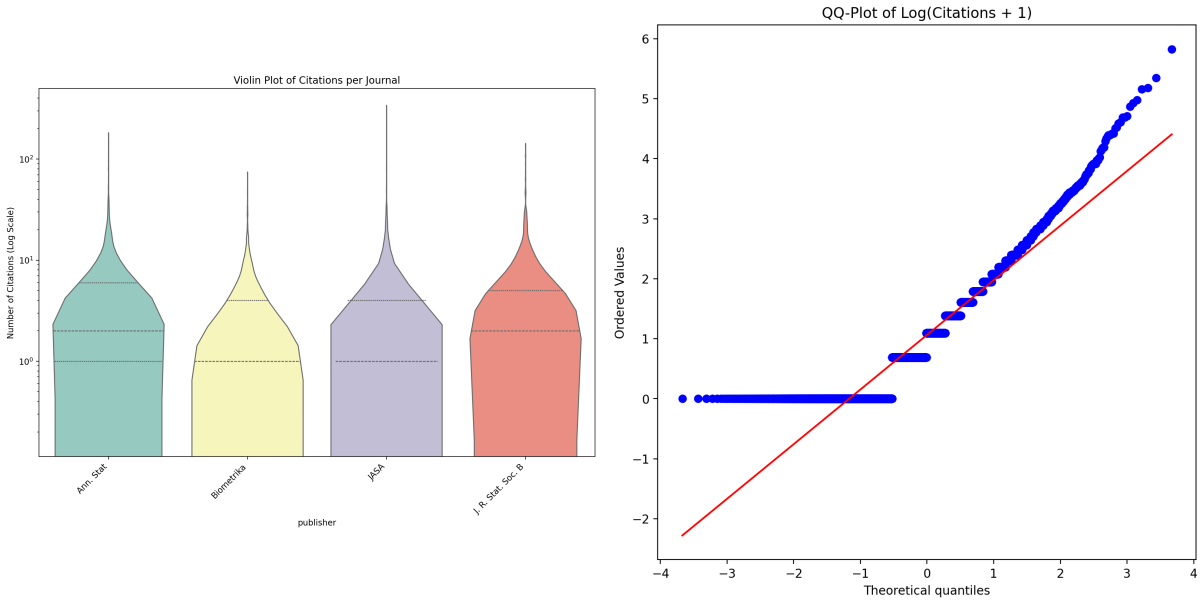
The bar chart shows clear differences in publication volume across the four journals. JASA publishes the largest number of papers, followed by Annals of Statistics and Biometrika, while JRSS-B has a noticeably smaller output. This imbalance indicates that journal size should be taken into account when interpreting citation counts and journal-level statistics in later analyses.

3.2 Citation Distribution and Inequality

Three complementary visualizations examine the overall distribution of citation counts:

- **Citations Histogram**: Displays the distribution of citation counts (in-degree) for all papers, showing how citations are spread across the dataset.
- **Violin plot of citations by journal (log scale)**: Compares citation distributions across the four journals, highlighting differences in spread, medians, and tail behavior rather than only average citations.
- **QQ-plot of log-transformed citations**: Assesses deviation from normality and confirms that citation counts follow a heavy-tailed distribution, justifying the use of nonparametric and network-based methods.



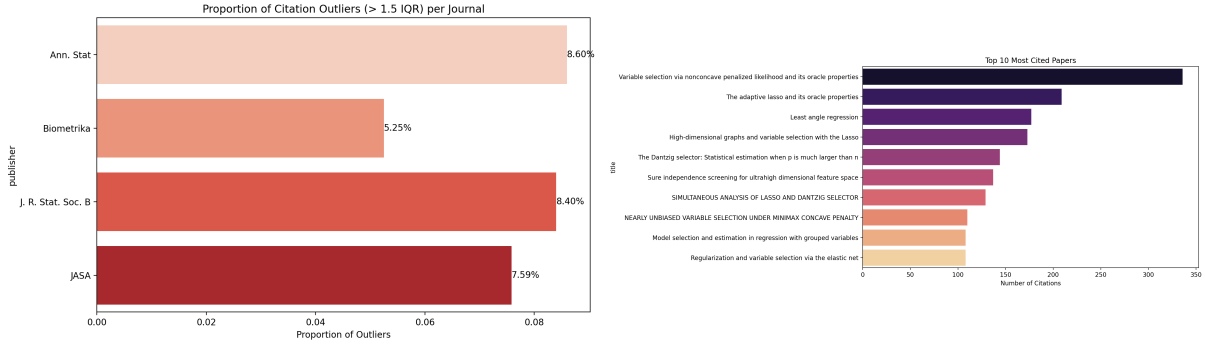


Key Findings: All three visualizations reveal consistent patterns. Citation distributions exhibit strong right-skewness with heavy tails across all journals, though with varying spread and concentration. The QQ-plot confirms substantial departures from normality, particularly a mass at zero and inflated upper tails. These characteristics indicate that citation data are highly non-Gaussian and uneven, justifying the use of nonparametric and network-based methods rather than mean-based assumptions in subsequent analyses.

3.3 Extreme Values and Outliers

Two visualizations focus on extreme citation outcomes:

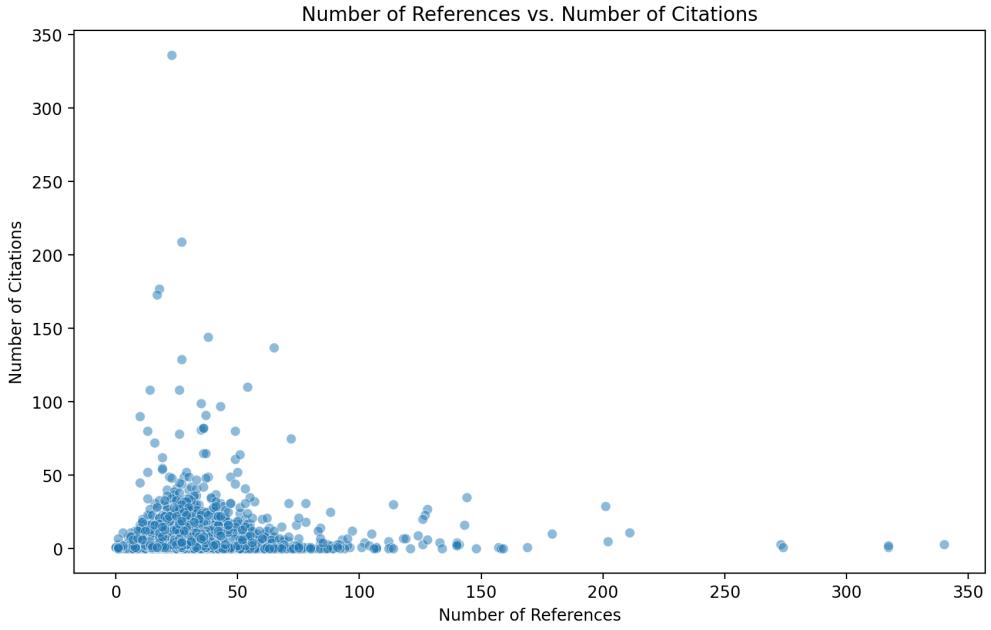
- **Bar chart of citation outlier proportion by journal**: Quantifies how often extremely highly cited papers occur in each journal, revealing journal-level differences in citation concentration.
- **Top-10 most cited papers (bar chart)**: Provides a concrete view of extreme citation outcomes and serves as a bridge from descriptive statistics to later network-based analyses.



Extremely highly cited papers constitute only a small fraction of publications, but their prevalence varies across journals. The top-10 most cited papers illustrate the strong concentration of citations, highlighting the importance of moving beyond descriptive statistics in subsequent network analyses.

3.4 Citation Behavior and Reference Patterns

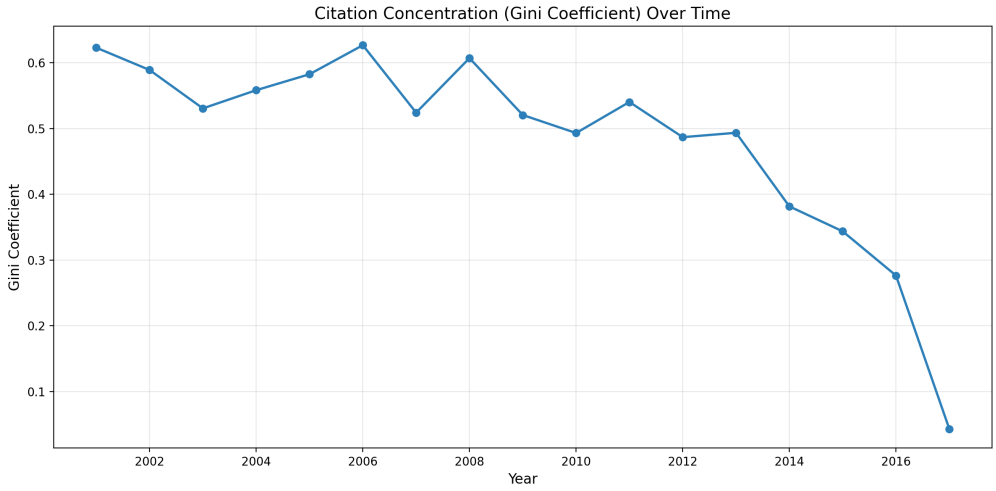
Scatter plot examining the relationship between the number of references in a paper and the number of citations it receives. This explores whether papers that cite more references also tend to accumulate more citations, testing a simple behavioral hypothesis.



Citation impact shows no clear dependence on the number of references, suggesting that extensively citing prior work does not guarantee higher influence.

3.5 Temporal Concentration of Citations

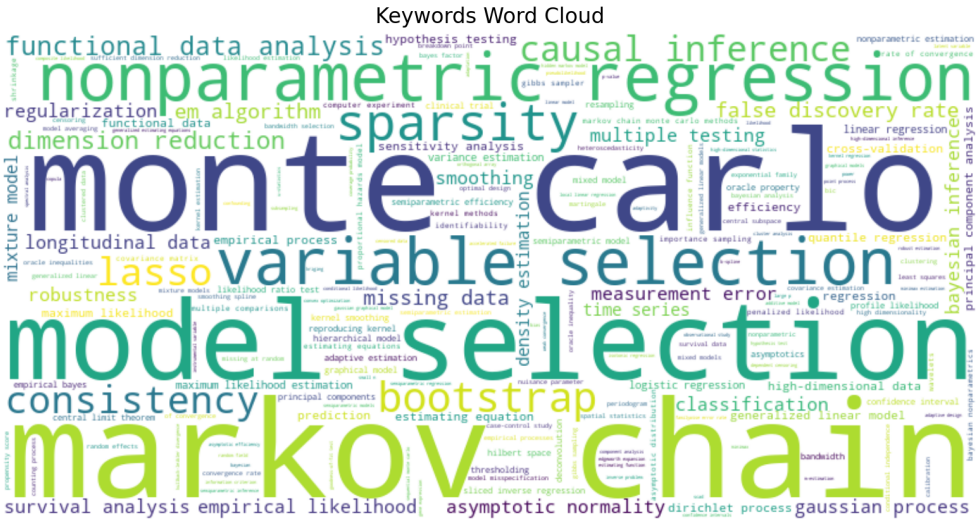
Line plot showing the Gini coefficient of citations over time (where year data is available). This tracks how unequal citation distributions are across years, indicating whether citation concentration has increased or decreased over the study period.



Citation inequality remains pronounced over time, but exhibits a decreasing trend in later years, indicating a possible reduction in citation concentration.

3.6 Content-Level Exploration

Keyword word cloud providing a high-level overview of dominant research themes and terminology in the dataset, without relying on detailed topic modeling. This visualization gives readers quick insight into the dominant research areas represented in the Big-4 statistics literature.



The dataset is dominated by core statistical methodology topics, confirming that the Big-4 journals primarily focus on theoretical and methodological research.

4 Further Analysis

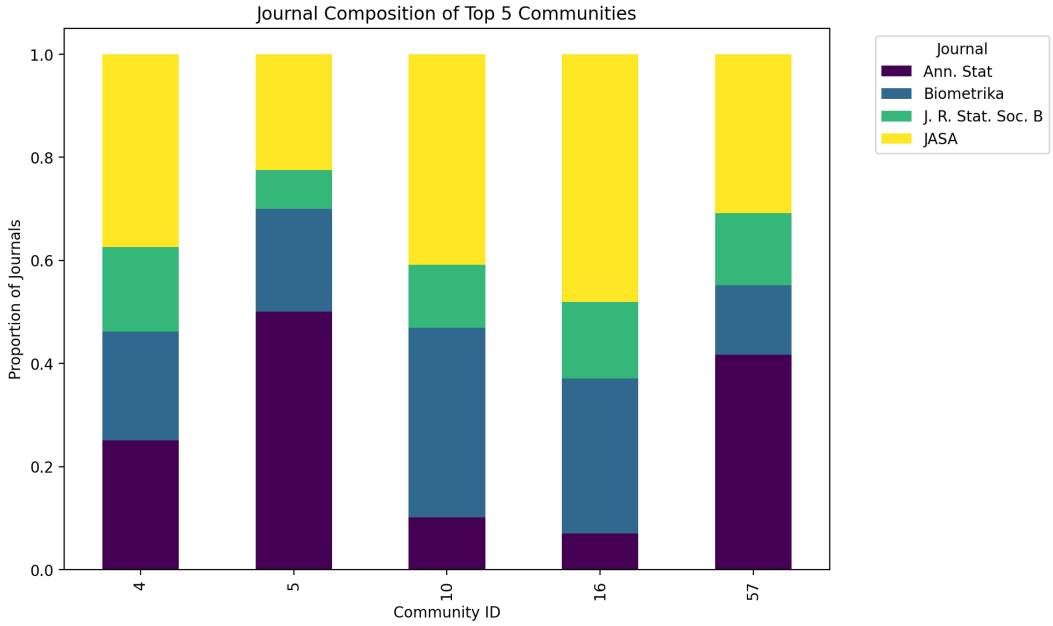
This study focuses on the **topological structure** of the citation network, aiming to reveal deep structures within the statistics community through a static network snapshot. We address four core research directions:

4.1 Community Detection and Topic Structure

Core Question: Does the statistics community exhibit clear divisions into "schools" or sub-fields? Are these communities organized by journal (e.g., theory vs. application) or by topic (e.g., Bayesian methods, high-dimensional inference)?

Methodology: We apply the Louvain or Leiden algorithm to cluster network nodes, validate clustering results against journal labels using Normalized Mutual Information (NMI), and visualize the network with community colors to observe group distributions.

To assess whether detected communities correspond to journal boundaries or cross-journal research topics, we examine the journal composition of the largest communities identified by the Louvain algorithm. This analysis focuses on the top five communities by size.



extbfFigure description: The stacked bar chart shows the proportion of papers from each journal within the five largest detected communities.

extbfInterpretation: All major communities consist of substantial contributions from multiple journals, rather than being dominated by a single outlet. While some communities show a higher presence of Annals of Statistics or JASA, none align cleanly with journal boundaries, indicating strong cross-journal mixing within communities.

4.2 Authority Identification in the Citation Network

Core Question: Who are the foundational pillars of statistics research? Beyond raw citation counts, we seek papers cited by influential works themselves.

Methodology: We use PageRank to identify high-influence nodes and apply the HITS algorithm to distinguish *authorities* (foundational theories such as Lasso, Bootstrap) from *hubs* (high-quality reviews or textbook-style papers), producing a network-based "must-read" list.

To identify foundational works and influential contributors beyond raw citation counts, we analyze authority structure in the Big-4 citation network at both the paper and author levels. By combining local citation neighborhoods, author-level aggregation, and HITS-based authority–hub scores, we provide a multi-scale view of influence within the statistical literature.

Top-10 Paper Citation Network

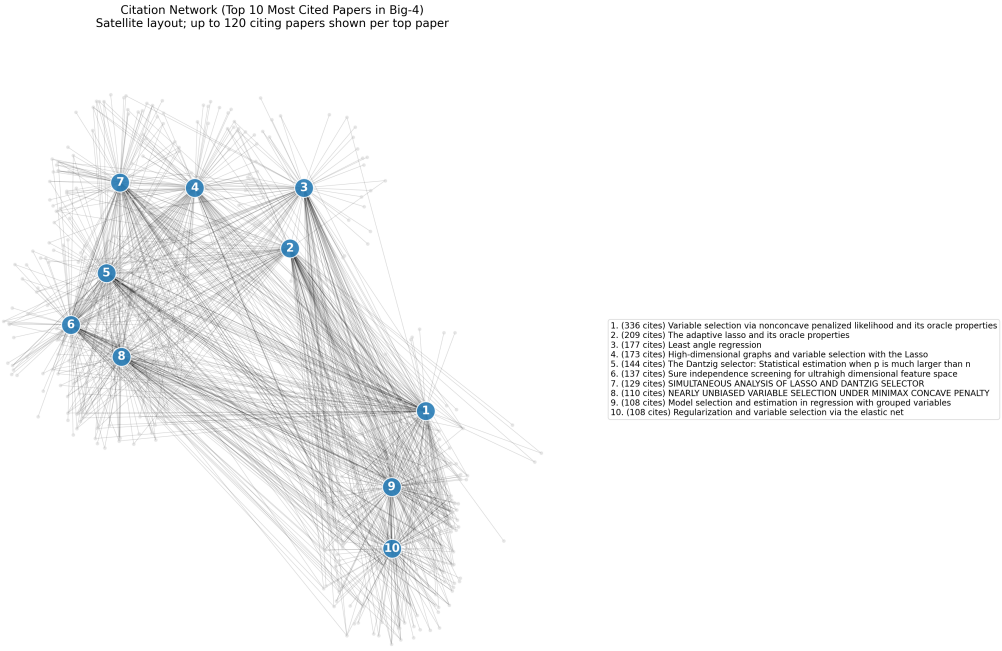


Figure description: The citation subnetwork centered on the top 10 most cited papers illustrates highly concentrated local citation neighborhoods. Each highlighted node represents a top-cited paper, while surrounding nodes correspond to papers that cite it within the Big-4 network.

Interpretation: The visualization reveals a dense and overlapping citation core, indicating that several seminal papers are frequently cited together. This suggests that influence is not isolated to individual works, but concentrated around a small set of canonical contributions that jointly shape the field.

Author–Author Citation Network

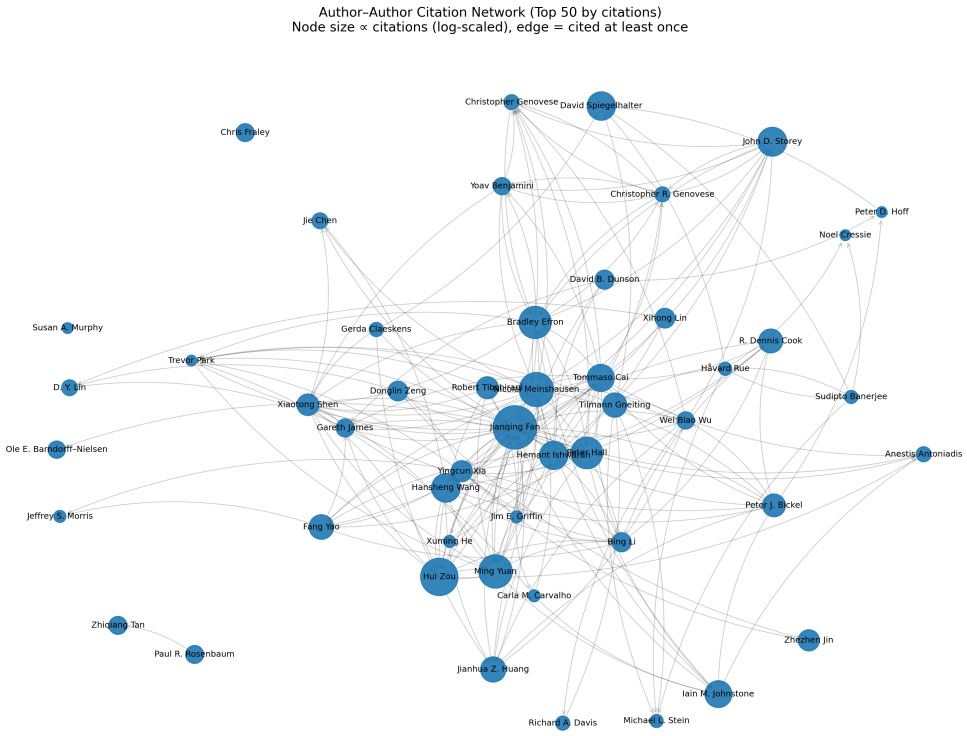


Figure description: The author–author citation network aggregates paper-level citations at the first-author level. Node size reflects total citations received (log-scaled), and edges indicate at least one citation between authors.

Interpretation: The network exhibits a centralized structure, with a group of highly influential authors occupying core positions and maintaining dense citation links. This mirrors the paper-level findings and suggests that authority in the Big-4 network is reinforced at the author level, driven by a relatively stable set of leading contributors.

Authority Score vs. Hub Score

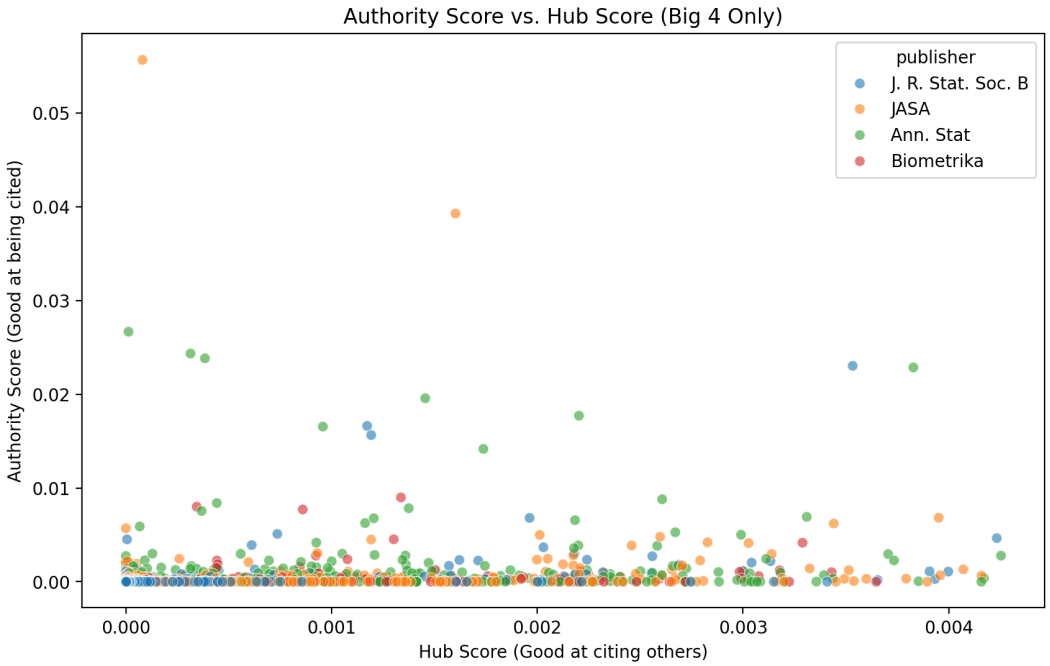


Figure description: The scatter plot compares HITS authority scores and hub scores for papers in the Big-4 network, colored by journal.

Interpretation: Most papers exhibit low scores on both dimensions, while a small number achieve high authority or hub scores. The weak correlation between the two axes indicates a clear distinction between being highly cited (authority) and extensively citing others (hub), supporting the use of network-based measures rather than raw citation counts alone.

4.2 Conclusion

Taken together, these results demonstrate that authority in the Big-4 citation network is highly centralized and multi-layered. Influential papers form a tightly connected core, influential authors occupy persistent central positions, and HITS scores distinguish foundational works from broadly citing connectors. This confirms that structural authority captures aspects of scientific influence not reflected by citation counts alone.

Definition of top-cited: in-degree within the Big-4 closed citation network

4.3 Journal-Level Citation Structure

Core Question: Do citation barriers exist among the Big-4 journals? For example, does the theory-oriented *Annals of Statistics* sit atop the citation hierarchy (highly cited but citing others less frequently)?

Methodology: We construct a 4×4 citation flow matrix (heatmap), perform hypothesis testing to assess whether JASA cites Biometrika significantly more than Annals, and normalize results to control for publication volume bias.

To examine whether citation barriers exist among the Big-4 journals, we analyze citation flows at the journal level. By aggregating paper-level citations into a journal–journal matrix and applying

normalization and statistical testing, we aim to distinguish raw citation volume effects from genuine citation preferences.

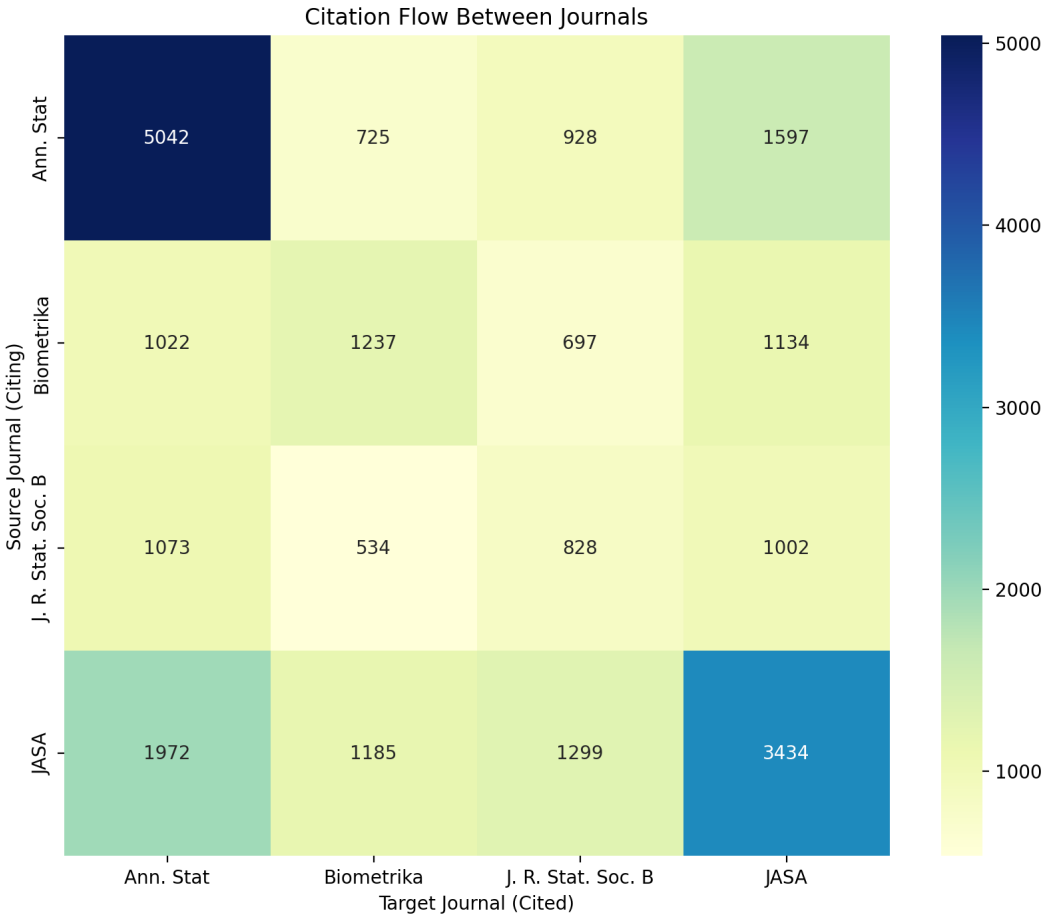


Figure description: The heatmap displays the raw number of citations exchanged between journals, where rows represent citing journals and columns represent cited journals.

Interpretation: Strong diagonal dominance indicates substantial self-citation within each journal, particularly for Annals of Statistics and JASA. Off-diagonal entries show asymmetric citation flows, suggesting potential hierarchy effects, but these raw counts are influenced by differences in publication volume.

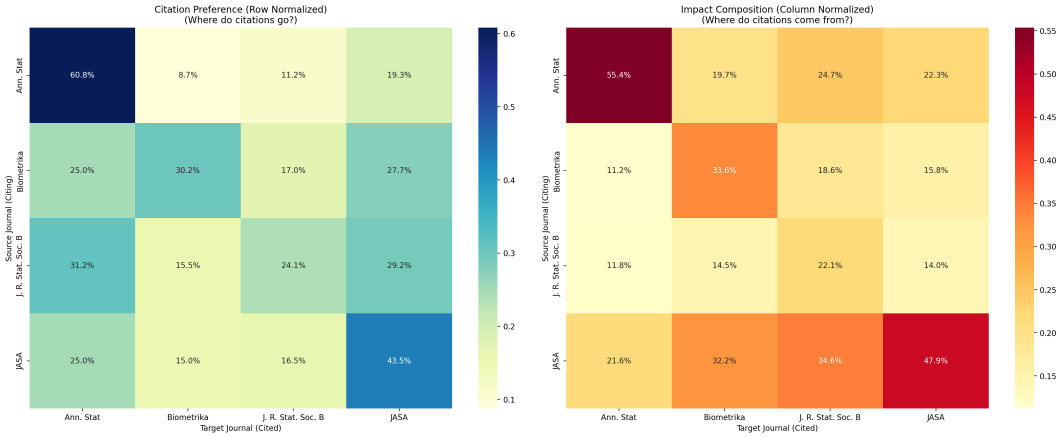


Figure description: The left heatmap normalizes citation flows by row, representing citation preferences of each journal, while the right heatmap normalizes by column, showing the composition of citations received by each journal.

Interpretation: After normalization, distinct citation patterns emerge. Annals of Statistics exhibits a strong inward citation structure, while JASA shows more diversified outgoing citations across journals. This suggests functional differentiation, with some journals acting as theoretical cores and others serving as integrative hubs.

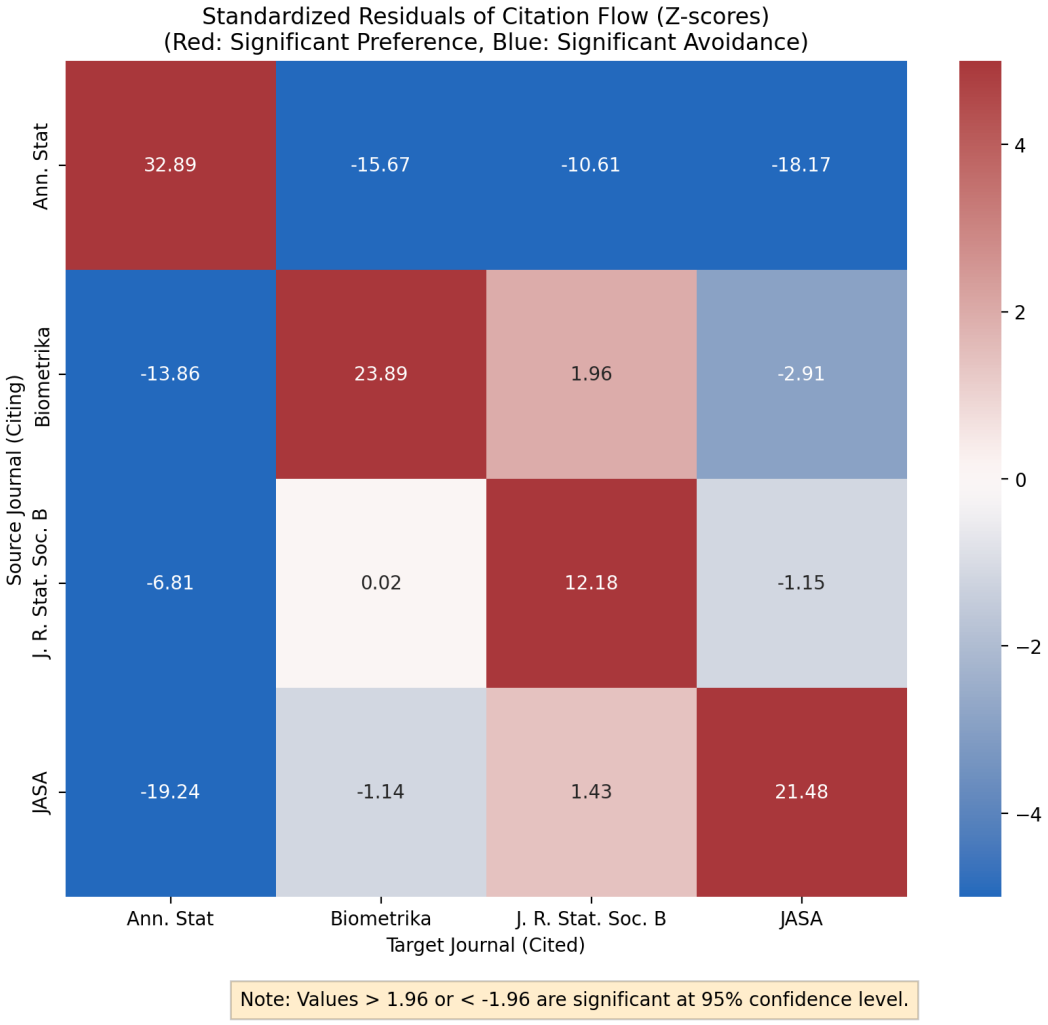


Figure description: The standardized residual heatmap highlights statistically significant deviations from expected citation flows under a null model of independence. Red cells indicate significant over-citation, while blue cells indicate significant under-citation.

Interpretation: Significant positive residuals along the diagonal confirm strong self-citation tendencies beyond what would be expected by chance. Several off-diagonal entries exhibit significant under-citation, providing evidence of citation avoidance between certain journal pairs and supporting the existence of structural citation barriers.

Overall, journal-level citation patterns in the Big-4 network are highly structured rather than random. While raw citation counts are influenced by publication volume, normalized and residual-based analyses reveal persistent self-citation behavior and asymmetric cross-journal interactions.

These results indicate that citation flows reflect functional roles and implicit hierarchies among journals, rather than a fully homogeneous citation ecosystem.

5 Discussion

The analyses reveal a highly skewed citation landscape within the Big-4 statistics journals. Heavy-tailed citation distributions and growing Gini coefficients show that influence is concentrated in a small subset of papers, while the majority receive few citations. Authority analysis reinforces this centralization: a compact core of papers and authors jointly anchors much of the network's structural influence. Community detection indicates that intellectual communities cut across journal boundaries, with mixed journal composition in the largest clusters. At the journal level, strong self-citation and asymmetric cross-journal flows persist even after normalization and residual analysis, suggesting structural citation preferences rather than random variation.

Several limitations should temper interpretation. First, the network is closed to citations within the Big-4, omitting external influence from other venues. Second, citation-based authority captures scholarly attention but not methodological quality or real-world impact. Third, static aggregation may obscure temporal dynamics such as topic shifts or delayed recognition. Finally, visualizations rely on available metadata and may be sensitive to missing or ambiguous identifiers.

Future work could expand the network to include adjacent journals, incorporate semantic embeddings for topic-aware community detection, and model temporal evolution of influence using dynamic network methods. Link-prediction or recommendation-style analyses could highlight potential cross-journal connections and identify under-cited but thematically relevant work. Robustness checks (e.g., alternative centrality metrics, bootstrapped community detection) would further validate the observed structural patterns.

6 Conclusion

This study maps citation structure within the four flagship statistics journals. By combining exploratory analysis with network methods (HITS authorities/hubs, community detection, journal-level flow analysis), we show that influence is highly concentrated, communities are cross-journal rather than journal-bound, and citation flows exhibit strong self-citation with asymmetric cross-journal patterns. These findings highlight both the coherence and the hierarchy of the statistics literature.

The results provide a network-informed view of "must-read" papers and influential authors, while revealing where journals function as theoretical cores versus integrative hubs. Extending the analysis beyond the Big-4, adding text-based topic models, and introducing temporal dynamics would deepen understanding of how statistical ideas diffuse and how citation practices evolve.