Generals or Soldiers? Scholars' Roles in Interdisciplinary Collaboration

Aoxia Xiao[*] and Nicolas Robinson-Garcia[**]

[*] *xax_929@whu.edu.cn*
ORCID: 0000-0003-2173-0996
School of Information Management, Wuhan University, P.R. China

[**] *elrobin@ugr.es*
ORCID: 0000-0002-0585-7359
Unit for Computational Humanities and Social Sciences (U-CHASS), EC3 Research Group, University of Granada, Spain

Abstract: Interdisciplinary research has become increasingly prevalent in academia, yet it faces numerous challenges, including barriers related to disciplinary boundaries, academic norms, and authorship practices. This study explores authorship dynamics across diverse research topics to better understand how scholars contribute to interdisciplinary endeavors. Using data from PLOS Publishers and ScienceDirect comprising over 750,000 publications and 2 million authors, we examine patterns of authorship and contribution across different research topics. Our analysis reveals consistent usage patterns of Contributor Roles Taxonomy (CRediT) categories across various research topics, indicating a degree of uniformity in author contributions. Through K-means clustering, our analysis identifies four distinct author clusters: "Sergeants," "Soldiers," "Generals," and "Field Commanders." Each cluster represents unique patterns of publication output, topic involvement, and CRediT category usage. These findings offer insights into the complexities of interdisciplinary collaboration, providing valuable knowledge for improving collaboration strategies and advancing interdisciplinary research initiatives.

Keywords: Interdisciplinary research, Authorship, Contributorship, Research collaboration.

1. Introduction

Today, interdisciplinary research is seen as innovative and progressive (Bruns, 2013). Scholars are increasingly transcending traditional disciplinary boundaries, engaging in collaboration across multiple subject areas. Despite growing interest in interdisciplinary research, various barriers hinder its full realization, such as disciplinary boundaries, academic norms, institutional structures, and funding mechanisms (Siedlok & Hibbert, 2014). And authors tend to collaborate primarily within their own discipline (Feng & Kirkley, 2020). Diverse disciplinary traditions shape varying norms, modes of sharing, collaboration, and interaction in research practices (Siedlok & Hibbert, 2014). These differences influence the willingness to cultivate skills for interdisciplinary collaboration, often differing from those requisites as a lone researcher or within a narrow disciplinary team (Jeffrey, 2003). Furthermore, differences in disciplinary traditions often lead to tensions and mistrust among members of interdisciplinary research teams, particularly regarding authorship attribution (Knight & Pettigrew, 2007). Misunderstandings and disputes over authorship are common, but can disrupt knowledge sharing and collaboration, blur research accountability, and result in erroneous attributions (Smith & Master, 2017). This highlights the need for a nuanced understanding of research contributions across diverse domains, and the crucial role in delineating individual contributions and developing interdisplinary collaboration. Authors contribute expertise, insights, and methodologies across multiple subject areas. Understanding authorship patterns across disciplines is vital for grasping interdisciplinary collaboration dynamics and assessing interdisciplinary research impact.

Based on these considerations, this study aims to investigate authorship across diverse research topics, emphasizing how different disciplines collaborate and how authors contribute to interdisciplinary research. We seek to understand authors' roles in interdisciplinary research to improve collaboration strategies in academia.

Central to our investigation are the following research questions:
(1) Whether there are differences in common contributor roles and division of labour across different disciplines.
(2) Whether authors demonstrate variations in their contributor roles when participating in research across different disciplines.

The introduction of author contribution lists in scholarly publishing allows researchers to list all study elements and contributors. Through a standardized "taxonomy," manuscript submission software can make it easy for researchers to assign contributor roles in a structured format during the paper development and publication (Allen et al., 2014). This approach enhances collaboration by clarifying expertise of each contributor (Singh Chawla, 2015). Subsequently, scientometric community used contribution information in contributorship and authorship analysis (Larivière et al., 2016; Lu et al., 2022; Walsh et al., 2019), credit allocation and author ranking (Rahman et al., 2017; Ding et al., 2021; Yang et al., 2022), researcher skills analysis and evaluation (Kong et al., 2019; Xiao et al., 2024). Thus, this study uses contribution information to analyse contributor roles and division of labour in interdisciplinary research.

2. Methods

2.1. Data Collection and Preprocessing

Data for this study were obtained from two primary sources: publications published by the PLOS Publishers between 2018 and 2023 and publications with structured author contributions from a copy of Elsevier's data provided by the International Centre for the Study of Research (ICSR) Lab. PLOS facilitated their dataset in December 2023, covering 97,819 publications from 2018 to 2023, all with author contribution information. The ICSR Lab dataset, collected from a Scopus snapshot on October 3, 2023, included 659,579 publications with structured author contributions from 2017 to 2024, along with SciVal Topics of Prominence[1] for each publication.

The data processing workflow and remaining data at each step are detailed in Figure 1 as follows:
(1) PLOS data were matched with corresponding records from the Scopus database using their DOIs. Topic information was retrieved at the publication level. Scopus Author Identifiers were then matched based on names, affiliations, and publications. This resulted in 94,637 publications involving 473,044 authors and 629,089 publication-author combinations.
(2) Regarding the structured author contributions data from ScienceDirect, topic information was also obtained at the publication level. This yielded 659,306 publications involving 1,914,291 authors and 3,592,875 publication-author pairs.
(3) The two datasets were merged, and data cleaning procedures were conducted. Elements of author contributions not in the 14 CRediT categories were removed. Publication-author combinations with missing contribution information were excluded. The final dataset comprised 753,937 publications involving 2,287,685 authors and 4,221,306 publication-author pairs.

It is important to note that the data cleaning process was conducted at the publication-author pair level, so not all authors for the 753,937 publications are included in the final dataset.

[1] https://www.elsevier.com/products/scival/overview/topic-prominence

Figure 1. Data processing workflow

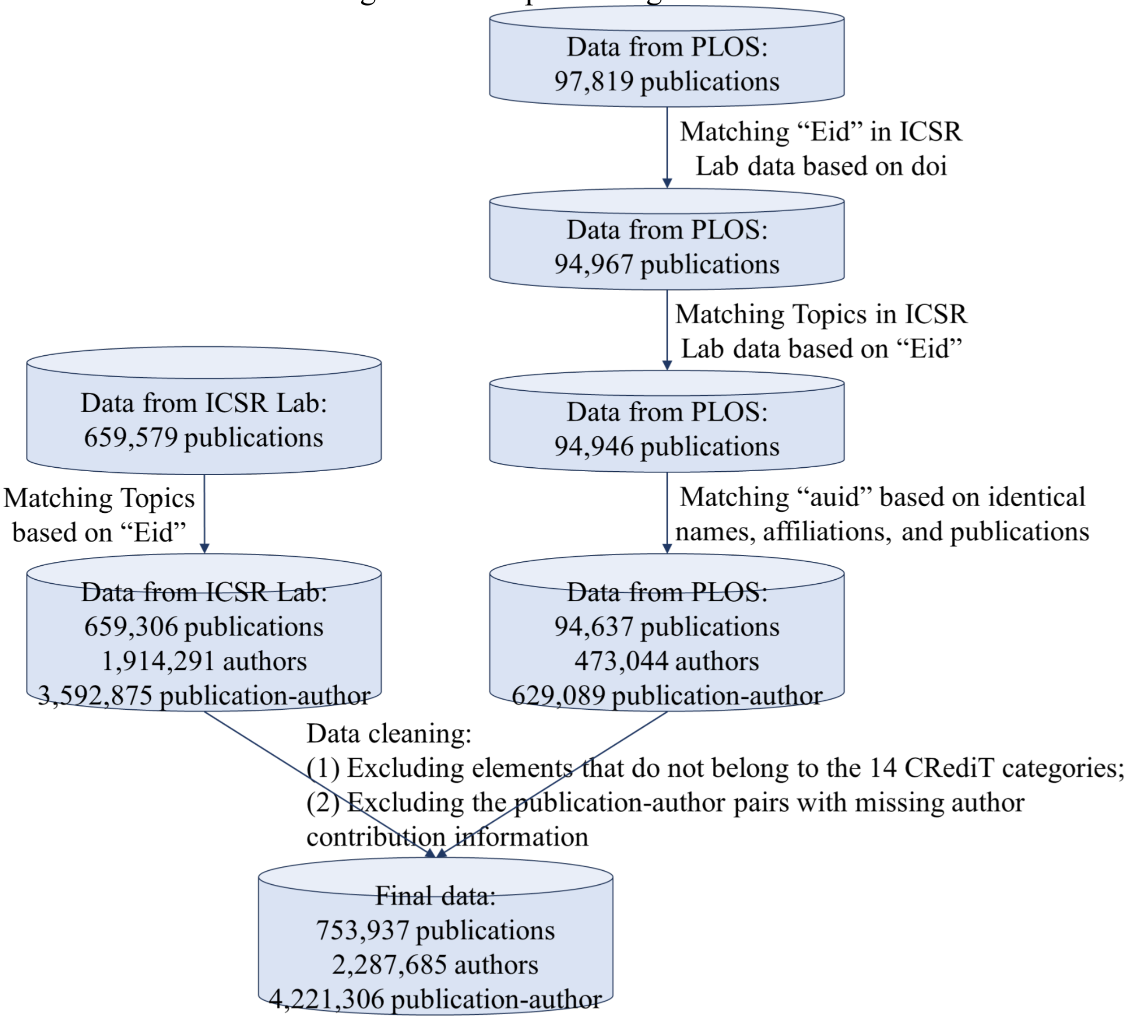

Data from PLOS: 97,819 publications
Matching "Eid" in ICSR Lab data based on doi
Data from PLOS: 94,967 publications
Matching Topics in ICSR Lab data based on "Eid"
Data from PLOS: 94,946 publications
Matching "auid" based on identical names, affiliations, and publications
Data from PLOS: 94,637 publications 473,044 authors 629,089 publication-author
Data from ICSR Lab: 659,579 publications
Matching Topics based on "Eid"
Data from ICSR Lab: 659,306 publications 1,914,291 authors 3,592,875 publication-author
Data cleaning:
(1) Excluding elements that do not belong to the 14 CRediT categories;
(2) Excluding the publication-author pairs with missing author contribution information
Final data: 753,937 publications 2,287,685 authors 4,221,306 publication-author

Table 1. The fields of the data used in this study

Field	**Description**	**Data Example**
Eid	Scopus publication ID.	85077340825
doi	Publication's DOI.	10.1371/journal.pone.0226925
Year	Publication year as used by Scopus and metrics.	2019
auid	Author ID in ICSR Lab data.	57191161224
CRediT	Author's contributor role in the publication.	["Writing – review & editing", "Validation", "Supervision", "Resources", "Methodology", "Formal analysis"]
TopicClusterId	Topic cluster ID (each publication belongs to one and only one cluster)	595

The final dataset contains publication, author, author contributions according to the CRediT standard, and topic information for each publication. Specifically, the fields in the dataset are outlined in Table 1.

Elsevier utilizes a vast citation network of millions of Scopus-indexed publications to establish approximately 96,000 Topics, defined by strong internal and weak external citation links. Employing direct citation analysis, these Topics are then grouped into 1,500 Topic Clusters, formed when citation link strength between Topics exceeds a threshold. The final dataset used in this study involved 1471 different topic clusters.

2.2. Analysis Methods

Statistical Analysis

Statistical tests were conducted to examine significant differences in the distribution of usage proportions of CRediT categories across different topic clusters. By calculating the number of publications contained within each topic cluster and the frequency of usage of each contributor role within the cluster (i.e., the number of publications within the cluster that contain each contributor role), the proportion of usage of each contributor role within each cluster was obtained. The null hypothesis stated no significant differences in these proportions.

Three tests were employed based on data characteristics:
(1) Friedman Test (Friedman, 1940): To assess differences in the medians of multiple related samples.
(2) Welch's t-test (Welch, 1947): To examine differences in the means of the samples.
(3) Kruskal-Wallis Test (Kruskal & Wallis, 1952): To compare differences in medians among multiple groups.

K-means Clustering

Our analysis focuses on evaluating authors' contribution patterns across different topics. To ensure the robustness of our analysis, we excluded authors contributing to only one publication, comprising approximately 67.4% of the dataset. After this exclusion, our dataset comprised 744,915 authors contributing to 2,678,536 publication-author pairs. We examined the distribution of remaining authors' publications across different topic clusters, quantifying the number of distinct topic clusters in which authors published. Notably, 204,479 authors (approximately 27.4%) exclusively published within a single topic cluster. Moreover, the author with the highest number of publications across topic clusters contributed to 41 distinct clusters.

We group authors with two or more publications using the K-means clustering method based on four attributes:
(1) The number of topic clusters in which the author has participated, indicating the breadth of the author's research interests and contributions across various topic areas.
(2) The total number of publications authored by the individual, providing an overview of the author's overall scholarly output.
(3) The total count of different CRediT categories used by the author, representing the diversity of roles undertaken by the author across all their publications. This value should be greater than or equal to 1 and less than or equal to 14, as per the CRediT taxonomy.
(4) The average number of CRediT categories used by the author within a single topic cluster, indicating the depth of the author's involvement and contribution within specific research topics.

The Elbow Method was applied, with additional consideration given to Silhouette Score and Calinski-Harabasz index for finding the most optimal clustering.

3. Results & Discussion

3.1. CRediT category usage across topics

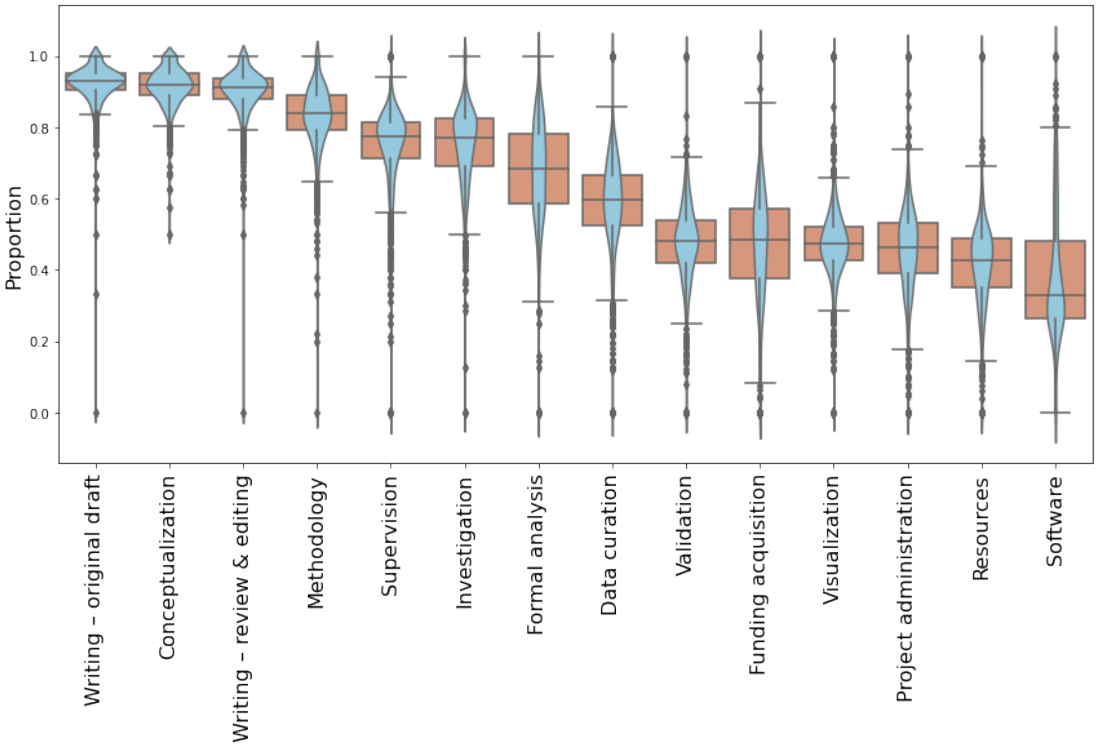
This study aims to analyse differences in CRediT category usage across various topic clusters. Results of statistical tests are shown in Table 2. P-values from all tests are relatively large, with Friedman and Kruskal-Wallis tests yielding a P-value of 1. Thus, the null hypothesis is not rejected, indicating no significant differences in the usage proportion of CRediT category distribution across topic clusters. In other words, while specific usage proportions may vary, overall distribution remains consistent.

Figure 2 illustrates the usage patterns of various contributor roles. Box plots and violin plots depict the distribution of each contributor role's usage within individual topic clusters. Overall, "Writing – original draft," "Conceptualization," and "Writing – review & editing" are the most frequently utilized roles. Following closely are "Methodology," "Supervision," "Investigation," and "Formal analysis." Conversely, "Data curation," "Validation," "Funding acquisition," "Visualization," "Project administration," "Resources," and "Software" are less commonly used. This distribution pattern is generally consistent across individual topic clusters as well.

Table 2. The results of the tests

Test Method	**Test Statistic**	**p-value**
Friedman test	1216.967	1.000
Welch's t-test	1.036	0.310
Kruskal-Wallis Test	1209.514	1.000

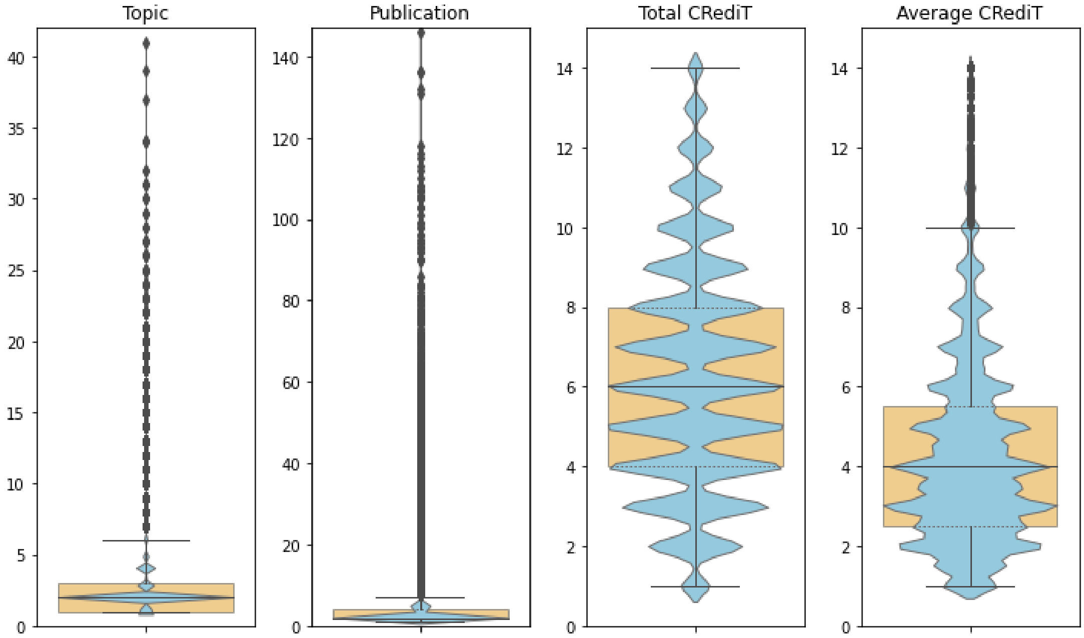
Figure 2. The usage of various contributor roles across different topic clusters



3.2. Difference of CRediT category usage across topics among author clusters

In this study, we analysed the value distribution of four clustering attributes for authors with two or more publications: the number of topic clusters, publication count, total CRediT usage, and average CRediT usage per topic cluster (see Figure 3). Our findings reveal that the majority of authors possess a limited number of publications, focusing their contributions on a select few topic clusters. Additionally, for most authors, average CRediT usage per topic cluster is lower than the total CRediT usage, indicating that authors are involved in different CRediT categories across different topic clusters.

Figure 3. Distribution of clustering attributes for authors with two or more publications



We use K-means clustering method to group authors with two or more publications based on the previous four attributes. To determine the optimal number of clusters we applied an elbow method. Within-Cluster Sum of Squares (WCSS), indicating cluster compactness, is plotted against different cluster numbers (K) (see Figure 4). The elbow point on the graph indicates the optimal K value, representing the point from which adding more clusters no longer significantly reduces WCSS. In our dataset, the elbow point may correspond to 3, 4, or 5 clusters. Furthermore, we computed the Silhouette Score and Calinski-Harabasz index (see Table 3.). Both indices peak at a cluster number of 4, indicating optimal clustering. The amount of data for each cluster is shown in Figure 5 and the box plot of normalized attribute data across clusters can be seen in Figure 6.

After clustering analysis, four distinct author clusters were identified based on publication and contribution patterns:

Cluster 1: "Sergeants" - Although numerous (250,085, comprising 33% of the dataset), their involvement in topic clusters and total publications are relatively low. They employ a variety of CRediT categories across publications, but the average usage within individual topic clusters is comparatively low. This suggests that while this group is sizable, their depth of engagement in research is modest. They contribute to various tasks of research, indicating a tendency to assume multiple roles but possibly lower professionalism. Authors in this cluster are similar to sergeants who are less involved in the overall battlefield, but contribute to the various tasks within the unit.

Cluster 2: "Soldiers" - Cluster 2 represents the most extensive group in the dataset, comprising 414,111 authors, accounting for 56% of the dataset. Authors in this group have publications in fewer topics with low overall productivity. They exhibit involvement in limited CRediT categories. And the difference between the total usage of CRediT and average CRediT usage per topic compared to Cluster 1 is minimal. This indicates their focus on specific tasks, contributing to fewer categories. They may contribute significantly within narrow research scopes or tasks, yet their overall research output remains relatively low, akin to soldiers tasked with individual missions.

Cluster 3: "Generals" - Authors in this group have publications across multiple topics with relatively high overall and per-topic publication rates. They employ numerous CRediT categories across publications. Although they constitute the smallest proportion of the dataset (6,495, 1%), they exhibit high productivity and contribution. This suggests broad engagement and high productivity across various fields, possibly indicating leadership or expertise in multiple domains, leading teams or projects, and making significant contributions in academics, similar to generals who oversee strategic operations.

Cluster 4: "Field Commanders" - Authors in this group have publications in multiple topics with moderate involvement. They represent a small portion of the dataset (74,224, 10%), with moderate overall and per-topic publication rates. They use various CRediT categories across publications, with relatively high average usage CRediT per topics. This indicates moderate contribution across multiple topics or research tasks. They may prefer broad engagement across fields rather than specialized focus, possibly serving as popular collaborators providing balanced contributions and expertise across different areas in academia, akin to field commanders who focus on specific areas of operation.

Figure 4. Within-Cluster Sum of Squares (WCSS) vs. Number of Clusters (K)

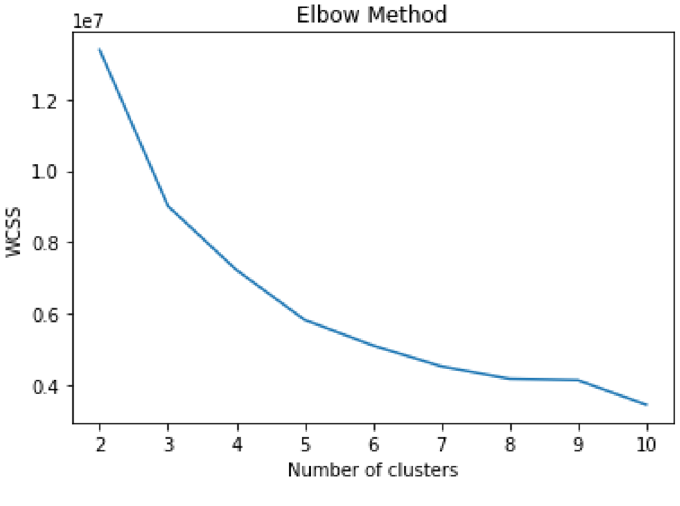


Table 3. Silhouette Score and Calinski-Harabasz index

Number of clusters	Silhouette Score	Calinski-Harabasz index
3	0.5732	0.5732
4	0.6014	0.6014
5	0.5197	0.5197

Figure 5. The amount of data for each cluster

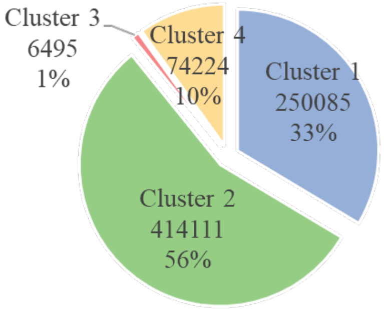
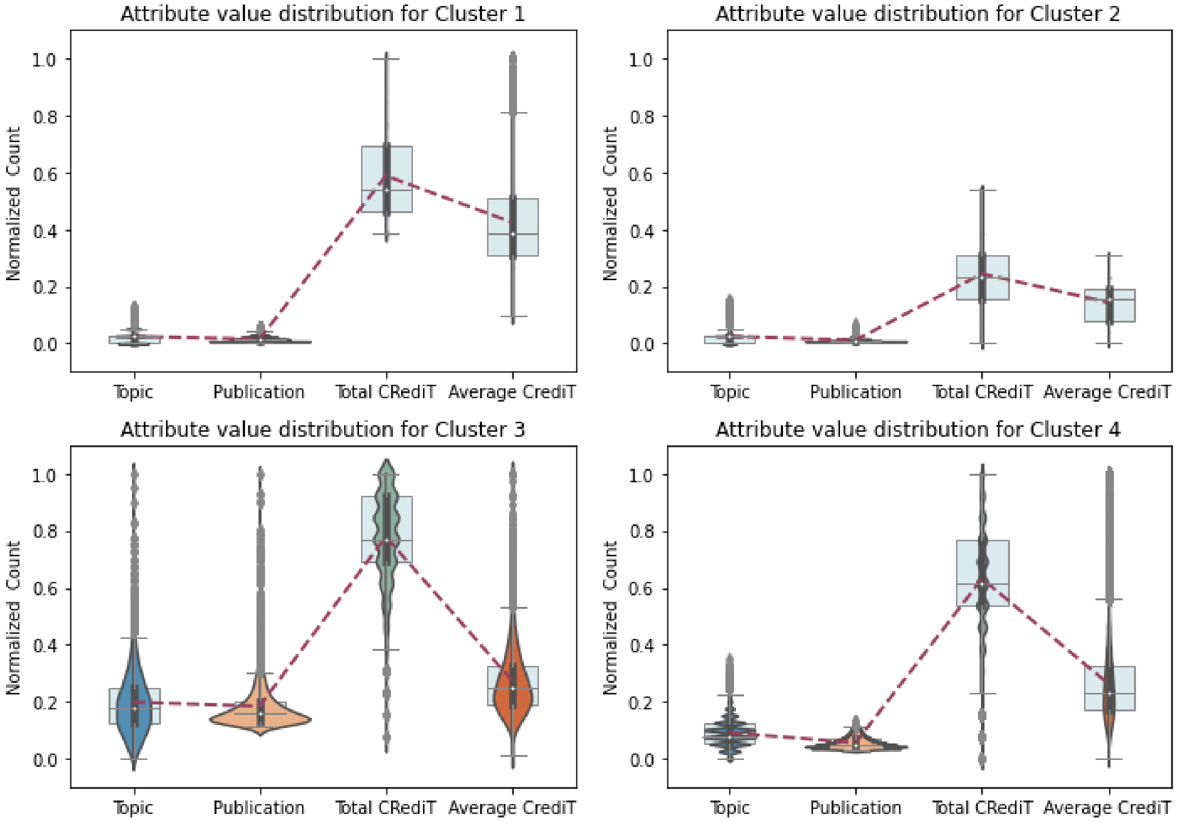


Figure 6. Distribution and mean value of attributes across different author clusters

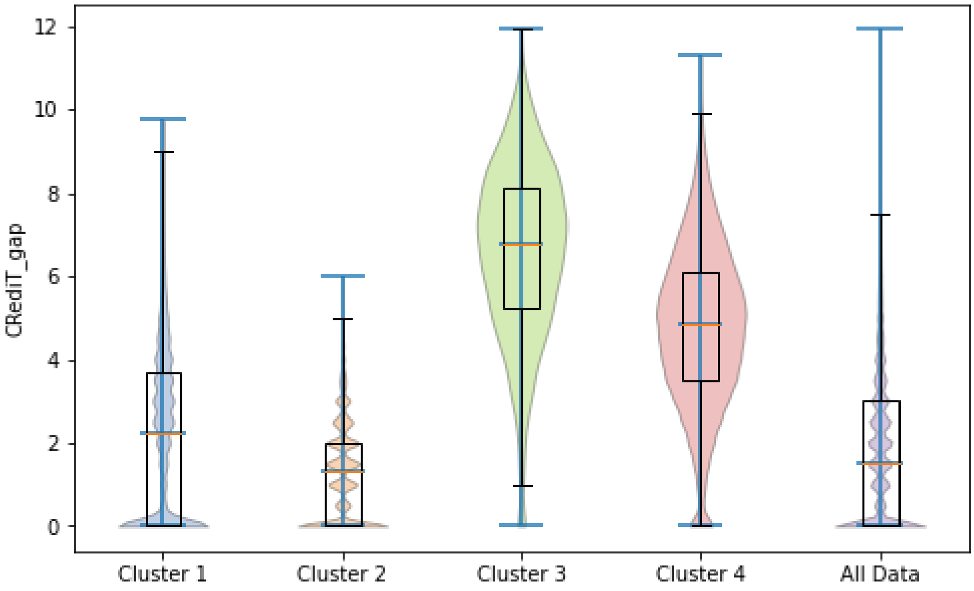


This study further compared CRediT category usage across different topic clusters among authors. We calculated authors' average CRediT category usage within topic clusters and compared it with their total CRediT category count. A larger difference may indicate varied category combinations across topics, while a smaller difference suggests consistent category usage across topics (see Figure 7). We excluded 204,479 authors who published in only one topic cluster, as meaningful comparison was not possible. Among the remaining 540,436 authors, Cluster 1 comprises 162,090 authors (30%), Cluster 2 contains 299,546 authors (56%), Cluster 3 includes 6,423 authors (1%), and Cluster 4 encompasses 72,377 authors (13%).

For most authors, the gap between the total CRediT category usage and average usage per topic is generally small, typically within 4 (as depicted in the "All data" violin plot in Figure 7). Cluster 3 and Cluster 4 show larger differences, whereas Cluster 1 indicates smaller differences. Cluster 2 exhibits the smallest difference, suggesting authors in Cluster 2 are less inclined to undertake varied tasks/roles across topics, consistent with clustering attribute distribution.

Cluster 1 and Cluster 2, representing the majority of authors, exhibit minimal variance between total and average CRediT usage, suggesting similar tasks across diverse topics. Cluster 2 authors, typical of most scholars, have lower publication output and undertake limited tasks. In contrast, Clusters 3 and 4 show larger differences in CRediT category usage, indicating diverse roles across topics. Despite having fewest authors, Cluster 3 displays high output and engages in a broad range of topics, indicating extensive interdisciplinary involvement. Additionally, distinct role differences exist across topics within this author cluster, with authors displaying pronounced variations in roles across topics. They participate significantly in interdisciplinary research, showcasing substantial diversity in roles and tasks across topics.

Figure 7. CRediT Category Count Difference Distribution



4. Conclusion

This study examines authors' contributorship across various research topics using data from PLOS Publishers and ICSR Lab, analysing a dataset comprising over 750,000 publications and 2 million authors. Our analysis reveals several notable findings. Firstly, while the specific usage proportion of the CRediT category varies across topic clusters, statistical tests show no significant differences in distribution. This suggests consistent CRediT category usage patterns across diverse topics. Secondly, using K-means clustering, we identify four author clusters based on publication and contribution patterns. Our analysis identifies four distinct author clusters: "Sergeants," "Soldiers," "Generals," and "Field Commanders." Each cluster represents unique patterns of publication output, topic involvement, and CRediT category usage. These findings underscore the diverse roles and engagement levels of authors across different research topics.

In summary, our findings highlight the complex nature of scholarly contributions across varied research fields. These insights are pivotal for fostering enhanced collaboration, refining author roles, and advancing interdisciplinary research. However, our study only provides an initial exploration, focusing on differences in authorship patterns across research topics without analysing variations among author types. Future research will explore specific contributor roles and tasks, such as conceptual work, practical tasks, and resource acquisition, to offer a more comprehensive analysis of labour division and skill utilization in interdisciplinary research.

Open science practices

In reflection on the use of open science practices within our research, it's important to note that our dataset comprises two distinct components: publications published by PLOS Publishers and publications obtained from a copy of Elsevier's data provided by ICSR Lab. The data from PLOS Publishers is OA data, retrievable directly from the HTML files on the PLOS website. However, the data from Elsevier, provided by ICSR Lab, is not openly accessible. Firstly, the metadata originates from Elsevier, which is proprietary. Secondly, the data has been structured by ICSR Lab, involving data curation efforts. The rationale behind incorporating non-public data lies in our research's aim to derive generalized conclusions from a substantial and representative dataset. Thus, the inclusion of non-public data was necessary to ensure an adequate volume and diversity of data for robust analysis.

Acknowledgments

Authors would like to acknowledge Veronique Kiermer, Chief Scientific Officer at PLOS for facilitating access to PLOS data on author contribution statements. They also acknowledge the ICSR Lab from Elsevier for granting access to ScienceDirect and Scopus data on author contribution statements and bibliographic metadata.

Author contributions

AX: Conceptualization, Data Curation, Formal Analysis, Methodology, Software, Visualization, Writing – original draft,
NRG: Conceptualization, Resources, Supervision, Writing – review & editing

Competing interests

Authors declare that they have no competing interests.

Funding information

This study is part of the COMPARE project funded by the Spanish Ministry of Science, (MCIN/AEI/https://doi.org/10.13039/501100011033) FSE invierte en tu future. Aoxia Xiao is supported by a scholarship from the China State Scholarship Fund. Nicolas Robinson-Garcia is supported by a Ramón y Cajal Fellowship granted by the Spanish Ministry of Science (PID2020-117007RA-I00).

References

Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. Nature News, 508(7496), 312. https://doi.org/10.1038/508312a

Bruns, H. C. (2013). Working Alone Together: Coordination in Collaboration across Domains of Expertise. Academy of Management Journal, 56(1), 62–83. https://doi.org/10.5465/amj.2010.0756

Ding, J., Liu, C., Zheng, Q., & Cai, W. (2021). A new method of co-author credit allocation based on contributor roles taxonomy: Proof of concept and evaluation using papers published in PLOS ONE. Scientometrics, 126(9), 7561–7581. https://doi.org/10.1007/s11192-021-04075-x

Feng, S., & Kirkley, A. (2020). Mixing Patterns in Interdisciplinary Co-Authorship Networks at Multiple Scales. Scientific Reports, 10(1), 7731. https://doi.org/10.1038/s41598-020-64351-3

Friedman, M. (1940). A Comparison of Alternative Tests of Significance for the Problem of m Rankings. The Annals of Mathematical Statistics, 11(1), 86–92. https://doi.org/10.1214/aoms/1177731944

Jeffrey, P. (2003). Smoothing the Waters: Observations on the Process of Cross-Disciplinary Research Collaboration. Social Studies of Science, 33(4), 539–562. https://doi.org/10.1177/0306312703334003

Knight, L., & Pettigrew, A. (2007). Explaining Process and Performance in the Co Production of Knowledge: A Comparative Analysis of Collaborative Research Projects.

Kong, X., Liu, L., Yu, S., Yang, A., Bai, X., & Xu, B. (2019). Skill ranking of researchers via hypergraph. PeerJ Computer Science, 5, e182. https://doi.org/10.7717/peerj-cs.182

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. Journal of the American Statistical Association, 47(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Larivière, V., Desrochers, N., Macaluso, B., Mongeon, P., Paul-Hus, A., & Sugimoto, C. R. (2016). Contributorship and division of labor in knowledge production. Social Studies of Science, 46(3), 417–435. https://doi.org/10.1177/0306312716650046

Lu, C., Zhang, C., Xiao, C., & Ding, Y. (2022). Contributorship in scientific collaborations: The perspective of contribution-based byline orders. Information Processing & Management, 59(3), 102944. https://doi.org/10.1016/j.ipm.2022.102944

Rahman, M. T., Regenstein, J. M., Kassim, N. L. A., & Haque, N. (2017). The need to quantify authors' relative intellectual contributions in a multi-author paper. Journal of Informetrics, 11(1), 275–281. https://doi.org/10.1016/j.joi.2017.01.002

Siedlok, F., & Hibbert, P. (2014). The Organization of Interdisciplinary Research: Modes, Drivers and Barriers. International Journal of Management Reviews, 16(2), 194–210. https://doi.org/10.1111/ijmr.12016

Singh Chawla, D. (2015). Digital badges aim to clear up politics of authorship. Nature, 526(7571), 145–146. https://doi.org/10.1038/526145a

Smith, E., & Master, Z. (2017). Best Practice to Order Authors in Multi/Interdisciplinary Health Sciences Research Publications. Accountability in Research, 24(4), 243–267. https://doi.org/10.1080/08989621.2017.1287567

Walsh, J. P., Lee, Y.-N., & Tang, L. (2019). Pathogenic organization in science: Division of labor and retractions. Research Policy, 48(2), 444–461. https://doi.org/10.1016/j.respol.2018.09.004

Welch, B. L. (1947). The Generalization of "Student's" Problem When Several Different Population Variances are Involved. Biometrika, 34(1–2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28

Xiao, A., Yang, S., Yue, M., & Jin, M. (2024). What Research Skills Do Scholars Excel at?—Based on Individual Contribution and External Recognition. In I. Sserwanga, H. Joho, J. Ma, P. Hansen, D. Wu, M. Koizumi, & A. J. Gilliland (Eds.), Wisdom, Well-Being, Win-Win (pp. 301–321). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-57860-1_21

Yang, S., Xiao, A., Nie, Y., & Dong, J. (2022). Measuring coauthors' credit in medicine field—Based on author contribution statement and citation context analysis. Information Processing & Management, 59(3), 102924. https://doi.org/10.1016/j.ipm.2022.102924