Independent Thinkers and Scientific Progress
An Analysis of Superstar Influence on Computer Science Research Dynamics

Filip Plonka
Supervisors: Hayley Hung, Vandana Agarwal, Chenxu Hao
EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Filip Plonka Final project course: CSE3000 Research Project
Thesis committee: Hayley Hung, Vandana Agarwal, Chenxu Hao, Megha Khosla

Abstract

In the scientific community, a few prominent researchers, known as "superstars," receive most of the attention, citations, and resources. However, it is unclear whether they promote true innovation. This study replicates and extends previous work analyzing how superstars influence their collaborators, focusing on the field of computer science. Using the Semantic Scholar Academic Graph dataset, we confirm that while connected researchers in computer science tend to publish more and receive more citations, their ideas are often less innovative. Unlike in the work we replicate, however, we observe this effect even before dissociating the connected researchers from superstars. We also develop a new metric to further clarify the impact of superstars on research diversity. The findings provide insights into the role of superstars in scientific innovation.

1 Introduction

Innovation is key to scientific progress, yet the distribution of attention and resources in academia tends to be unequal, with a small group of prominent researchers, or "superstars," receiving much of it [1]. While these superstars are often highly productive and influential, their impact on the broader research community is not always clear. This work seeks to understand the role of independent thinkers in creating scientific innovation and whether association with superstars helps or hinders this process.

In a paper titled "Don't Follow the Leader: Independent Thinkers Create Scientific Innovation," Sean Kelty et al. [2] explored the effect of superstars on their collaborators in the field of physics. They introduced measures to quantify novelty, innovation, and impact from scholarly citation networks, and their findings indicated that while connected researchers tend to be more productive and receive more citations, they also produce less innovative and more redundant ideas after dissociating from their superstars. On the other hand, independent authors who have not associated with superstars tend to produce more innovative and diverse content.

The idea that independent thinkers can foster innovation aligns with the view that creative processes benefit from diverse perspectives [3]. However, hierarchies in academic collaborations can sometimes inhibit such diversity, a phenomenon observed in academic publishing as the Matthew Effect, where success begets success [4]. While some argue that this effect ensures "epistemic security" by rewarding top contributors proportionally [5], others caution that it may concentrate resources and attention on a few individuals, limiting broader innovation [6]. The exact nature of how these dynamics play out in the field of computer science remains unclear.

To clarify these questions, we replicate the work of Kelty et al. in the field of computer science by using the Semantic Scholar Academic Graph (S2AG) dataset [7]. We also extend this previous work in two ways: first, by also using a different way of representing papers as vector embeddings, to check the findings of Kelty et al. for robustness; and second, by introducing a new metric for diversity, aiming to explore further the impact of superstars on not just innovation but also the diversity of research output.

This paper is structured as follows: We first provide a detailed background on the work of Kelty et al. to provide context for our replication effort. We then describe our methodology, including data collection, preprocessing, and the computation of various metrics. Following this, we present our results, comparing them with those found by Kelty et al. Then we consider the ethical aspects of our work, and finally, we discuss the implications of our findings for the computer science research community and suggest directions for future work.

By replicating and extending the analysis of superstar influence in a different scientific domain, our work contributes to the broader understanding of how influential researchers shape the landscape of academic research. Our findings have potential implications for research policy, collaboration strategies, and the evaluation of scientific impact in computer science.

2 Background

This section provides an overview of the work by Kelty et al: we describe their dataset, key metrics, analyses, and conclusions, focusing on aspects relevant to our replication in the field of computer science. Additionally, we explain our rationale behind how we chose to extend their work.

2.1 Work by Kelty et al.

Kelty et al. conducted a study on the influence of superstar researchers in physics using the American Physical Society (APS) corpus. Their work aimed to understand how these highly influential scientists impact the productivity, visibility, and innovation of their collaborators and the broader research community. The study utilized the APS corpus, which contains articles published in APS journals since 1893. Superstars were defined as authors in the top 0.1% by h-index, resulting in 303 superstars among 292,394 authors.

To quantify novelty and innovation in scientific papers, Kelty et al. used three key metrics. The Shannon Entropy measures the topic diversity in a paper's abstract, with higher values indicating greater diversity. The Reference/Citation Diversity metric quantifies the dispersion of topics in a paper's references or citations. The Innovation metric counts new term combinations introduced by a paper. The precise formulation of these metrics follows in our discussion on methodology in Section 3.3.

Using these metrics, Kelty et al. performed several analyses to understand the impact of superstars. They found that superstars significantly outperformed other academics on all metrics, with particularly striking differences in innovation scores. At the individual level, researchers who frequently cited superstars showed higher metric scores, more citations, and more publications. However, when excluding papers co-authored with superstars, the benefits to the number of citations received and papers published largely disappeared, and innovation actually decreased as an author cited superstars more frequently.

Another important analysis they performed was the comparison of early-career researchers. They defined two groups: "early collaborators" who frequently co-authored with superstars in their first five years, and "early innovators" who had high innovation scores but no superstar collaboration in their first five years. Early collaborators initially had more citations per paper and more publications. However, when excluding papers co-authored with superstars, early innovators published more and received comparable citation rates. Importantly, early innovators maintained higher innovation scores throughout their careers.

From these analyses, Kelty et al. concluded that while collaboration with superstars can boost short-term productivity and visibility, it may not foster long-term independent success or innovation. Their results suggest that researchers who develop innovative ideas independently, without early reliance on superstar collaborations, tend to maintain higher levels of innovation throughout their careers and achieve comparable citation rates when controlling for superstar effects.

2.2 Motivation for Extending Previous Work

While Kelty et al.'s findings are illuminating for the field of physics, the dynamics of academic collaboration and innovation might differ across disciplines. Our motivation to replicate this study in computer science stems from the desire to understand if similar patterns hold in a field characterized by rapid technological advancements and a different collaboration culture.

To ensure the robustness of our findings, we extend the methodology of Kelty et al. by incorporating an additional method for representing papers as vectors. This choice is largely motivated by the specifics of how research diversity metrics are computed and the suitability of cosine similarity for different vector embeddings, with details provided in Section 3.3. Additionally, using another method to represent research papers allows us to check whether our results depend on the specifics of the modeling technique. Consistent results across both representations would indicate that our findings are more robust and likely reflect a real effect, rather than being artifacts of methodological choices. This is especially true given that Kelty et al. do not provide an argument for why their chosen embedding method is particularly appropriate; their choice seems largely motivated by the need for *some* embedding, to be able to calculate the chosen metrics. Absent a reason to favor one embedding type over another, we replicate the relevant results with both metrics to check for robustness.

Additionally, we introduce a new metric for diversity to deepen our understanding of research output. While the key conclusions of Kelty et al.'s mostly concern innovation and citation counts, we aim to balance this by examining diversity more thoroughly. The diversity metric we propose assesses the variance in an author's work by calculating the spread of their papers in a high-dimensional embedding space. This approach has the benefit of capturing diversity at the per-author level, rather than the per-paper level, which is more natural when performing analyses comparing different groups of authors, as is the case in our work. This way, we aim to assess superstar impact on the breadth and variety of research contributions more effectively.

3 Methodology

This section outlines our approach to replicating and extending the work of Kelty et al. We describe our data collection and filtering process, methods for topic modeling and embeddings, computation of key metrics for evaluating research diversity and innovation, and analyses of superstar influence. Throughout, we highlight where our methodology aligns with or deviates from the original study.

3.1 Data Collection and Filtering

We utilized the Semantic Scholar Academic Graph (S2AG) datasets [7], which are part of the Semantic Scholar Open Data Platform [8]. Specifically, we use the *abstracts*, *authors*, *citations*, *embeddings-specter_v2*, and *papers* datasets. The *papers* dataset contains metadata including the field of study for each paper. To restrict ourselves to the field of computer science, we filtered the dataset to include only papers where the majority of the `s2fieldsofstudy` entries indicated Computer Science.

To manage computational resource constraints, we randomly sampled 10% of these papers, resulting in a subset of 718,355 papers. For related datasets (*citations*, *authors*, *abstracts*, and *embeddings-specter_v2*), we filtered entries to include only those related to the selected papers.

3.2 Topic Modeling and Embeddings

Following Kelty et al., to analyze the thematic content of each paper, we employed Latent Dirichlet Allocation (LDA). LDA is an unsupervised machine learning technique that models a corpus of documents as mixtures of topics, where each topic is a distribution over words [9]. This approach allows us to represent each paper by a set of topics. Metrics computed using these topics are then how we quantify features like the diversity of a paper.

Our methods for performing LDA directly replicate the work of Kelty et al. We use the abstracts of the papers in our dataset, first preprocessing by removing stop words and lemmatizing the remaining words to reduce them to their base forms. We use the scikit-learn implementation of LDA using $k = 25$ as the number of topics [10]. This choice of value for k is guided by the use of the UMass coherence measure, which indicates that there are diminishing returns to higher values of k [11]. The output for each document is a 25-dimensional vector v_u, where each element v_u^i represents the frequency of topic i in the abstract of paper u.

In addition to LDA, we also utilized SPECTER embeddings to provide an alternative representation of the research papers. SPECTER embeddings are dense vector representations of paper content generated using a pre-trained SPECTER2 model [12]. Each paper is represented as a 768-dimensional vector v'_u, available as part of the *embeddings-specter_v2* dataset in S2AG. This provides a different perspective on the paper content, as discussed in section 2.2: by using both LDA and SPECTER embeddings, we can compute some metrics using either method, allowing us to compare results and ensure our findings are robust across different embedding techniques.

3.3 Computation of Metrics

We computed four key metrics for each paper to evaluate research diversity and innovation: Shannon Entropy, Reference/Citation Diversity, Innovation, and our proposed Pairwise Diversity. The first three metrics directly follow the definitions used by Kelty et al., while the fourth is a new metric that we propose.

Shannon Entropy Shannon entropy quantifies the diversity of topics within a paper. For a given document u, it is defined as:

$$I_u^{(S)} = -\sum_{i=1}^{k} v_i^u \ln v_i^u$$

Higher values of Shannon entropy indicate greater topic diversity, reflecting the breadth of subjects covered in the paper.

Reference and Citation Diversity Reference and citation diversity measure the variance in topics of references and citations, respectively. These metrics are computed using cosine similarity:

$$I_u^{(X)} = \frac{1}{|X^u|} \sum_{l \in X^u} \left[1 - \cos(\mathbf{v}^l, \overline{X^u})\right],$$

where $\cos(\mathbf{v}^l, \overline{X^u})$ is the cosine similarity between the vector embedding of a reference or citation and the average vector of all references or citations. These metrics evaluate the extent to which a paper draws from diverse sources or inspires a wide range of subsequent research.

Innovation Innovation measures the degree to which a paper introduces new combinations of terms into the literature. It is defined as:

$$I_u^{(I)} = \frac{1}{2} \sum_{w_1 \neq w_2 \in u} \mathcal{I}(w_1, w_2; u),$$

where $\mathcal{I}(w_1, w_2; u)$ is an indicator function that is 1 if terms w_1 and w_2 are first seen together in paper u and 0 otherwise. This metric captures the novelty of conceptual combinations within a paper.

Pairwise Diversity Pairwise Diversity is a new metric we propose to extend the analysis of research diversity. This metric is defined for author a as:

$$D_a = \frac{2}{|X^a|(|X^a| - 1)} \sum_{p,q \in X^a} (1 - \cos(\mathbf{v}^p, \mathbf{v}^q)),$$

where X^a represents the set of papers by author a, $\mathbf{v}^p$ is the embedding of paper p, and $\cos(\mathbf{v}^p, \mathbf{v}^q)$ is the cosine similarity between the embeddings of papers p and q. This metric captures the diversity in a researcher's output by measuring how "spread out" their work is in the high-dimensional vector space used by the embedding, as quantified by average pairwise cosine distance. Unlike the other metrics, which quantify an aspect of a single *paper*, this metric associates each *author* with a measure of how diverse their papers are.

Examining these metric definitions, we see that diversity metrics utilize cosine distance. Applying this with LDA vectors, where each element represents the proportion of a specific topic within a paper, is theoretically less appropriate, as LDA vectors are sparse and optimized for topic modeling rather than similarity measures. In contrast, SPECTER embeddings are dense and specifically designed for capturing semantic similarity in a high-dimensional space, making cosine distance a more theoretically sound choice. This consideration motivates our use of SPECTER embeddings as a complementary approach, as further elaborated in section 2.2.

3.4 Superstar Statistics

Following Kelty et al. in examining the differences in research diversity and innovation between superstars and non-superstars, we computed Shannon Entropy, Reference/Citation Diversity, and Innovation for each paper and Pairwise Diversity for each author. All metrics (except Innovation, which does not involve vector embeddings) were calculated both using LDA and SPECTER.

For each author, we averaged the value of each metric across their papers. Pairwise Diversity did not require averaging as it is already an author-level metric. We then compared these mean values between superstars and non-superstar authors and performed statistical tests to determine the significance of observed differences. Additionally, we analyzed correlations between pairs of metrics computed using LDA to assess if they captured different aspects of a paper's quality. The findings are discussed in detail in Section 4.2.

3.5 Superstar Influence

The most important part of our replication effort concerns assessing the impact of superstars on various dimensions of scientific performance. To this end, we performed two analyses, both directly replicating the methodology of Kelty et al.

In the first, we examined the relationship between how much authors cite superstars and their research output. Authors were grouped into five bins based on the proportion of their papers that cited superstars: 0-20%, 20-40%, 40-60%, 60-80%, or 80-100%. For each author, we found the average metrics values as in Section 3.4, as well as citation counts and publication counts. Then, for each bin, we found the average value of each metric across the authors in that bin. This lets us determine how e.g. Innovation varies between authors who cite superstars less and more often. Finally, we repeated this process, this time excluding papers co-authored with superstars, and visualized the two trends (including and excluding superstar papers) for each metric.

In the second analysis, we focused on early-career researchers. This involved identifying two groups, as defined by Kelty et al.:

- **Early Collaborators**: Researchers who, in their first five years of publishing, collaborated with superstars in at least half of their publications.
- **Early Innovators**: Researchers who, in their first five years of publishing, did not collaborate with or cite superstars but ranked in the top 10% for Innovation.

Then, for each group, we computed the average number of citations received after 1, 2, 3, ..., 40 years, as measured from the time t_0 of first publication. This slightly extends Kelty et al.'s analysis, which only covered up to 25 years. We performed these calculations twice: once including all papers, and once excluding superstar collaborations. Both sets of results were then visualized for comparison.

4 Results and Discussion

This section presents our findings. We first discuss the results of our topic modeling. We then compare metrics between superstar and non-superstar researchers. Next, we examine how citing and collaborating with superstars affects various aspects of research output. We also analyze the career trajectories of early-career scientists in relation to superstar influence. Finally, we address the limitations of our work.

4.1 Topic Modeling

Using Latent Dirichlet Allocation (LDA), we identified 25 distinct topics from the abstracts of computer science papers in our dataset. Table 1 below shows the keywords for some of the identified topics.

Topic	**Keywords**
Topic 0	device, information, method, unit, first, second, user, invention, display, mobile
Topic 1	resource, cloud, computing, attack, application, device, energy, service, user, security
Topic 2	software, test, process, development, system, game, testing, project, tool, agent
Topic 3	security, model, key, protocol, scheme, event, based, simulation, authentication, analysis
...	...
Topic 24	problem robot algorithm task dynamic approach agent policy path environment

Table 1: Keywords for selected topics identified using LDA.

The LDA topic modeling results seem sensible and logical, showing important areas of study in computer science. Each topic includes a specific set of keywords, indicating that the LDA method effectively sorted the main themes in our dataset. The complete list of topics is listed in Appendix A.1.

4.2 Superstar Statistics

We compared research output metrics between superstar and non-superstar authors. Superstars consistently outperformed non-superstars across all metrics, as summarized in Table 2.

Metric	**Our Work (%)**	**Kelty et al. (2023) (%)**
Entropy (LDA)	5%	2%
Entropy (SPECTER)	2%	-
Citation Diversity (LDA)	67%	15%
Citation Diversity (SPECTER)	64%	-
Reference Diversity (LDA)	14%	20%
Reference Diversity (SPECTER)	4%	-
Pairwise Diversity (LDA)	179%	-
Pairwise Diversity (SPECTER)	450%	-
Innovation	135%	900%

Table 2: Summary of differences in mean metrics for superstars vs. non-superstars. The percentages indicate how much higher each metric is for superstars compared to non-superstars.

All differences were statistically significant ($p < 0.01$), with effect sizes ranging from small to medium (Appendix A.3.2). The new pairwise diversity metric showed the largest effect size ($d = 0.761$). Correlation analysis (Appendix A.3.1) revealed low interdependencies between metrics, suggesting they capture distinct aspects of research output. Additionally, given the importance of the Innovation metric in the work of Kelty et al., we examine the changes over time and the distribution of this metric in Appendix A.2.1.

We also plotted the distributions of each metric for superstars and non-superstars, as shown in Appendix A.2. This visualization reveals that for the diversity metrics, much of the difference between superstars and non-superstars is driven by superstars more rarely having a "trivial" diversity score of zero. We theorize that this is due to decisions made in initial data preprocessing, as elaborated in our discussion of the limitations of our work in Section 4.5.

Our results broadly align with those of Kelty et al: we find a small difference in mean

Entropy, a moderate difference in measures of diversity, and a large difference in Innovation. The most notable discrepancy is that while the difference we find in Innovation is large, it's not as extreme as the ten-fold difference that Kelty et al. find. This suggests that innovation in computer science is less concentrated among the few top researchers than in physics. For Pairwise Diversity, the large differences observed suggest that our proposed metric has potential as a robust indicator of academic potential.

4.3 Effect of Superstar Influence

We examined how citing superstars affects various metrics of research output. Authors were grouped into five bins based on the proportion of their papers citing superstars (0-20%, 20-40%, 40-60%, 60-80%, 80-100%). For each metric, we plotted the average value across these bins, both including and excluding papers co-authored with superstars. The results for Entropy, Citation Diversity, Reference Diversity, and Pairwise Diversity are depicted in Figure 1.

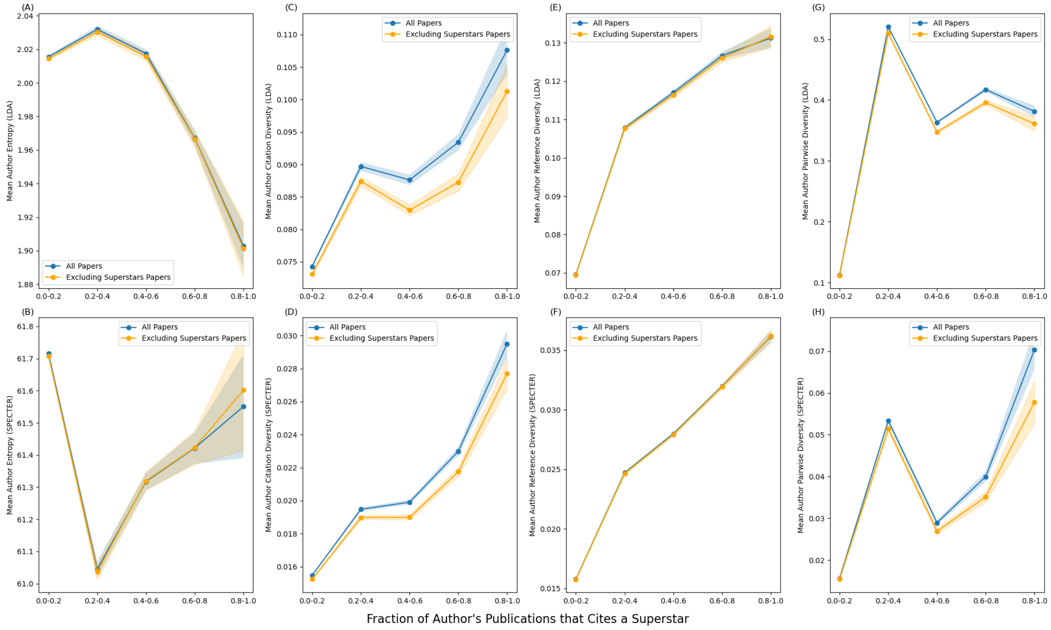


Figure 1: Averages of the novelty and innovation metrics across authors as a function of the fraction of their papers that cite superstars. The blue trend lines include all papers, while the orange ones exclude superstar collaborations. The metrics are, from left to right: Entropy, Citation Diversity, Reference Diversity, and Pairwise Diversity. For each metric, the upper plot was computed using LDA, and the bottom plot with SPECTER. Shaded areas represent ±1 standard error of the mean, indicating the uncertainty in the estimated averages.

Our results show that citation and reference diversity increase significantly as authors cite superstars more frequently. Pairwise diversity peaks in the 20-40% bin when using LDA, then plateaus, while it shows a tentative increase with SPECTER. Entropy decreases with LDA but increases with SPECTER as authors cite superstars more often. The largest difference between all-paper and excluding-superstar-papers trends is observed for citation diversity with LDA, but otherwise, the differences are small across metrics.

These findings largely align with Kelty et al.'s results, who reported increasing Reference Diversity, Citation Diversity, and Entropy, with small differences between the two trends. Our results confirm the positive association between citing superstars and research diversity in computer science, although they show an unclear trend for Entropy, which trends differently depending on the embedding method used.

A case could be made that the trend for Entropy computed using LDA topic vectors is more meaningful, as a high entropy for such a vector means a more even spread across the topics in a paper's abstract and as such a more "surprising" paper, while it's not clear that the entropy of a dense semantic embedding is similarly meaningful. However, for Entropy our trend is opposite in direction to that found by Kelty et al., and in both cases the differences between the 0-20% and 80-100% bins are small. As such, we hesitate to draw conclusions about the effect of superstar influence on the ability of authors to produce work that is surprising or covers diverse topics.

The trends for Innovation, Citation Count, and Publication Count are shown in Figure 2.

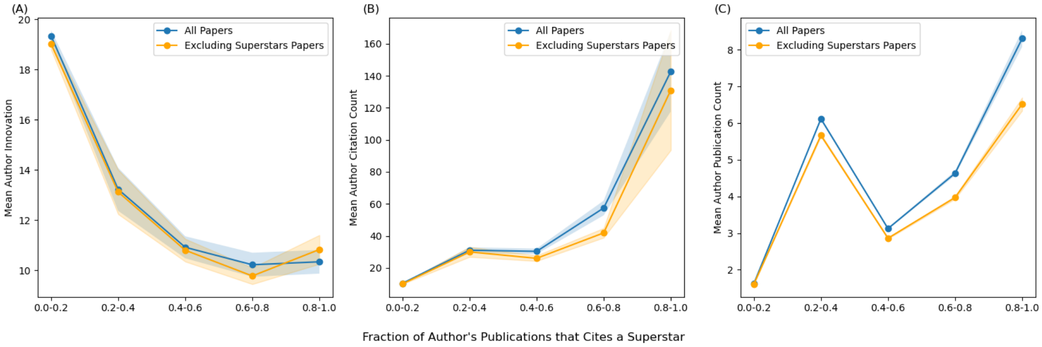


Figure 2: Averages of the metrics across authors as a function of the fraction of their papers that cite superstars. From left to right: Innovation, Citation Count, Publication Cdount.

We observe a clear decrease in innovation as authors cite superstars more frequently, with approximately a 50% reduction in the 80-100% bin compared to the 0-20% bin. Conversely, we observe a substantial increase in citations received, with more than a tenfold difference between the highest and lowest bins. Publication count also increases, showing about a fourfold difference between the extreme bins, with a notable spike in the 20-40% bin. Interestingly, the trends excluding superstar papers closely mirror those including all papers across all metrics.

These results partially diverge from Kelty et al.'s findings. While they reported flat innovation for all papers but a decreasing trend when excluding superstar papers, we observe a consistent decrease. Our citation and publication count trends align with their all-papers results but contrast with their findings when excluding superstar papers. We find a much weaker impact of excluding superstar papers across these metrics.

As such, for these three metrics, we are only partially able to replicate the results of Kelty et al. The most important trends we observe are a clear benefit in citation count to citing superstars often, as well as the negative impact of doing so on Innovation. As we do not see a strong effect of removing papers collaborated with superstars, it appears that these effects aren't strongly driven by direct collaboration. Rather, our results point to distinct patterns among researchers in computer science. Those who frequently cite superstars may be working in high-profile areas of the field, which could lead to increased visibility and productivity. However, this focus on prominent research directions may come at the cost of originality. This suggests that the observed trends may not solely reflect the direct influence of superstars, but rather indicate a broader phenomenon where some researchers engage more with mainstream, highly cited work. While this approach may yield higher citation counts and publication rates, it appears to correlate with lower levels of innovation.

While best efforts were made towards producing a direct replication, it is also possible that some of the discrepancies between our results and those of Kelty et al. arise due to subtle methodological differences rather than different research dynamics in computer science as compared to physics. As such, we believe that considering such an analysis more carefully is a promising direction for future research.

4.4 Superstar Collaboration and Early Career Scientists

We compared the career trajectories of two groups of early-career researchers: "early collaborators" who frequently collaborated with superstars in their first five years, and "early innovators" who produced highly innovative work without citing or collaborating with superstars. For each group, we plotted the average number of citations received after 1, 2, 3, ..., 40 years from the time of first publication, and then repeated this process again but excluding superstar collaborations, as described in Section 3.5. The results are shown in Figure 3.

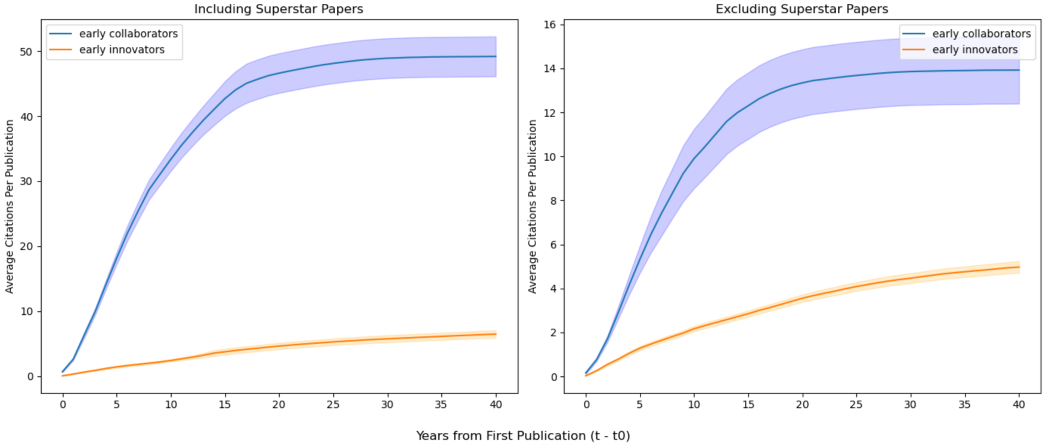


Figure 3: Citations for early collaborators and early innovators, calculated at 1, 2, 3, ..., 40 years from the time of first publication. The blue trend line is for early collaborators, while the orange is for early innovators. The left subplot shows the trend when including all author papers, and the right when excluding superstar collaborations.

The data show that early collaborators receive significantly more citations than early innovators over time when including all papers, with more than a seven-fold difference after

40 years. When excluding superstar collaborations, both groups receive fewer citations, but this effect is more pronounced for early collaborators, reducing the gap to less than a three-fold difference after 40 years.

The main difference between our results and those of Kelty et al. is that their study showed a smaller gap between early collaborators and early innovators in citation counts. Early collaborators received only slightly more citations than early innovators when including all papers, and early innovators actually had more citations when excluding superstar papers. This discrepancy is likely related to our findings in Section 4.2, where we observed that superstars in computer science do not show as large a difference in Innovation as those in physics. Given that observation, it's not surprising that early innovators "underperform" in this analysis. However, the critical effect replicated: excluding superstar papers shifts the citation advantage toward early innovators.

It is important to note that our cutoffs for defining superstars and early innovators are somewhat arbitrary: the top 0.1% in h-index for superstars and the top 10% in Innovation for early innovators. These thresholds could in principle be adjusted to balance the citation trends when including all papers, suggesting that the coinciding of these trend lines in the analysis Kelty et al. is not particularly meaningful.

Another observation worth noting is that early innovators' citation counts plateau more slowly over time: this is what prompted us to extend the analysis to 40 years, compared to the 25 used by Kelty et al. While this could suggest that early innovators produce more durably useful work, it is more likely a consequence of selecting for innovation also selecting for authors who published earlier. This trend is due to the observed decrease in average paper Innovation per year (see Appendix A.2.1): it is increasingly difficult to produce novel keyword combinations over time, as more of them have already been encountered.

Overall, while our results largely replicate the key findings of Kelty et al., some differences in trends were observed. These differences may reflect distinct disciplinary norms or methodological variations. Nevertheless, our findings underscore the importance of supporting early-career researchers who demonstrate independent innovation, as their work can achieve significant impact without reliance on superstar collaborations. This suggests that fostering a diverse range of early-career paths may be beneficial for long-term innovation and citation success in computer science.

4.5 Limitations of Our Work

While our study provides valuable insights, several limitations must be acknowledged.

First, our data preprocessing approach, involving a random sampling of 10% of computer science papers, introduced potential biases. This method resulted in many authors having a small subset of their work included, which could skew the representation of their research output: in our computed subset, authors on average have less than two corresponding papers each. This means that when computing for an author the percentage of their papers that cite a superstar, as in Section 4.3, we often get "round" results like 100%, 50%, 33%, 67%, 25%, etc. So when an author, for example, falls in the 60-80% bin, this often means they cited superstars in 2 of their 3 papers, and thus are an author with an above number of publications for our dataset. This effect causes citation count to impact the binning, which slightly biases our results, and could explain some of the spikes in the 20-40% bin that we observe. Additionally, authors who only have one associated paper trivially have a Pairwise Diversity of zero, and much of the difference between superstars and non-superstars in this metric seems to be driven by the amount of researchers with this trivial diversity score (see

Appendix A.2), suggesting this metric is confounded by publication count.

A preprocessing approach that would have been superior in hindsight is to select a smaller subset of *authors*, and then for each author include data about all of their computer science papers. The trade-off is that then for some papers we are missing data about some of their authors, but this turns out to be less important for our analyses, which much more often consider all the papers of a given researcher. This consideration can inform future research by suggesting that a focus on complete author datasets may be important for obtaining unbiased results.

Second, our replication effort does not encompass all analyses performed by Kelty et al. We focused on the most critical aspects related to their conclusions, such as the examination of diversity and innovation metrics, but did not replicate their redundancy metric or their analysis at the group level (see Section II.D in their work) [2]. Additionally, while we examined citation counts over time for early-career researchers, Kelty et al. also analyzed Innovation over time, which we did not. However, the results they find show a rather predictable advantage for early innovators, as they are by definition selected for high Innovation.

Third, our proposed Pairwise Diversity metric, while showing potential, requires further exploration. The large differences observed for superstars indicate its promise as an indicator of academic potential. However, the results in Section 4.3 do not provide a clear trend, potentially due to the preprocessing issues mentioned earlier. Future work could refine this metric and assess its robustness across different datasets and disciplines.

Fourth, while we aimed to follow Kelty et al.'s methodology closely, some discrepancies may arise from differences in implementation. For example, the exact method for computing Innovation involved point-wise mutual information to filter out spurious term combinations. Although we applied this technique, the lack of precise details in Kelty et al.'s description could lead to some variations in our results. For this reason, we make our source public at https://github.com/fplonka/cse3000, enabling other researchers to verify and build upon our work.

5 Responsible Research

In conducting this study, we carefully considered several aspects of responsible research practice. This section outlines our approach to data privacy, reproducibility, and the broader implications of our work.

Our dataset contains author names, but we deliberately chose not to use or disclose this information in our analysis or results, aligning with ethical guidelines for protecting individual privacy in research [13]. We aggregated data and presented results at a group level to minimize the risk of identifying specific researchers. However, it's important to acknowledge that even anonymized data can potentially be de-anonymized through various techniques [14].

To promote reproducibility, we provide detailed descriptions of our methodology and have made our source code publicly available at https://github.com/fplonka/cse3000. This allows other researchers to verify our results and improve on our work. We also openly discuss the limitations of our study, including potential biases introduced by our sampling method. These efforts align with growing calls in the scientific community for increased transparency in research [15].

We caution against overgeneralization of our findings. Our study examines correlations and trends, not causal relationships. Factors beyond superstar influence likely play significant roles in academic success [16]. We encourage policymakers and institutions to consider our findings as part of a broader approach to supporting diverse research paths. Furthermore, we acknowledge the limitations of metrics like citation counts and h-index as measures of academic impact [17]. We encourage future work to explore additional or alternative measures of research impact and innovation.

By addressing these aspects, we aim to contribute not only to the understanding of superstar influence in computer science but also to the ongoing dialogue about ethical and impactful research practices in academia.

6 Conclusions and Future Work

Our study replicates and extends the work of Kelty et al., investigating the influence of superstars in computer science. Our findings align with Kelty et al. in showing that superstars outperform non-superstars across several metrics, notably innovation and diversity. However, the smaller differences in innovation suggest that computer science might have less concentrated innovation compared to physics.

We find that frequent citation of superstars correlates with higher citation and publication counts but lower innovation. This trend persists even when excluding papers co-authored with superstars, indicating that researchers engaging more with highly cited, mainstream work might do so at the cost of originality. This highlights the need for researchers to explore diverse ideas instead of just following top-cited work.

Our analysis of early-career researchers reveals that those who frequently collaborate with superstars initially receive more citations but are more affected when superstar papers are excluded. In contrast, early innovators who do not rely on superstar collaborations are less impacted in terms of citation counts when these papers are excluded. This suggests that policies supporting independent early-career researchers could enhance long-term innovation in computer science by reducing dependency on superstar collaborations.

Future research should address the limitations of our study, such as the potential biases introduced by data sampling. Our proposed Pairwise Diversity metric shows promise but requires further exploration and validation. Additionally, examining the influence of superstars across other scientific disciplines would help generalize our findings and inform research policies aimed at promoting broad-based innovation. Refining data preprocessing methods, such as selecting complete datasets for specific authors, could also improve the accuracy of future studies.

Overall, our study contributes to understanding how superstars shape the academic landscape in computer science. We successfully replicated many of Kelty et al.'s findings, confirming the role of superstars in driving citations and publication counts, but also highlighting the trade-off with innovation. Our use of SPECTER embeddings proved useful as a robustness check, providing a more theoretically sound basis for diversity metrics. However, our Pairwise Diversity metric did not provide conclusive insights and warrants further investigation in future studies. Our results indicate that academia may benefit from policies that balance supporting superstars with encouraging independent, innovative research paths. This approach could foster a more diverse and innovative research environment, ultimately enhancing the long-term impact and progress of the field.

References

[1] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569â16572, November 2005.

[2] Sean Kelty, Raiyan Abdul Baten, Adiba Mahbub Proma, Ehsan Hoque, Johan Bollen, and Gourab Ghoshal. Don't follow the leader: Independent thinkers create scientific innovation, 2023.

[3] Robin J. Ely and David A. Thomas. Cultural diversity at work: The effects of diversity perspectives on work group processes and outcomes. *Administrative Science Quarterly*, 46(2):229–273, 2001.

[4] Robert K. Merton. The matthew effect in science. *Science*, 159(3810):56–63, 1968.

[5] Weihua Li, Tomaso Aste, Fabio Caccioli, and Giacomo Livan. Early coauthorship with top scientists predicts success in academic careers. *Nature Communications*, 10:5170, 11 2019.

[6] Pierre Azoulay, Joshua Graff Zivin, and Jialan Wang. Superstar extinction. *The Quarterly Journal of Economics*, 125(2):549–589, 2010.

[7] Alex D. Wade. The semantic scholar academic graph (s2ag). In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 739, New York, NY, USA, 2022. Association for Computing Machinery.

[8] Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. The semantic scholar open data platform. *ArXiv*, abs/2301.10140, 2023.

[9] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. volume 3, pages 601–608, 01 2001.

[10] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[11] David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In Regina Barzilay and Mark Johnson, editors, *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics.

[12] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. Specter: Document-level representation learning using citation-informed transformers, 2020.

[13] Michael Zimmer. "but the data is already public": On the ethics of research in facebook. *Ethics and Information Technology*, 12:313–325, 12 2010.

[14] Arvind Narayanan and Vitaly Shmatikov. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, 2008.

[15] Marcus R. Munafo, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1), January 2017.

[16] Samuel F. Way, Allison C. Morgan, Aaron Clauset, and Daniel B. Larremore. The misleading narrative of the canonical faculty productivity trajectory. *Proceedings of the National Academy of Sciences*, 114(44), October 2017.

[17] Diana Hicks, Paul Wouters, Ludo Waltman, Sarah de Rijcke, and Ismael Rafols. Bibliometrics: The leiden manifesto for research metrics. *Nature*, 520(7548):429â431, April 2015.

A Appendix

A.1 LDA Topics

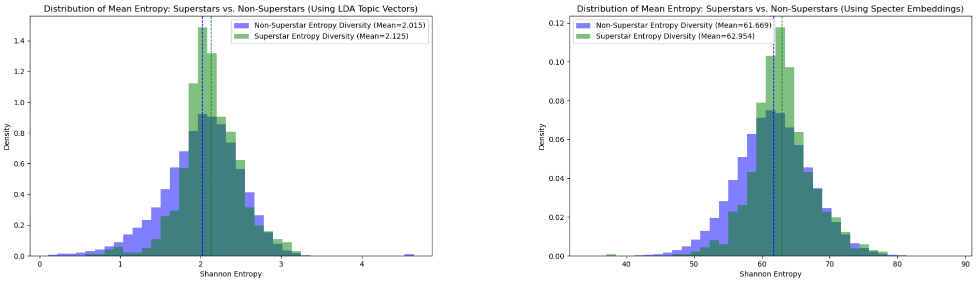
The table below lists all 25 topics identified using Latent Dirichlet Allocation (LDA) from the abstracts of computer science papers in our dataset. Each topic is represented by the top 10 keywords associated with it.

Topic	**Keywords**
Topic 0	device, information, method, unit, first, second, user, invention, display, mobile
Topic 1	resource, cloud, computing, attack, application, device, energy, service, user, security
Topic 2	software, test, process, development, system, game, testing, project, tool, agent
Topic 3	security, model, key, protocol, scheme, event, based, simulation, authentication, analysis
Topic 4	network, node, traffic, wireless, scheme, protocol, performance, routing, channel, packet
Topic 5	method, based, set, algorithm, result, rule, clustering, approach, proposed, pattern
Topic 6	technology, computer, application, research, development, digital, new, library, paper, information
Topic 7	user, interaction, human, interface, virtual, visual, design, environment, 3d, interactive
Topic 8	module, communication, channel, control, information, base, data, station, system, terminal
Topic 9	system, control, based, time, controller, real, paper, design, performance, using
Topic 10	video, graph, frame, temporal, cache, sequence, view, kernel, representation, action
Topic 11	user, information, web, database, query, content, search, social, data, based
Topic 12	service, model, based, application, process, design, approach, framework, domain, paper
Topic 13	method, proposed, detection, based, algorithm, result, estimation, image, noise, using
Topic 14	data, technique, large, analysis, time, mining, processing, set, application, big
Topic 15	feature, recognition, image, object, classification, face, method, detection, using, based
Topic 16	server, network, file, vehicle, client, storage, management, request, peer, connection
Topic 17	time, error, rate, problem, number, algorithm, scheduling, result, delay, probability
Topic 18	learning, study, student, research, knowledge, question, result, based, evaluation, online
Topic 19	signal, power, design, performance, frequency, circuit, architecture, high, system, speech
Topic 20	language, code, program, text, word, programming, document, source, translation, type
Topic 21	model, network, learning, neural, training, method, task, performance, prediction, deep
Topic 22	image, memory, block, coding, compression, color, bit, quality, processing, resolution
Topic 23	algorithm, method, function, proposed, optimization, problem, based, paper, result, new
Topic 24	problem, robot, algorithm, task, dynamic, approach, agent, policy, path, environment

Table 3: List of LDA topics identified from the abstracts of computer science papers. Each topic is represented by the top 10 keywords associated with it.

A.2 Metric Distributions

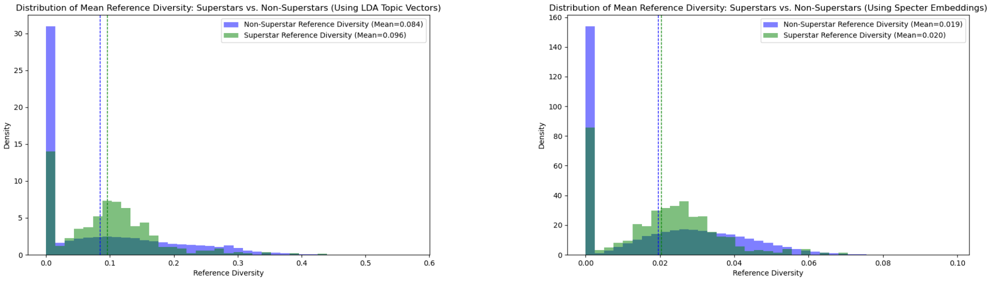
In this section, we provide additional plots for the metrics discussed in our results (Section 4.2). These plots include plots for metrics computed using both LDA and SPECTER embeddings.



(a) LDA Entropy

(b) SPECTER Entropy

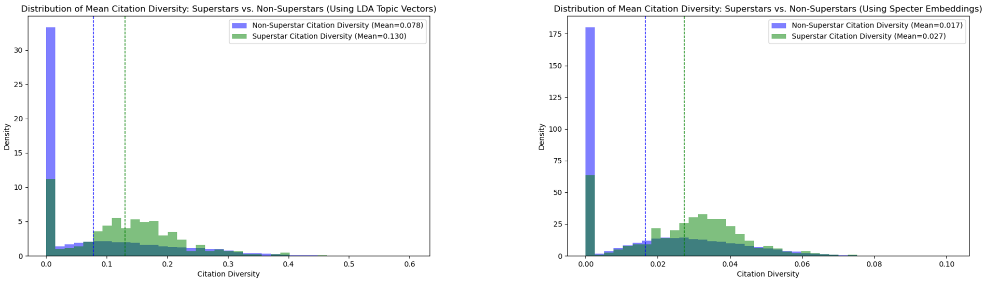
Figure 4: Distribution of mean Entropy across authors: superstars vs. non-superstars.



(a) LDA Reference Diversity

(b) SPECTER Reference Diversity

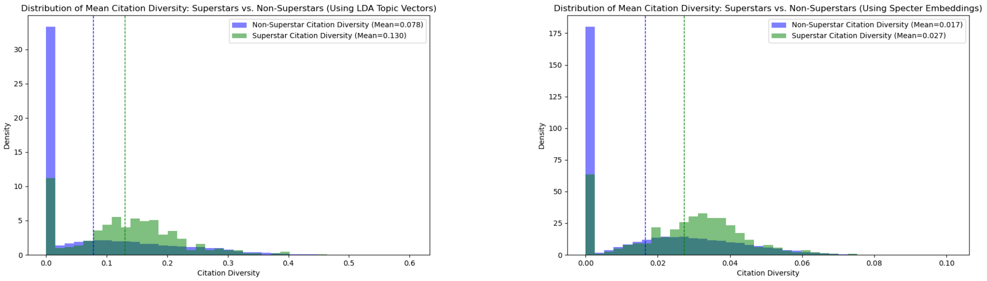
Figure 5: Distribution of mean Reference Diversity across authors: superstars vs. non-superstars.



(a) LDA Citation Diversity

(b) SPECTER Citation Diversity

Figure 6: Distribution of mean Citation Diversity across authors: superstars vs. non-superstars.



(a) LDA Pairwise Diversity

(b) SPECTER Pairwise Diversity

Figure 7: Distribution of mean Pairwise Diversity across authors: superstars vs. non-superstars.

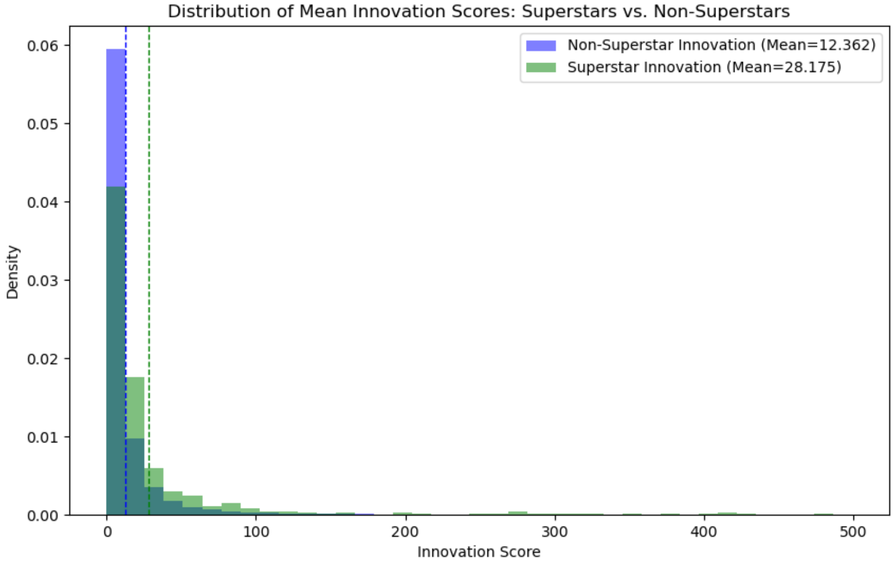


Figure 8: Distribution of mean innovation scores across authors: superstars vs. non-superstars.

A.2.1 Changes in Innovation Over Time

Of the metrics used in the work of Kelty et al., superstars showed the largest difference with non-superstars in innovation. This prompts a closer examination of trends in this metric. Figure 9 show the mean innovation score per year and the distribution of innovation scores across papers.

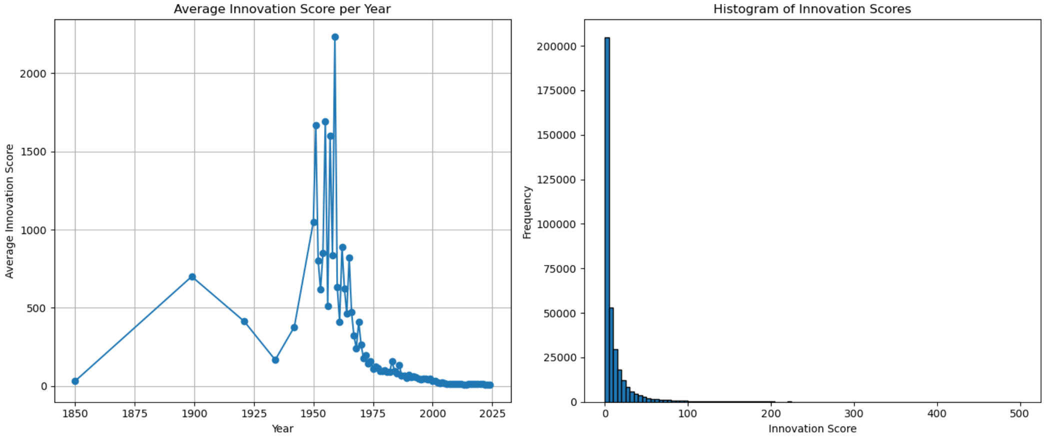


Figure 9: (Left) Mean paper innovation per year. (Right) Distribution of innovation scores across papers.

The left plot demonstrates a decrease in mean innovation scores over time, reflecting the increasing difficulty of introducing new keyword pairs as the field matures. The right plot shows a power law distribution, with most papers exhibiting low innovation and a few papers showing high innovation.

A.3 Statistical Analysis Results

This section presents the results of the statistical analyses conducted on the dataset, including pairwise correlations of metrics derived from papers and comparative statistics between superstar and non-superstar authors across various metrics.

A.3.1 Pairwise Correlations of Metrics

The correlations between various metrics computed from the papers are presented in the table below. Where relevant the LDA version of the metric was used. Pairwise diversity is excluded here because it is a per-author metric, while the other metrics are per-paper.

Metric	Innovation	Entropy	Citation Div.	Reference Div.	Citation Count
Innovation	1.000	0.034	0.019	0.006	0.002
Entropy	0.034	1.000	0.086	0.115	0.001
Citation Div.	0.019	0.086	1.000	0.109	0.188
Reference Div.	0.006	0.115	0.109	1.000	0.030
Citation Count	0.002	0.001	0.188	0.030	1.000

Table 4: Pearson correlations between metrics derived from paper data.

A.3.2 Comparison of Metric Means between Superstars and Non-Superstars

Metric	Superstar Mean	Non-Superstar Mean	P-Value	Cohen's d
Innovation	42.77	18.17	0.0053	0.115
Entropy	2.13	2.02	4.51×10^{-20}	0.262
Citation Diversity	0.13	0.08	1.46×10^{-39}	0.544
Reference Diversity	0.10	0.08	0.00018	0.131
Citation Count	86.65	12.62	2.47×10^{-12}	0.259
Pairwise Diversity	0.35	0.13	7.60×10^{-80}	0.761

Table 5: Comparison of metrics between superstar and non-superstar authors.