Supplementary material for

Thirty-Two Years of IEEE VIS: Authors, Fields of Study and Citations

Hongtao Hao
hongtaoh@cs.wisc.edu

Yumian Cui
ycui53@wisc.edu

Zhengxiang Wang
jackwang196531@gmail.com

Yea-Seul Kim
yeaseul.kim@cs.wisc.edu.

July 30, 2022

Contents

1 Methods

1.1 Get DOIs for VIS 2021 papers

For a paper whose DOI as obtained from Crossref contains the IEEE prefix, i.e., `10.1109`, and whose title as obtained from Crossref matches the queried title, we regarded the query result as correct. There are twenty-seven papers whose DOI contained this prefix, but whose title as obtained from Crossref ID did not match that on VIS 2021. By comparing these titles manually, we found that four of them were indeed mismatches; the rest were correct results with insignificant variations in titles. There were twenty-three papers whose DOIs as obtained from Crossref did not contain `10.1109`; Their query results were obviously wrong. In sum, $23 + 4 = 27$ papers had incorrect query results. We manually collected their DOIs from IEEE Xplore.

1.2 Choose the right data sources

We first ruled out JSTOR and PubMed. JSTOR does not contain reference and citation information at all. PubMed has impressive coverage on VIS papers; IEEE Xplore even links each publication to its PubMed page. Unfortunately, PubMed provides very limited data on authors and citations and provides no data on references at all.

We then examined Web of Science and Scopus. IEEE Xplore displays each publication's citation metrics as measured by these two databases, whose coverage on VIS papers, however, is not desirable. We randomly selected 100 papers from the 3,242 VIS papers and collected their citation metrics data as displayed on IEEE Xplore. Only 71 of them had citation counts from Scopus; this figure for Web of Science was 68. We validated this result by identifying VIS papers via DOI query on Web of Science. Among all 3,242 papers, 25% were not identifiable. Some of the inaccessible papers could be identified by title query but this approach is not only labor-intensive and error-prone but also hard to automate and reproduce. Apart from this low coverage on VIS papers, we gave up Web of Science and Scopus also because of their paywalls. Both are proprietary and therefore not easily accessible. Even if we collect information from them through web-scraping, we are not allowed to share our data publicly, which defies our wish to make our work reproducible.

Crossref is free and has surprisingly good coverage on VIS papers. Among all 3,242 papers, only one is missing from its database. A closer examination of their data, however, revealed Crossref is not an ideal source of data for our present study either. For all the first authors, 99% of them miss affiliations, rendering Crossref useless for author-related data collection. In addition, Crossref provides information on fields of study only at the journal level rather than at the paper level. This means that on Crossref, VIS papers either miss fields of study, or "subject" as Crossref codes it, or they have exactly the same fields of study, which are assigned to the journal of *IEEE Transactions on Visualization and Computer Graphics*.

Semantic Scholar [1] provides a powerful API. It provides citation and reference counts and titles of both referenced and citing papers. What it lacks, however, is the author data and the fields of study. Because these two types of data are important in our study, Semantic Scholar is not a viable choice.

This left us only three options: Google Scholar, Microsoft Academic Graph (MAG), and OpenAlex [2]. Google Scholar is the most popular database in terms of web traffic [3].

[1] https://www.semanticscholar.org/

[2] https://openalex.org/

[3] https://blogs.lse.ac.uk/impactofsocialsciences/2021/05/27/goodbye-microsoft-academic-hello-open-research-infrastructure/

It provides the citation counts, which we collected, but no other useful information for this study. Between MAG and OpenAlex we chose the latter mostly because Microsoft discontinued MAG recently [4]. If we get our data from MAG, it is very likely that our workflow will not be useful for future researchers. Like MAG, OpenAlex is a scholarly database freely accessible to the public. A large amount of its data is directly from MAG. OpenAlex is new but growing and well maintained. One special advantage of OpenAlex over MAG for our study is that it provides author affiliations' type, i.e., education, company, government, etc., which is not available on MAG.

1.3 How OpenAlex collects concepts data

OpenAlex uses the word "Concept" to describe fields of study. OpenAlex built a multi-class classifier based on MAG's fields of study data [3] to assign concepts to each publication indexed in OpenAlex. OpenAlex employs 65K concepts taken from `wikidata.org`. These concepts have six levels. There are 19 Level-0 concepts, each indicating the highest level concept, for example, computer science, mathematics, medicine, physics, chemistry, etc. The larger the number of the level, the more detailed this concept is. OpenAlex detailed how they operate concepts tagging in their whitepaper [1]. They also released their codes of concept tagging [2].

2 Results

2.1 Changes in average number of authors per paper

Fig. 1 shows the changes in the average number of authors per paper by year. It is clear that it has an upward trend.

2.2 Distribution of L0 concepts in VIS, referenced, and citing papers

The following figures, namely Fig. 2, Fig. 3, and Fig. 4 present the distribution of Level 0 Concepts among VIS, referenced, and citing papers. From these figures, it is clear that in all of them, Computer Science is the most salient part, followed by Mathematics.

2.3 Concepts co-occurrence in VIS papers

The following figures, namely, Fig. 5, Fig. 6, Fig. 7, Fig. 8 show which concepts (L0-L3) in VIS papers co-occurred. From these results, it is clear that at each level only a few Concepts co-occurred, with the highest-frequency Concept at each level attracting the majority of co-occurrences.

[4] https://www.microsoft.com/en-us/research/project/academic/articles/microsoft-academic-to-expand-horizons-with-community-driven-approach/

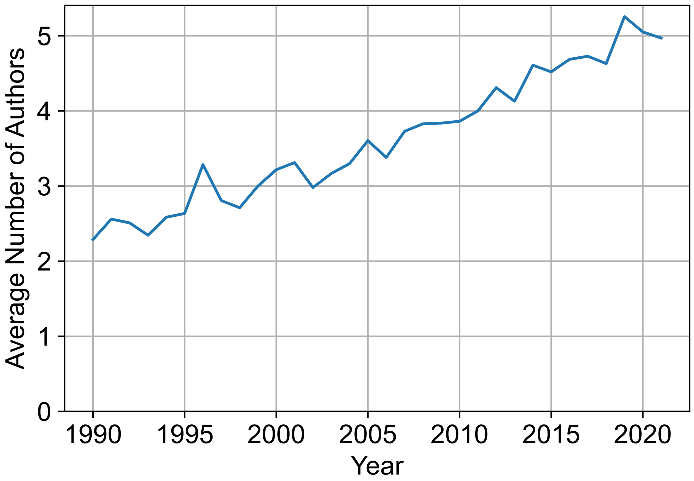

Figure 1: Average number of authors by year

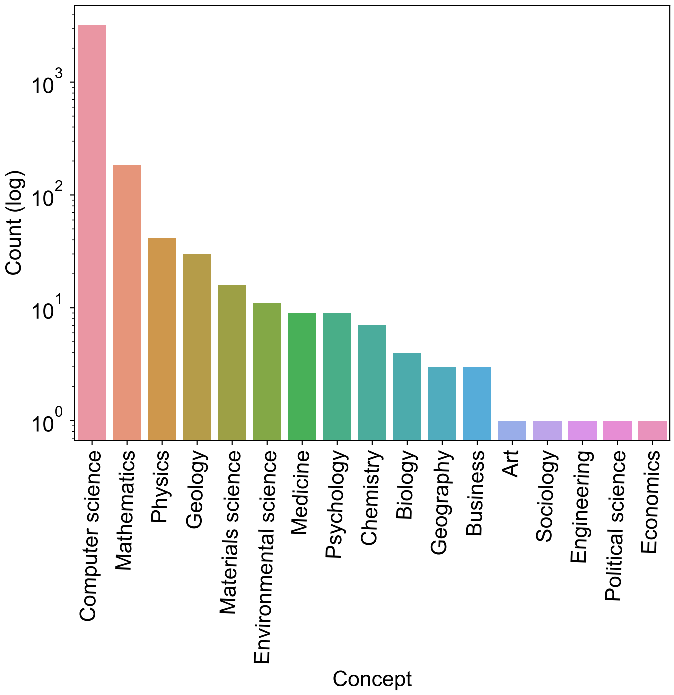

Figure 2: Distribution of L0 concepts among VIS papers

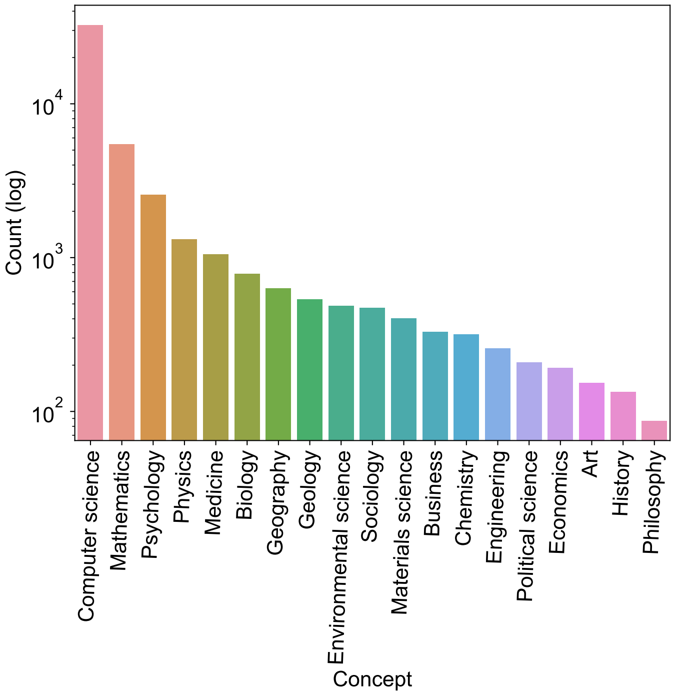

Figure 3: Distribution of L0 concepts among referenced papers

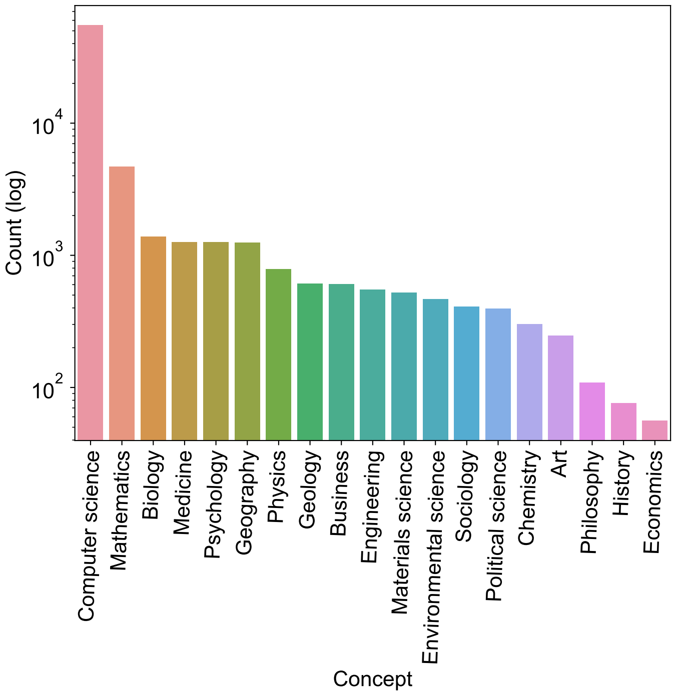

Figure 4: Distribution of L0 concepts among citing papers

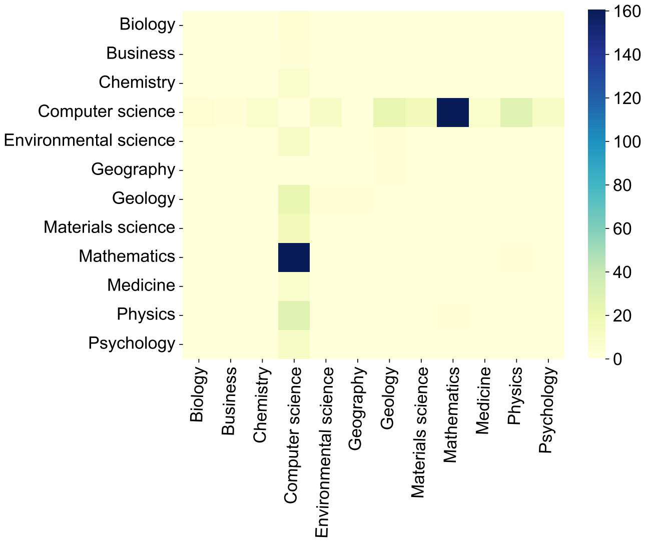

Figure 5: Co-occurrence of L0 concepts in VIS; pairs appeared at least once.

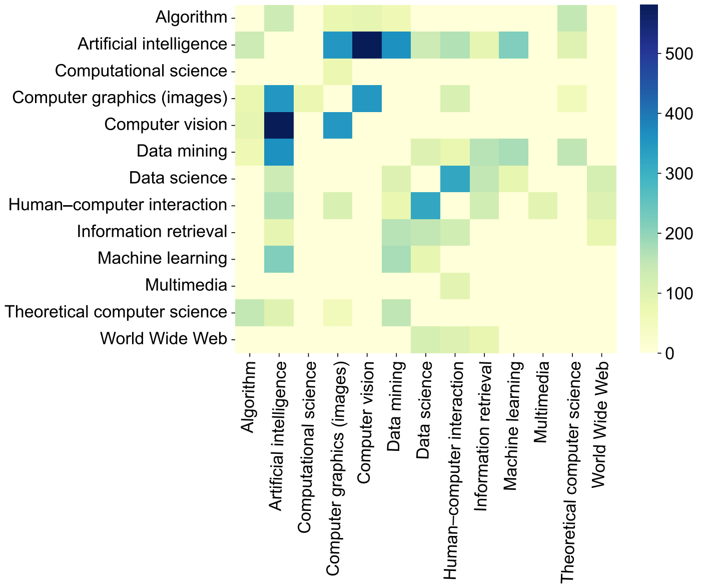

Figure 6: Co-occurrence of L1 concepts in VIS; pairs appeared at least 50 times.

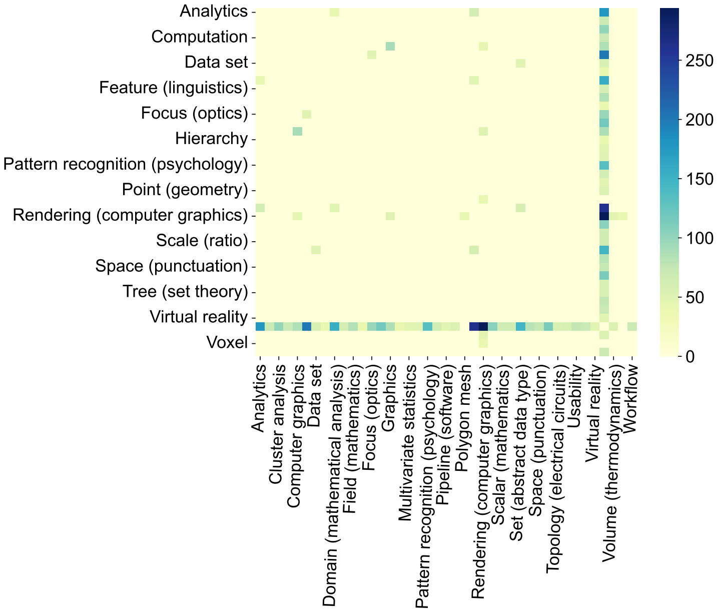

Figure 7: Co-occurrence of L2 concepts in VIS; pairs appeared at least 40 times.

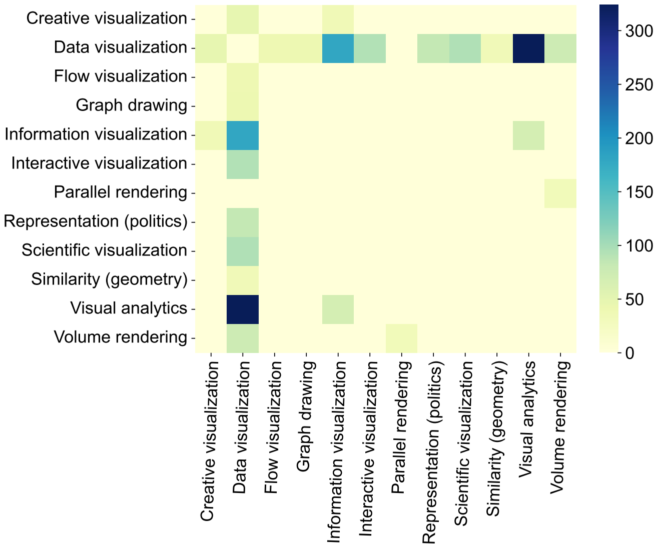

Figure 8: Co-occurrence of L3 concepts in VIS; pairs appeared at least 30 times.

2.4 Group-wise citation analysis

Given the highly skewed nature of citation counts, we employed non-parametric tests to investigate differences in groups. We rely on citation counts from OpenAlex, and will report results if Google Scholar citation counts yielded a different result.

Mann-Whitney tests showed that neither papers written by authors from affiliations of different types ($U = 606,271, p = 0.86$), nor those by authors from different countries ($U = 822,967, p = 0.29$), obtained more citations than their counterparts. Papers involving authors from the United States ($U = 1,058,864, p = 0.69$) did not get more impacts than their counterparts either. There were also no differences in the citations between journal and conference papers ($U = 1,201,856, p = 0.33$). Google Scholar citations, however, showed that cross-country collaboration papers got significantly fewer citations ($U = 783,863, p < 0.01$), and that conference papers got more citations than journal papers ($U = 1,078,280, p < 0.001$). Citations from both OpenAlex ($U = 292,878$) and Google Scholar ($U = 287,031$) both revealed that award-winning papers are more impactful ($p < 0.001$).

We ran a one-way ANOVA to compare the effect of conference tracks on the number of citations. Test results showed there were significant differences among groups ($F(3,3066) = 35.19, p < 0.001$). InfoVis papers got the highest number of citations, followed by Vis, and VAST. SciVis papers were the lowest. Tukey's HSD test revealed that all group pairs had significantly different citations except for that between Vis and VAST ($p = 0.21$).

2.5 Regression results for citation analysis

Regression results are as follows:

2.5.1 OpenAlex citations

2.5.2 Google Scholar citations

2.5.3 Log10 transformation on OpenAlex citations

References

[1] OpenAlex. *Automated concept tagging for OpenAlex, an open index of scholarly articles.* https://docs.google.com/document/d/1OgXSLriHO3Ekz0OYoaoP_h0sPcuvV4EqX7VgLLblKe4/edit#. 2022.

[2] OpenAlex. *OpenAlex Concept Tagging.* https://github.com/ourresearch/openalex-concept-tagging. 2022.

[3] Zhihong Shen, Hao Ma, and Kuansan Wang. “A web-scale system for scientific knowledge exploration”. In: *arXiv preprint arXiv:1805.12216* (2018).

Table 1: Regression results with OpenAlex citations. Estimates are unstandardized coefficients with standard error and p values.

	Dependent variable:
	Number.of.Citations
Year.Distance.from.2020	2.507^{***}
	(0.370)
ConferenceSciVis	-42.747^{***}
	(5.646)
ConferenceVAST	-14.599^{***}
	(4.919)
ConferenceVis	-37.252^{***}
	(4.298)
PaperTypeJ	22.250^{***}
	(4.866)
Number.of.Authors	-0.289
	(0.884)
Cross.type.CollaborationTrue	-1.323
	(3.630)
Cross.country.CollaborationTrue	-3.952
	(3.826)
With.US.AuthorsTrue	4.049
	(3.076)
AwardTrue	28.179^{***}
	(6.705)
Constant	28.486^{***}
	(8.196)
Observations	3,070
R^2	0.057
Adjusted R^2	0.054
Residual Std. Error	80.168 (df = 3059)
F Statistic	18.445^{***} (df = 10; 3059)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

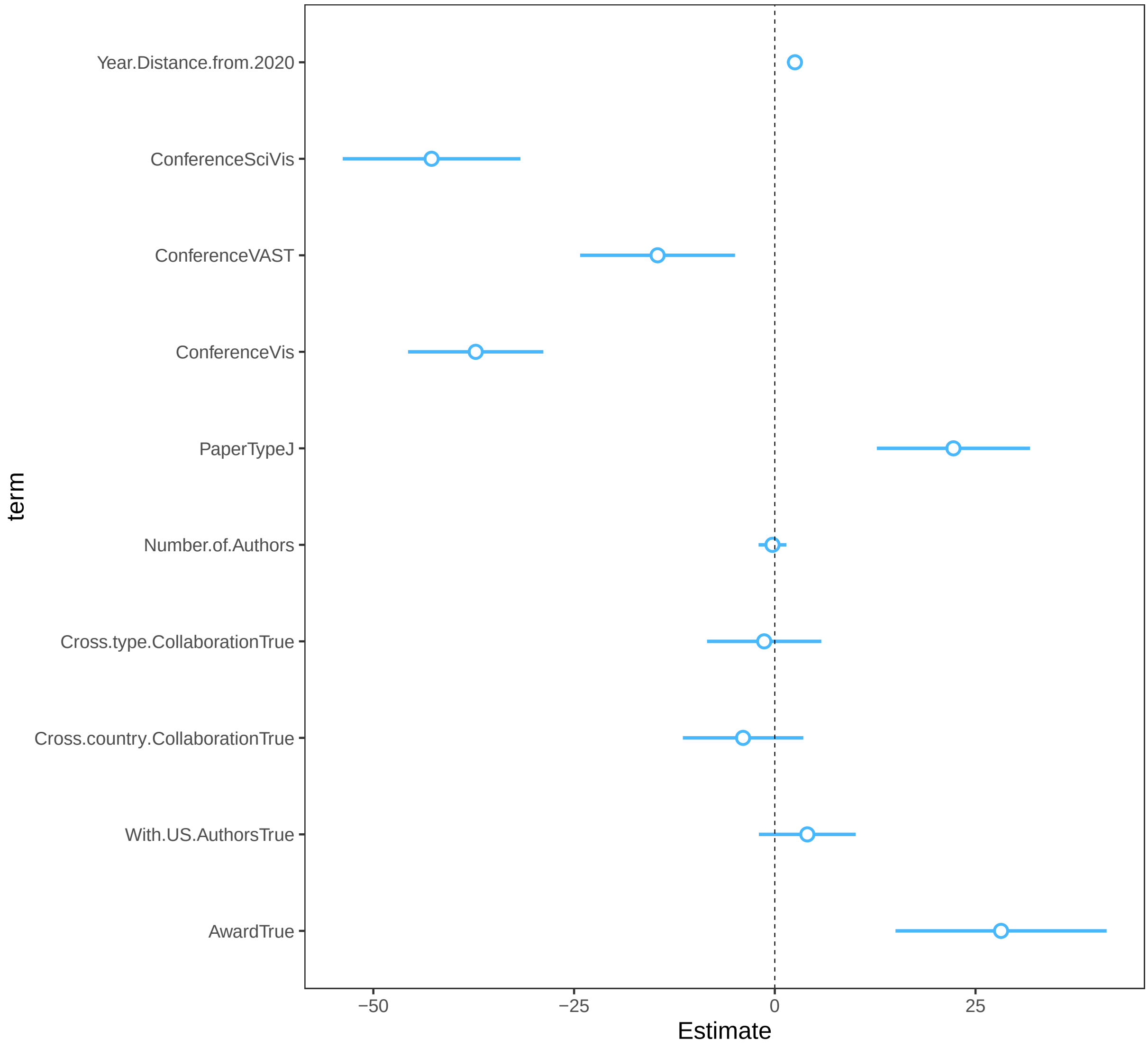

Figure 9: Regression results with OpenAlex citations. Estimates are unstandardized coefficients with 95% confidence intervals.

Table 2: Regression results with Google Scholar citations. Estimates are unstandardized coefficients with standard error and p values.

	Dependent variable:
	Citation.Counts.on.Google.Scholar
Year.Distance.from.2020	4.769^{***}
	(0.654)
ConferenceSciVis	-69.884^{***}
	(9.983)
ConferenceVAST	-28.003^{***}
	(8.698)
ConferenceVis	-64.277^{***}
	(7.601)
PaperTypeJ	25.274^{***}
	(8.605)
Number.of.Authors	-0.751
	(1.563)
Cross.type.CollaborationTrue	-4.143
	(6.418)
Cross.country.CollaborationTrue	-5.817
	(6.765)
With.US.AuthorsTrue	7.075
	(5.439)
AwardTrue	51.561^{***}
	(11.857)
Constant	52.831^{***}
	(14.493)
Observations	3,070
R^2	0.063
Adjusted R^2	0.060
Residual Std. Error	141.763 (df = 3059)
F Statistic	20.517^{***} (df = 10; 3059)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

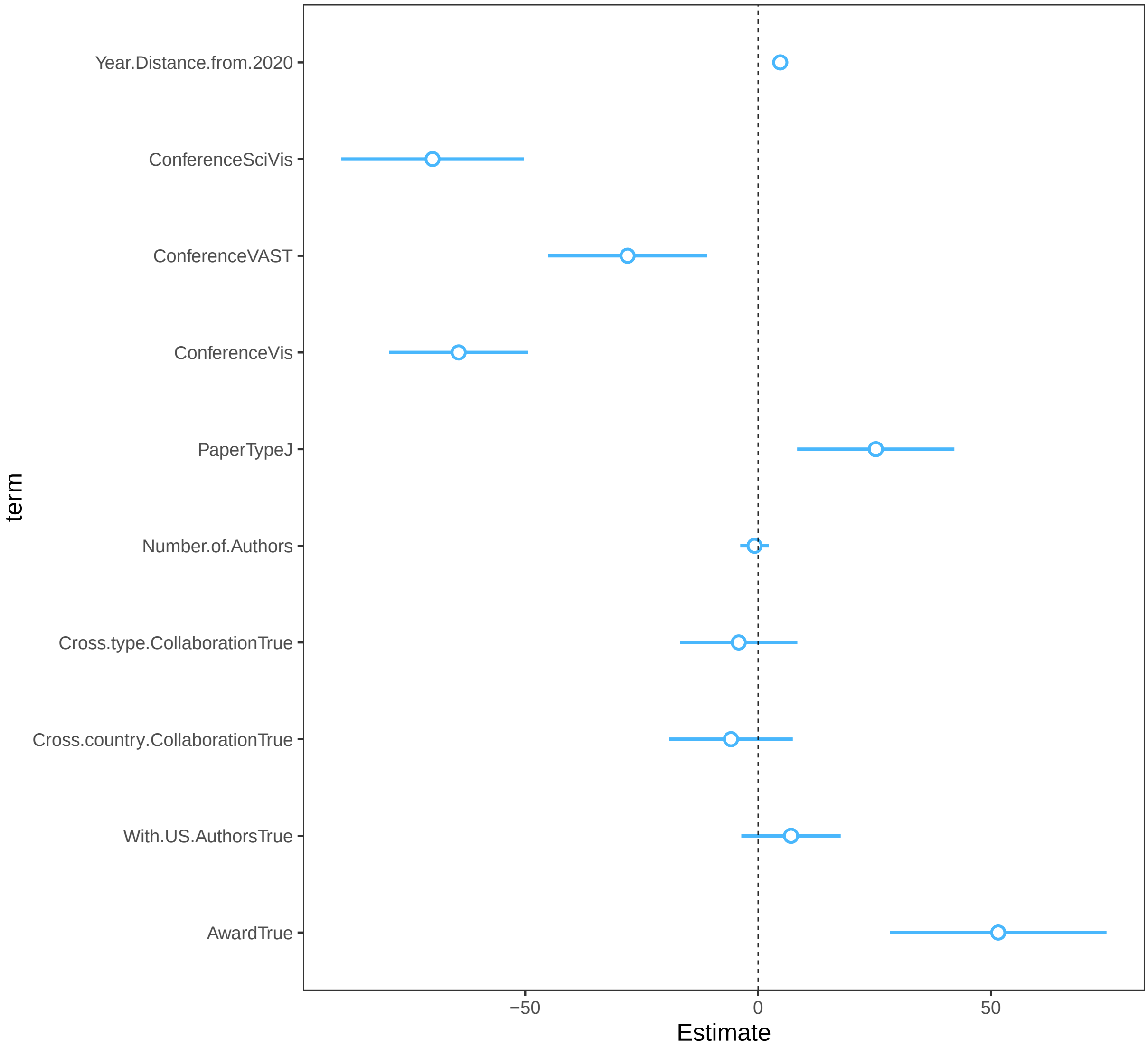

Figure 10: Regression results with Google Scholar citations. Estimates are unstandardized coefficients with 95% confidence intervals.

Table 3: Regression results with log10 transformation on OpenAlex citations. Estimates are unstandardized coefficients with standard error and p values.

	Dependent variable:
	citenum_log10
Year.Distance.from.2020	0.015^{***}
	(0.002)
ConferenceSciVis	-0.442^{***}
	(0.035)
ConferenceVAST	-0.080^{***}
	(0.031)
ConferenceVis	-0.216^{***}
	(0.027)
PaperTypeJ	0.190^{***}
	(0.031)
Number.of.Authors	-0.001
	(0.006)
Cross.type.CollaborationTrue	0.016
	(0.023)
Cross.country.CollaborationTrue	0.002
	(0.024)
With.US.AuthorsTrue	-0.014
	(0.019)
AwardTrue	0.254^{***}
	(0.042)
Constant	1.287^{***}
	(0.052)
Observations	3,070
R^2	0.086
Adjusted R^2	0.083
Residual Std. Error	0.504 (df = 3059)
F Statistic	28.628^{***} (df = 10; 3059)
Note:	*p<0.1; **p<0.05; ***p<0.01

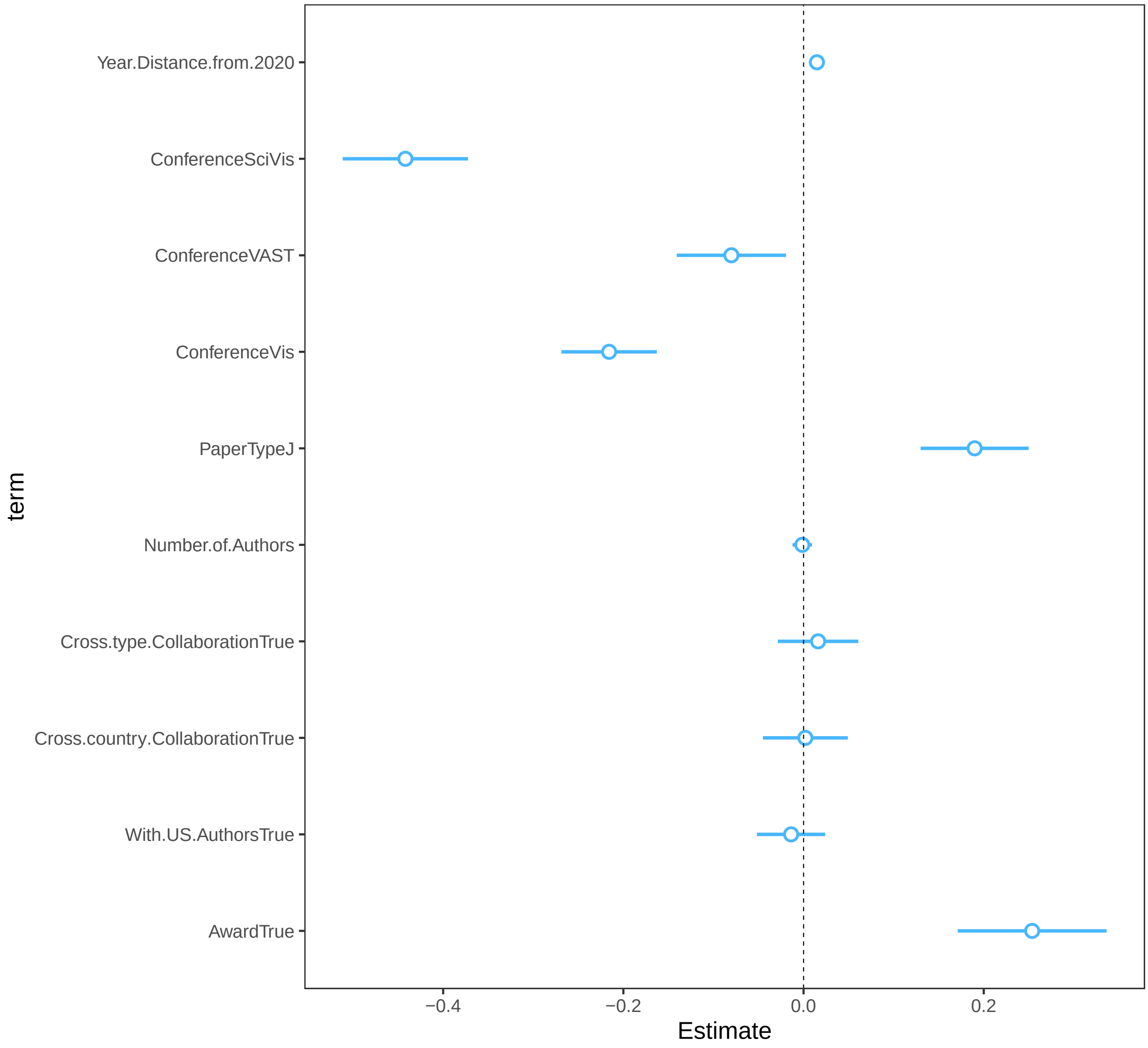

Figure 11: Regression results with log10 transformation on OpenAlex citations. Estimates are unstandardized coefficients with 95% confidence intervals.