

NRC Publications Archive Archives des publications du CNRC

Geographic citation gaps in NLP research

Rungta, Mukund; Singh, Janvijay; Mohammad, Saif M.; Yang, Diyi

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. /
La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

Publisher's version / Version de l'éditeur:

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 1371-1383, 2022-12

NRC Publications Archive Record / Notice des Archives des publications du CNRC :

<https://nrc-publications.canada.ca/eng/view/object/?id=10681e02-0c99-4b07-bad5-c57c87918e1b>

<https://publications-cnrc.canada.ca/fra/voir/objet/?id=10681e02-0c99-4b07-bad5-c57c87918e1b>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at

PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

Geographic Citation Gaps in NLP Research

Mukund Rungta^{✦♣}, Janvijay Singh^{✦♣}, Saif M. Mohammad[♡], Diyi Yang[◇]

[◇] School of Interactive Computing, Georgia Institute of Technology

[♡] National Research Council Canada

[◇] Stanford University

{mrungta8, iamjanvijay}@gatech.edu

saif.mohammad@nrc-cnrc.gc.ca

diyi@cs.stanford.edu

Abstract

In a fair world, people have equitable opportunities to education, to conduct scientific research, to publish, and to get credit for their work, regardless of where they live. However, it is common knowledge among researchers that a vast number of papers accepted at top NLP venues come from a handful of western countries and (lately) China; whereas, very few papers from Africa and South America get published. Similar disparities are also believed to exist for paper citation counts. In the spirit of “*what we do not measure, we cannot improve*”, this work asks a series of questions on the relationship between geographical location and publication success (acceptance in top NLP venues and citation impact). We first created a dataset of 70,000 papers from the ACL Anthology, extracted their meta-information, and generated their citation network. We then show that not only are there substantial geographical disparities in paper acceptance and citation but also that these disparities persist even when controlling for a number of variables such as venue of publication and sub-field of NLP. Further, despite some steps taken by the NLP community to improve geographical diversity, we show that the disparity in publication metrics across locations is still on an increasing trend since the early 2000s. We release our code and dataset here: <https://github.com/iamjanvijay/acl-cite-net>.

1 Introduction

Progress in science is accelerated by a sharing of ideas. However, there have been numerous instances in history where the predominance of one group of people in science, and the silencing of others, has led to the publication of harmful pseudoscience (Gould et al., 1996; Saini, 2019). Particularly egregious examples include the publication of theories and ideas on racial hierarchy (Plutzer,

2013), male superiority (Huang et al., 2020), gender binary (Darwin, 2017), and eugenics (Cottrol, 2015). It has also been shown that a lack of inclusion in invention and discovery leads to fewer technologies for the excluded group. For example, Koning et al. (2021) show how fewer technologies and health products are designed for women and Bender (2011), Bird (2020) and Mohammad (2019) show how a number of language technologies are designed for only a small number of languages.

In this paper, we explore geographic inclusion in Natural Language Processing (NLP) research. Our premise is that in a fair world, people have equitable opportunities to education, to conduct scientific research, and to publish, regardless of where they live. However, researchers in the field know that a vast number of papers accepted at top NLP conferences and journals come from a handful of western countries and (lately) China. On the other hand, very few papers with African and South American authors are published.

Further, the papers that get a majority of citations tend to be from a small number of institutions. Highly funded universities and research labs also tend to garner greater early visibility for their papers. Some of these papers might be cited more simply because the affiliate university or lab is perceived as prestigious (Amara et al., 2015; Hurley et al., 2013). Price (1965) examined the growth of citation networks and showed that papers with more early citations are likely to be cited more in the future (the “*rich get richer*” phenomenon).

Citations received by a research article serve as one of the key quantitative metrics to estimate its impact. Citations-based metrics, such as h-index (Hirsch, 2005; Bornmann and Daniel, 2009), can have a considerable impact on a researcher’s career, funding received, and future research collaborations. Citation metrics are also commonly taken into consideration in determining university rankings and overall scientific outcomes from a country.

[♣]Equal contribution.

Thus, the degree of equity in citations across geographic regions can act as one of the barometers of fairness in research. Furthermore, geographic location directly correlates to the languages spoken in an area. Therefore, to increase the reach of NLP beyond high-resource languages, it is important to elevate the research pursued in languages from these under-represented regions.

In this work, we investigate the impact of a researcher’s geographic location on their citability for the field of NLP. We examine tens of thousands of articles in the ACL Anthology (AA) (a digital repository of public domain NLP articles), and generate citation networks for these papers using information from Semantic Scholar, to quantify and better understand disparities in citation based on the geographic location of a researcher. We consider a set of candidate factors that might impact citations received and perform both qualitative and quantitative analyses to better understand the degree to which they correlate with high citations.

However, it should be noted that we do not explore the cause of citation disparities. Reasons behind such location-based disparities are often complex, inter-sectional, and difficult to disentangle. Through this work we aim at bringing the attention of the community to geographic disparities in research. We hope that work in this direction will inspire actionable steps to improve geographic inclusiveness and fairness in research.

2 Dataset

As of January 2022, the ACL Anthology (AA) had 71,568 papers.¹ We extracted paper title, names of authors, year of publication, and venue of publication for each of these papers from the repository. Further, we used information about the AA papers in Semantic Scholar² to identify which AA papers cite which other AA papers — *the AA citation network*. Since the meta-information of the papers in AA and Semantic Scholar does not include the affiliation or location of the authors, we developed a simple heuristic-based approach to obtain affiliation information from the text of the paper.

We refer to our dataset as the *AA Citation Corpus*. It includes the AA citation graph, author names, unique author ids (retrieved from Semantic Scholar), conference or workshop title, month and year of publication, and country associated with

the author’s affiliation. *We make the AA Citation Corpus freely available.*

Detailed steps in the construction of the citation network and the extraction of affiliated country information are described in the subsections below.

2.1 Citation Graph Construction

To create the citation graph, we collected the BibTeX entries of all the papers in the anthology. We filtered out the entries which were not truly research papers such as forewords, prefaces, programs, schedules, indexes, invited talks, appendices, session information, newsletters, lists of proceedings, etc. Next, we used Semantic Scholar APIs³ to identify unique Semantic Scholar ID (SSID) corresponding to each paper in the BibTeX. For this, we queried the Semantic Scholar APIs in two ways: (a) Using the ACL ID present in BibTeX, which ensures that correct SSID was retrieved for a paper in BibTeX; and (b) for papers whose SSID cannot be retrieved using ACL ID, we searched the paper using the paper title mentioned in BibTeX. In (b), to ensure correctness of the retrieved SSID, we take the fuzzy string matching score⁴ between title in BibTeX and that retrieved from Semantic Scholar. SSIDs with fuzzy score greater than 85% are marked as correct. For the remaining retrieved SSIDs, we manually compared the title in BibTeX and the one retrieved from Semantic Scholar.

We were able to retrieve correct SSIDs for **98.63%** of the papers in the ACL Anthology. Finally, we queried the Semantic Scholar APIs with the SSIDs of each of the AA papers to retrieve the SSIDs of the papers cited in the AA papers. With this information, we created the AA citation graph.

2.2 Country Information Extraction

We inferred the authors’ affiliated country from the textual information extracted from the research paper PDFs. We used SciPDF Parser⁵, a python parser for scientific PDF, to extract text from any PDF. Section-based parsing by this tool helps us to concentrate only on the header, which contains information about the author’s affiliation. The considerable differences in templates of papers published across different venues and years presented several challenges. We first compiled an exhaustive list of countries and their universities from the web.

¹<https://aclanthology.org/>

²<https://www.semanticscholar.org/>

³<https://www.semanticscholar.org/product/api>

⁴<https://pypi.org/project/fuzzywuzzy>

⁵https://github.com/titipata/scipdf_parser

# Countries	Number of papers
0: no country	14,818
1: one country	48,815
>1: multiple countries	7,062

Table 1: Count of papers by the number of automatically inferred affiliated countries.

(Details in Appendix C.) For each paper, we examine the affiliation section to identify mentions of a country (using our list of countries).⁶ Using this approach we were able to map each paper with its affiliated country. Table 1 shows the number of papers having n -country tags, where $n = \{0, 1, >1\}$ represents no country, one country and multiple countries respectively. Further, as the mapping of paper to the country was automatically constructed, the authors manually annotated the ground truth country tag for 1000 papers. This was done to analyze the correctness of the automatically identified country tags. These papers were selected at random from the dataset. Out of 1000 papers, country tags for 845 (84.5%) exactly match the ground truth. For most of the remaining unmatched cases, the algorithm either missed one country from the list or was unable to find any country tag for the paper.

3 Disparity in Citation based on Location

We use the *AA citation Corpus* to answer a series of questions on disparity of publications and citations across geographic location. We start with a look at the number of publications from around the world, followed by an examination of their citations.

Q1. Is there a disparity in the number of NLP publications across different countries? How does the amount of publications correlate with linguistic diversity?

A. We used counts of papers from the AA Citation Corpus to determine the number of papers from each country, as visualized in Figure 1. For an even coarser view, we also examined a partition of the world into ten regions.⁷ We calculated the total

⁶Even if an author has multiple affiliations (countries) we only consider the ones mentioned in the paper.

⁷One can partition the world map into regions in many ways. We made use of the partition provided by the United Nations Geo-scheme: https://en.wikipedia.org/wiki/List_of_countries_by_United_Nations_geoscheme. This list includes seventeen subregions, and we combine some of these subregions into ten coarser regions for simplicity.

number of papers from each region by aggregating papers from all countries present in this region. We also aggregate citation counts of papers by region.

Discussion Figure 1 shows huge disparities for the number of publications among countries. The western world which includes United States, Canada, United Kingdom, France, Germany, etc. dominates the network with high publication count. On the other hand, most countries in Africa, South America, Eastern Europe, South East Asia, and Middle East remain in the red zone with very few publications till date. When examining language diversity⁸ (indicated by size of yellow dot), we see that countries in the red zone have the highest language diversity. Higher linguistic diversity indicates larger number of different languages spoken in that geographic region. For example, the list of countries with the highest number of languages includes: Indonesia (710), Nigeria (524), India (453), and Brazil (228).⁹

More work on these languages is needed, by local researchers in partnership with the language communities. One recent effort in this regard is project Masakhane, a grassroots organisation whose mission is to strengthen and spur NLP research in African languages, for Africans, by Africans.¹⁰ This analysis showcases the huge disparity in the number of publication from each country. Through the questions ahead, we further uncover geographic patterns in citations, across these mid-tier and top-tier publishing countries.

Q2. How has the citation count ("influence" of NLP research) of papers from different regions changed over the years?

A. To study this question, we examine the following metric: mean citation count per paper for each country until certain year. Formally, this metric can be defined as follows:

$$MC_{(j,k)} = \frac{\sum_{i \in P_k} C_k(i) \mathbb{I}_{i \in j}}{\sum_{i \in P_k} \mathbb{I}_{i \in j}}$$

where $MC_{(j,k)}$ indicates mean-citation count of country- j until year- k . $C_k(i)$ indicates citation count of paper- i until year- k . $\mathbb{I}_{i \in j}$ is 1 if paper- i belongs to country- j otherwise 0. P_k indicates

⁸https://en.wikipedia.org/wiki/Linguistic_diversity_index

⁹https://en.wikipedia.org/wiki/Number_of_languages_by_country

¹⁰<https://www.masakhane.io/>

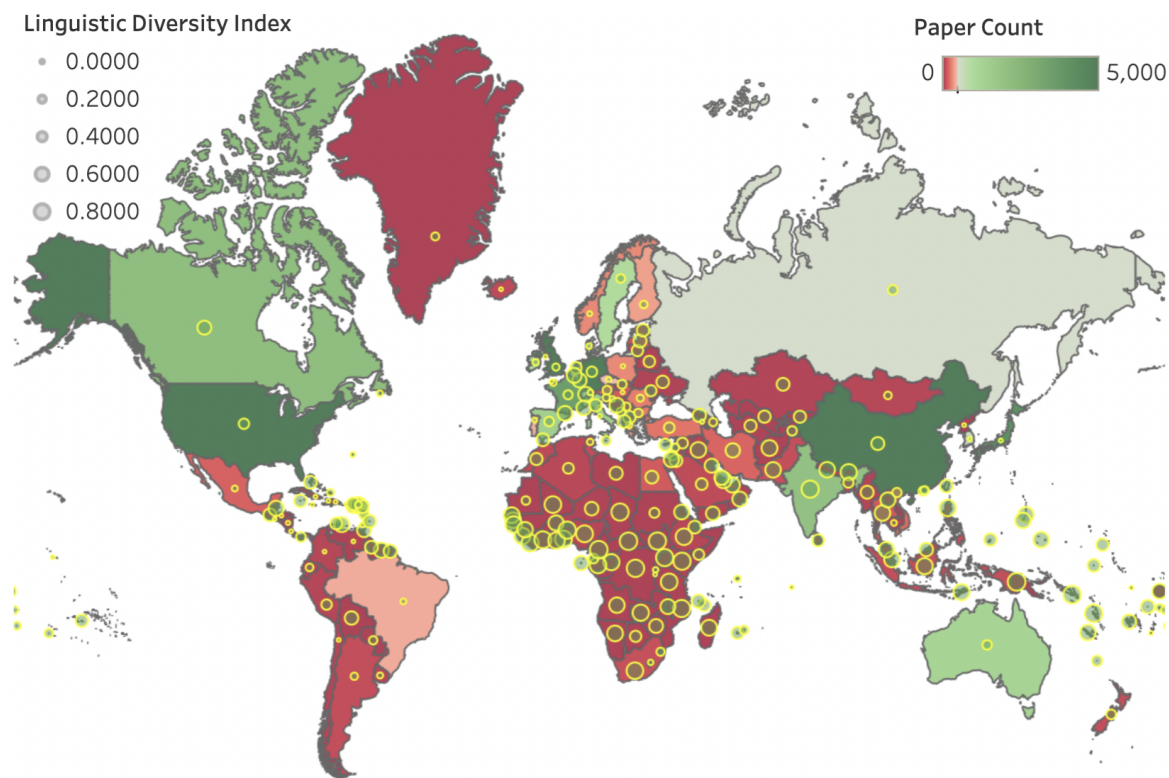


Figure 1: World-map showing variation of linguistic diversity (shown through circles of different sizes) and volume of NLP research (shown through color shades from red to green) across the globe.

the set of all the papers published until year- k . To calculate this metric, we create multiple citation networks each containing papers published only until a certain year. For instance, a citation graph associated with year-2000 will only contain papers published until 2000 — with papers as nodes and references as edges. Using such a citation network associated with year- k , we can then easily evaluate $MC_{(j,k)}$ for any country- j .

In Figure 2, we plot the total citation count received by all the papers published by different regions of the world across a span of ~ 21 years from year 2000 to 2021.¹¹ Similarly, in Figure 3, we plot the mean-citation metric for top-10 countries across these years. We restrict the analysis to just the top-10 countries for the sake of simplicity. Also, note that in Figures 2 and 3, for each point in the x-axis, the corresponding point in the y-axis pertains to the citation graph composed of papers published up till that year.

To understand how the influence of a paper changes across the years after its publication, for each country, we calculate the average number of

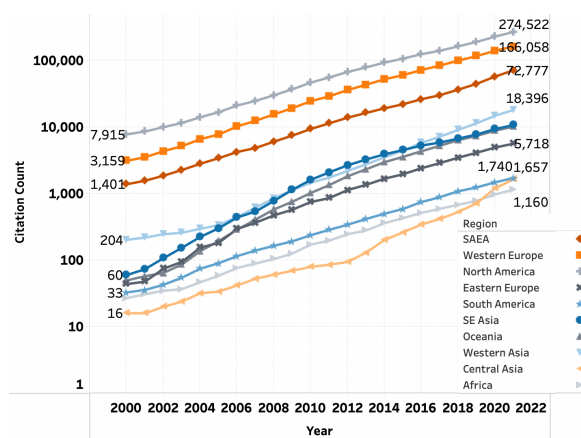


Figure 2: Citation count for each world region. Note the log scale for the y axis. SE Asia: South Eastern Asia; SAEA: South Asia and Eastern Asia.

citations its papers receive one year after publication, two years after publication, and so on. We will refer to this time span as the age of a paper. We do so by building citation networks for each year from 2000 till 2021 and aggregate the number of papers and citations for each age-value (no. of years since the paper was published). Figure 4 shows the plot of average citation vs. age-of-paper.

¹¹We use shapes and colour-blind friendly palettes in all the figures to ensure that the figures are *colour-blind friendly*. An *interactive version* of all figures is accessible in Appendix B.

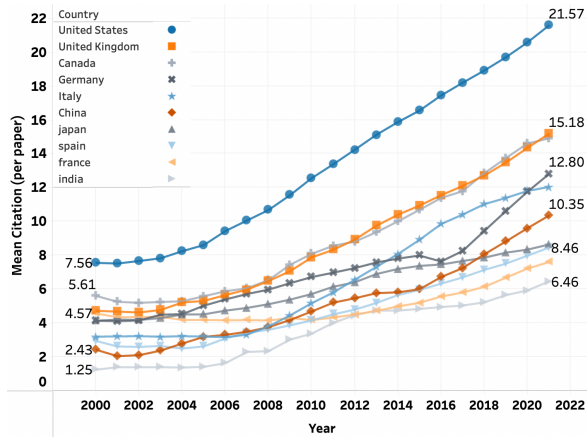


Figure 3: Variation of mean citations (per publication) across time for the top-10 publishing countries.

Discussion Figure 2 clearly depicts the huge disparities existing in citations received by different regions. Further, the trends have largely remained the same over the years, with the notable exception that central asia has made moderate gains since 2013, whereas the gains made by south east asia around the 2010s, have diminished since about 2012.

From Figure 3, we make following observations: *Firstly*, the plot shows that US (by a large margin), as well as, UK and Canada dominate the mean citation metric across the considered time-span. *Secondly*, for these three countries as well as for Germany, the growth-rate of the mean-citation metric is remarkably higher than that of other countries. This growth-rate is indicated by the slope of the curve. *Thirdly*, China and Germany have shown a significant increase in growth rate over the past eight years. *Overall*, this plot highlights that papers from certain countries have had markedly higher citations (on average) compared to other countries, across the years. Additionally, we also note that the gap in mean-citation metric across the top-10 countries has more than doubled (increasing from 6.31 to 15.11) in the past 21 years. This also indicates that much of the "influential" research in NLP from the past two decades is heavily concentrated in US, UK, and Canada.

We also explore median citation statistics in Figure 11 (Appendix A). In contrast to the mean, difference between medians is relatively smaller (for instance, 6 for US vs 3 for India in year 2021). This suggests that countries such as US, UK, and Canada have more very highly cited papers that push up their means. We also observe that the median has increased faster for countries like US, UK,

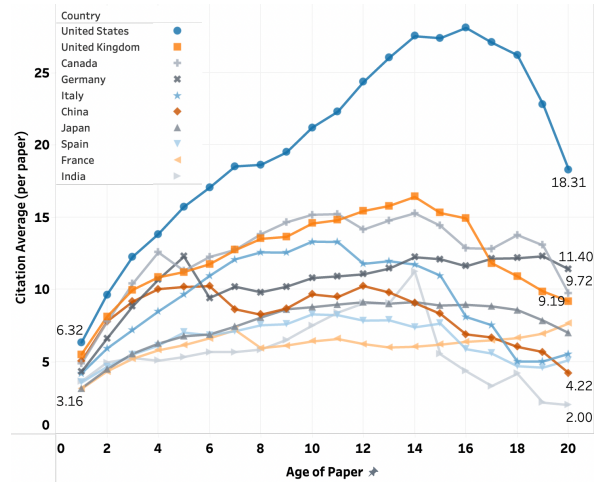


Figure 4: Variation of mean citations (per paper) across age of publication for the top-10 publishing countries.

and Canada compared to India, Spain, and France. Although the differences in medians are smaller, they are suggestive of marked citation disparity among the publication from different countries.

Figure 4 shows that in the initial years after a paper's publication (1–2 years), the gap in mean-citation between the most and least cited top-10 countries is ~ 3.5 , but it explodes to ~ 23 after 15 years. This suggest that, even among the top-publishing countries, the influence of some countries (like India and Spain) is surprisingly short-lived. Whereas, papers from US, UK, and Canada rapidly gain influence as the paper ages and their influence is long lasting.

Q3. How do countries cite each other? What inter-country citation patterns contribute to citation disparities?

A. We analyze the following metric for all publications until year 2021: fraction of references from country- j in an average paper from country- k . We will call it the *citation fraction*. Mathematically, it is defined as follows:

$$F(k, j) = \frac{\sum_{i \in P} \sum_{r \in R(i)} \mathbb{I}_{r \in j} \mathbb{I}_{i \in k}}{\sum_{i \in P} |R(i)| \mathbb{I}_{i \in k}}$$

where $R(i)$ indicates the set of all the references of paper- i , $\mathbb{I}_{r \in j}$ is 1 if reference- r (a paper) belongs to country- j otherwise 0, P is the set of all the papers. Intuitively, $F(k, j)$ indicates the average fraction of references from country- j in a paper from country- k . We compute the metric $F(k, j)$ from our citation graph with papers as nodes and references as

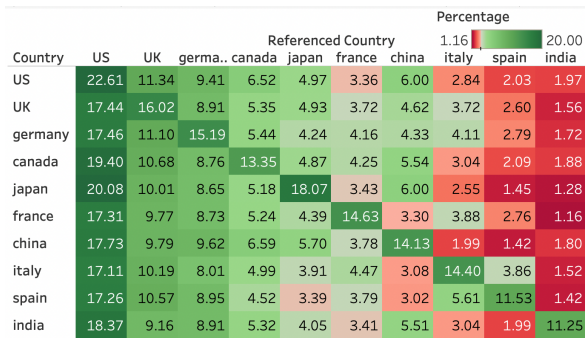


Figure 5: Heatmap depicting inter-country citation spread among top-10 publishing countries, until 2021.

edges. In Figure 5, we plot $F(k, j)$ (times 100 or %) to study how a country contributes to the citation count of another country. Here, country- k is shown in each row, with country- j in each column. Moreover, we also study how country- k 's citation fractions' $F(k, j)$ dispersion has changed over years. To study this, we plot Gini-coefficient (Dorfman, 1979) for each each country for each year from 2000. Again, we restrict the plot to the top-10 publishing countries for simplicity.

Discussion We make following observations from Figure 5. *Firstly*, for each country "self-citation" seems to be the highest contributor. In other words, a paper's most citations are from its' original country. *Secondly*, all the countries cite US, UK, Germany, and Canada, which is shown by high density in columns corresponding to US (18.47%), UK (10.86% avg.), Germany (9.51%) and Canada (6.25%). These convey that not only "self-citation" but significantly larger citation from all the other countries contribute to higher citation statistics for papers from some countries (like US, UK, Germany, and Canada). Note that the sum of all the rows for a country is ~ 71 . If all the referenced countries were plotted, then these row-sums would have resulted in a perfect 100. This means that $\sim 29\%$ of citations from the top-10 publishing countries are carried to the rest of the world.

In Figure 6, we plot the Gini-coefficient for these citation fractions, to show how the spread of citations has changed over years. A decreasing value of Gini-coefficient indicates that the spread is becoming closer to equal (a proxy for research-inclusivity and equitable opportunities). The Gini-coefficient of citation-fraction drops almost linearly for all top-10 countries. Specifically, the average of Gini-coefficients drops from 0.5985 to 0.3887 from 2000 to 2021. This is encouraging as a decrease in Gini-

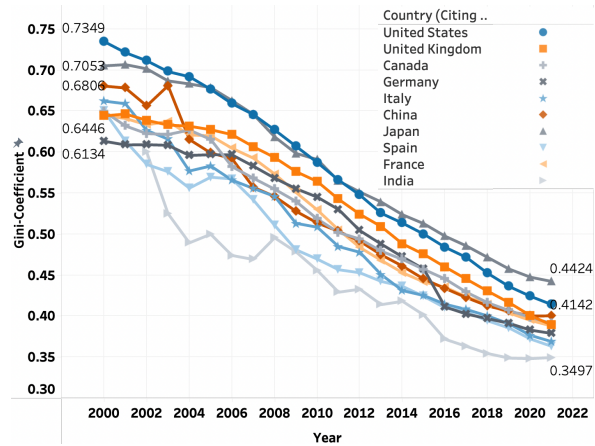


Figure 6: Gini-coefficient depicting dispersion of citation fractions among top-10 publishing countries.

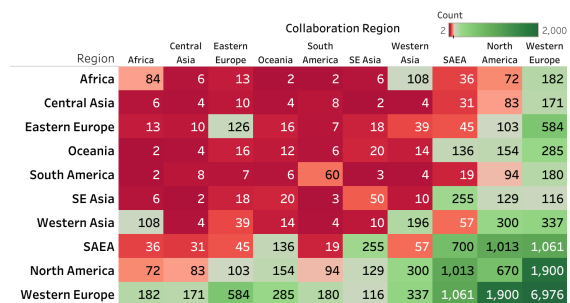


Figure 7: Heatmap indicating the collaboration between two regions with number of papers till 2021. SE Asia: South Eastern Asia; SAEA: South Asia and Eastern Asia.

coefficient indicates inclusivity in research. A linear interpolation, suggests that ideally it would still take ~ 39 years to reach zero for these Gini-coefficients, if the rate of change stays the same.

Q4. How does geographic location drive the collaboration between researchers? How has this changed over time?

A. For this analysis, we choose to measure inter-country collaboration, i.e., only papers with authors from more than one country are considered. Figure 7 shows the heatmap with the count of the number of papers for all pairs of regions. This count indicates the number of papers belonging to the collaboration between the countries from a given region pair. Note that 'collaboration between the same region' includes only those papers having a collaboration between different countries in that region, and papers with all authors from the same country are excluded from this analysis.

Again, we use the Gini coefficient to quantitatively account for the collaboration between re-

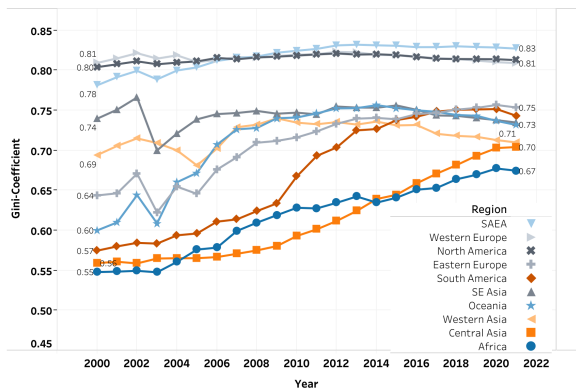


Figure 8: Trend of Gini-coefficient depicting collaboration among different regions. SE Asia: South Eastern Asia & SAEA: South Asia and Eastern Asia.

searchers across geographic locations. This coefficient will be close to 0 if the regions collaborate roughly equally with other regions, and close to 1 for a very selective collaboration. We computed the Gini coefficient across time; and, to determine the coefficient for a given year, we only considered the papers published up till that year. Figure 8 shows the trend of Gini coefficient over years.

Discussion We can make three notable observations from the heatmap of collaborations between continents (Figure 7). Firstly, different regions of Europe collaborate well with each other, resulting in Europe having the highest inter-continent collaboration. Collaboration in North America can be mostly attributed to the collaboration between the United States and Canada. Secondly, North America has more collaboration with Europe (2,003) than Asia (1,525), indicating the preferential disparity of North American researchers towards Europe and Asia to some extent. Thirdly, South America and Africa have very few research collaborations with the rest of the world; within this limited collaboration, these regions have higher collaboration with Asia and Europe than with North America. These results highlight the dire need for a wider research collaboration.

The Gini coefficient estimating the disparity in collaboration across geographic regions is shown in Figure 8. From 2000 to 2021, Gini coefficient for Africa rose from 0.37 to 0.53 and for South America, it rose from 0.41 to 0.61. Wide and meaningful collaboration is crucial for exchange of ideas and learning of new skills. Thus this decreased collaboration with researchers from Africa and South America is concerning.

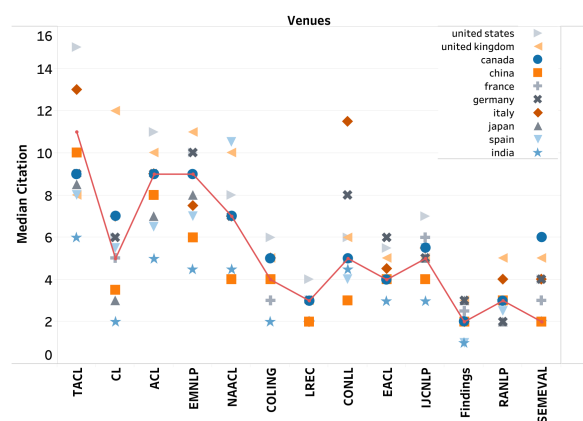


Figure 9: Median citation of papers published by different countries across different venues

Q5. How do the citation statistics of different countries vary across venues?

A. The venue of publication is believed to correlate somewhat to the quality of the publication and its impact. Very roughly speaking, papers published in a venue would be of similar quality and thus may be expected to be cited similarly. For example, papers published in journals, such as TACL, would be very similar to other papers in the venue, versus a paper published in a workshop. Here, we study citations of countries across venues to determine whether citation gaps exist even when one controls for the venue of publication.

We computed the median citation received by all papers published in a venue grouped by the country of publication. We choose the median as the measurement metric because it is less influenced by extremely large citation counts. Figure 9 shows the plot for the top-10 countries.

Discussion We observe that the United States is above this median line (indicated by a solid red line) for all 13 of 13 venues and the same can be said for the United Kingdom (10 of 13) and Germany (6 of 13). On the other hand, Japan is below the median line for 12 of 13 venues, and China and India are below the median line for all the venues. Even though the later countries publish fairly in all these venues, they receive significantly low citations compared to their peers. Also Figure 9 brings out the point that median citation for conferences like ACL, EMNLP, NAACL is significantly higher than Findings, RANLP, SEMEVAL. Overall this analysis shows the huge disparity in the citation received by the median paper from different coun-

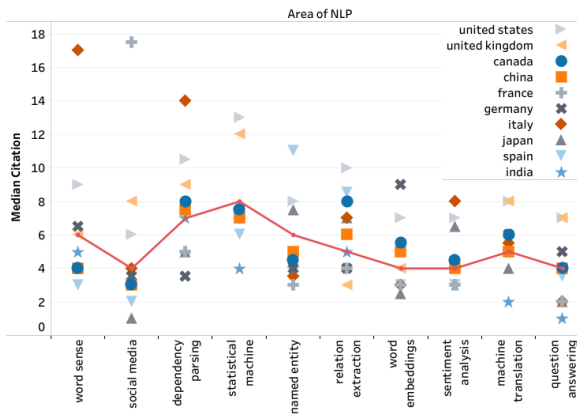


Figure 10: Median citation of papers published by different countries across different areas of NLP

tries published in the same venue. Thus geographic citation gaps exist even when we control for the venue of publication.

Q6. Is disparity in citation statistics consistent across research areas within NLP? Or is the gap simply because some countries work in areas that receive low numbers of citations (overall)?

A. In order to sample papers from various sub-field/areas of NLP, we consider the word bigrams from the title of the paper. We follow the same approach used by [Mohammad \(2020\)](#). We consider top-10 bigrams based on the number of papers sampled for each bigram, each representing an area of NLP. Figure 10 shows the median citation received by the publications from different countries across different areas of NLP. The median citation received by all papers published in an area is represented by a solid red line in Figure 10.

Discussion: Here we aim to understand whether low-cited countries work mostly in the areas that are less-cited. Conversely, we can also argue that highly cited countries mostly work in the areas that receive a high citation. From Figure 10 we can see that the United States is above the median line for all 10 areas, the United Kingdom for 5 of 10. On the other hand, countries like China (3 of 10), India (1 of 10), and Japan (3 of 10) stay below the median line for the majority of areas. In order to fairly compare the difference in median citations among countries, we consider the number of papers published by any country in a particular area. Table 2 compares the publication count of the United States and China for 3 areas of NLP and their median citation. China has a fairly similar number of

Area	United States		China	
	Count	Median	Count	Median
Sentiment Analysis	82	7.0	65	4.0
Dependency parsing	96	10.5	68	7.5
Relation Extraction	100	10.0	81	6.0

Table 2: Total papers and mean citation statistics for US and China across areas of research.

publications for these areas, but the median citation is significantly lower than the United States. This highlights the disparity in citations received by papers based on their geographic location.

4 Related Work

Bias in citation networks is a long standing problem that arises by a combination of a myriad number of factors. [Tahamtan et al. \(2016\)](#) categorised these factors into three buckets: firstly, paper related factors which involves paper quality ([Bucla-Casal and Zych, 2010](#)), field of study ([Costas et al., 2009](#)), novelty, length of paper ([Antonioni et al., 2015](#); [Falagas et al., 2013](#)); secondly, journal/conference related factors, like impact factor ([Callaham et al., 2002](#)), language of publication ([Lira et al., 2013](#)); and thirdly, author related factors, like number of authors ([Della Sala and Brooks, 2008](#); [Bosquet and Combes, 2013](#)), collaboration ([Nomaler et al., 2013](#)), self-citation ([Costas et al., 2010](#)) author’s reputation ([Collet et al., 2014](#)), affiliation ([Sin, 2011](#)), gender, race and age ([Ayles and Vars, 2000](#); [Leimu and Koricheva, 2005](#)).

[Nielsen and Andersen \(2021\)](#) analyzed a few subfields of science to show that the concentration of citation by the top 1% of the authors increased from 14 to 21% between 2000 and 2015 and that the Gini coefficient for citation imbalance has risen from 0.65 to 0.70. They also observed that these top 1% authors mostly reside in Western Europe and Australasia. Similarly, [Lee et al. \(2010\)](#) and [Pasterkamp et al. \(2007\)](#) presented a similar hypothesis that researchers from USA and UK are cited much more than rest of the world. This finding is well established across different disciplines of science. Using correlation analysis, [Lou and He \(2015\)](#) have shown that there is a high negative correlation between the author’s affiliation and uncitedness. They also show that along with the affiliation, the venue of publication greatly impacts the citation received by any paper ([Callaham et al., 2002](#)). [Schluter \(2018\)](#) shows, using a mentor-

mentee network, that there is a significant time gap for the female researchers to acquire the mentorship status. Vogel and Jurafsky (2012); Mohammad (2020) measures the gender bias existing in the NLP community by considering papers published in ACL Anthology. However, Mohammad (2020) considers a larger domain to examine the gender bias specifically from the point of view of female first and last authors.

In the broader science community, extensive research has shown the existence of gender bias in citation network and its impact on the career trajectory of female researchers (Chatterjee and Werner, 2021; Llorens et al., 2021). Although research is done in quantifying the citation disparity across geographic location for other fields (Paris et al., 1998; Nishioka et al., 2022), not many studies have systematically studied geographic gap in NLP, and our work makes an original contribution in analyzing this disparity at scale in the NLP community.

5 Conclusion

We compiled a large dataset of meta-information associated with NLP papers, that includes the country name associated with an author’s affiliated institution, number of citations, and the papers cited by each paper. We use this dataset to systematically examine citation disparities across geographic regions. We show that substantial geographic disparities exist, even when controlling for factors such as venue of publication. Authors from North America and Western Europe have published 65% of the total number of papers; whereas, authors from Africa, South America, and South-Eastern Asia (combined) have published only 3.7% of the total papers. We also find markedly lower levels of cross-region author–author collaboration across the two sets of regions. Even among the top-ten publishing countries, we find marked differences. Very little research from countries such as India, Spain, Italy, and China is cited by their peers from the top-five publishing countries.

Citation gaps across genders and geographic regions occur due to various complex, intersectional, and structural reasons. This paper does not explore the reasons behind them. However, Pan et al. (2012); Skopec et al. (2020); Wuestman et al. (2019) have examined some of the causes for geographic citation gaps in scientific literature such as self citation, co-location of institutions, and national funding for research & development. The

above reasons are not exhaustive and determining the degree of impact of each reason is extremely difficult. Further, the reasons can be different across different regions of the world. We hope that future qualitative research sheds greater light on these complexities. Understanding them is crucial for greater inclusion in research. We also hope that such work will engender concrete actions to address citation gaps. Initiatives taken by the NLP community toward diversity, inclusion, and collaboration, especially actions that encourage local NLP research across the globe, can make meaningful change and help address geographic inequities.

6 Ethics Statement

Since the attribution of paper to countries is done through an automatic process, it is possible that a small number of attributions might suffer from noisy labels. Further, the only inferences drawn in this work are at aggregate-level, and not about individual papers or authors.

7 Limitations

ACL Anthology does not provide the affiliation (or some location information) associated with the authors. Since our work majorly relies on authors’ location, we came up with a heuristic approach (described in Section 2) to automatically identify the country of the authors’ associated with each research paper. Additionally, we also analyzed the correctness of automatically identified country tags. In our analysis, we found that for 84.5% of the papers country tags were perfectly retrieved. For the remaining unmatched papers, the algorithm either missed one country from the list or was unable to find any country tag for the paper. Since we use an exhaustive list of countries and universities, these retrieval errors are not country-specific. But rather these errors are specific to the formatting of the research papers, which is dependent on the publication venue. In order to facilitate future research, we suggest that ACL should make location-associated meta-information of the authors publicly available. We also suggest that conferences and journals should publicly report summary statistics of geographic diversity of accepted papers.

In this work, we only used the papers published in the ACL Anthology. These are usually papers from large international conferences, and always written in English. However, it should be noted that there exist vibrant local communities that publish at

local venues, often in non-English languages. Furthermore, NLP research is also published at other non-ACL global venues, such as AAAI, ICLR, ICML, WWW. The conclusions drawn from our experiments therefore apply to international English NLP conferences and journals, and further work is needed to explore the landscape of local sub-communities across the world.

Citations are heterogeneous and can be categorized in different ways. For example, semantic scholar categorizes citations into background, method, and result citations. In the current work, we do not distinguish between types of citations.

The first-author position in a paper is usually reserved for the researcher that has done the most work and writing. The last author position is often reserved for the most senior or mentoring researcher. However, this work does not explore the distributions of these positions by region.

Finally, even though this work only looks at published papers, future work can also examine rejection rates by geographic location (if that information becomes accessible). See Church (2020) for a discussion on improving reviewing, and how organizers of early EMNLP conferences (in the early 2000s) actively accepted papers from Asia at time when that was less common. Church (2020) argue that this was one of the reasons for the success of EMNLP and its reputation as a conference that is more accepting of new and diverse ideas.

References

- Nabil Amara, Réjean Landry, and Norrin Halilem. 2015. What can university administrators do to increase the publication and citation scores of their faculty members? *Scientometrics*, 103(2):489–530.
- George A Antoniou, Stavros A Antoniou, Efstratios I Georgakarakos, George S Sfyroeras, and George S Georgiadis. 2015. Bibliometric analysis of factors predicting increased citations in the vascular and endovascular literature. *Annals of vascular surgery*, 29(2):286–292.
- Ian Ayres and Fredrick E Vars. 2000. Determinants of citations to articles in elite law reviews. *The Journal of Legal Studies*, 29(S1):427–450.
- Emily M. Bender. 2011. [On achieving and evaluating language-independence in nlp](#). *Linguistic Issues in Language Technology*, 6.
- Steven Bird. 2020. [Decolonising speech and language technology](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Lutz Bornmann and Hans-Dieter Daniel. 2009. The state of h index research: Is the h index the ideal way to measure research performance? *EMBO reports*, 10(1):2–6.
- Clément Bosquet and Pierre-Philippe Combes. 2013. Are academics who publish more also more cited? individual determinants of publication and citation records. *Scientometrics*, 97(3):831–857.
- Gualberto Buela-Casal and Izabela Zych. 2010. Analysis of the relationship between the number of citations and the quality evaluated by experts in psychology journals. *Psicothema*, pages 270–276.
- Michael Callaham, Robert L Wears, and Ellen Weber. 2002. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *Jama*, 287(21):2847–2850.
- Paula Chatterjee and Rachel M Werner. 2021. Gender disparity in citations in high-impact journal articles. *JAMA Network Open*, 4(7):e2114509–e2114509.
- Kenneth Ward Church. 2020. [Emerging trends: Reviewing the reviewers \(again\)](#). *Natural Language Engineering*, 26(2):245–257.
- François Collet, Duncan A Robertson, and Daniela Lup. 2014. When does brokerage matter? citation impact of research teams in an emerging academic field. *Strategic Organization*, 12(3):157–179.
- Rodrigo Costas, Maria Bordons, Thed N Van Leeuwen, and Anthony FJ Van Raan. 2009. Scaling rules in the science system: Influence of field-specific citation characteristics on the impact of individual researchers. *Journal of the American Society for Information Science and Technology*, 60(4):740–753.
- Rodrigo Costas, Thed van Leeuwen, and María Bordons. 2010. Self-citations at the meso and individual levels: effects of different calculation methods. *Scientometrics*, 82(3):517–537.
- Robert J. Cottrol. 2015. [The Myth of Race: The Troubling and Persistence of an Unscientific Idea](#). By Robert Wald Sussman. *Journal of Social History*, 49(3):740–741.
- Helana Darwin. 2017. Doing gender beyond the binary: A virtual ethnography. *Symbolic Interaction*, 40(3):317–334.
- Sergio Della Sala and Joanna Brooks. 2008. Multi-authors’ self-citation: A further impact factor bias? *Cortex; a journal devoted to the study of the nervous system and behavior*, 44(9):1139–1145.
- Robert Dorfman. 1979. A formula for the gini coefficient. *The review of economics and statistics*, pages 146–149.

- Matthew E Falagas, Angeliki Zarkali, Drosos E Karageorgopoulos, Vangelis Bardakas, and Michael N Mavros. 2013. The impact of article length on the number of future citations: a bibliometric analysis of general medicine journals. *PLoS One*, 8(2):e49476.
- Stephen Jay Gould et al. 1996. *The Mismeasure of Man*. WW Norton & company.
- Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.
- Junming Huang, Alexander J Gates, Roberta Sinatra, and Albert-László Barabási. 2020. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, 117(9):4609–4616.
- Lori A Hurley, Andrea L Ogier, and Vetle I Torvik. 2013. Deconstructing the collaborative impact: Article and author characteristics that influence citation count. *Proceedings of the American Society for Information Science and Technology*, 50(1):1–10.
- Rembrand Koning, Sampsa Samila, and John-Paul Ferguson. 2021. Who do we invent for? patents by women focus more on women's health, but few women get to invent. *Science*, 372(6548):1345–1348.
- Shi Young Lee, Sanghack Lee, and Sung Hee Jun. 2010. Author and article characteristics, journal quality and citation in economic research. *Applied Economics Letters*, 17(17):1697–1701.
- Roosa Leimu and Julia Koricheva. 2005. What determines the citation frequency of ecological papers? *Trends in ecology & evolution*, 20(1):28–32.
- Rodrigo Pessoa Cavalcanti Lira, Rafael Marsicano Cezar Vieira, Fauze Abdulmassih Gonçalves, Maria Carolina Alves Ferreira, Diana Maziero, Thais Helena Moreira Passos, and Carlos Eduardo Leite Arieta. 2013. Influence of english language in the number of citations of articles published in brazilian journals of ophthalmology. *Arquivos Brasileiros de Oftalmologia*, 76:26–28.
- Anaïs Llorens, Athina Tzovara, Ludovic Bellier, Ilina Bhaya-Grossman, Aurélie Bidet-Caulet, William K Chang, Zachariah R Cross, Rosa Dominguez-Faus, Adeen Flinker, Yvonne Fonken, et al. 2021. Gender bias in academia: A lifetime problem that needs solutions. *Neuron*, 109(13):2047–2074.
- Wen Lou and Jianguan He. 2015. Does author affiliation reputation affect uncitedness? *Proceedings of the Association for Information Science and Technology*, 52(1):1–4.
- Saif M. Mohammad. 2019. The state of nlp literature: A diachronic analysis of the acl anthology. *arXiv preprint arXiv:1911.03562*.
- Saif M. Mohammad. 2020. **Gender gap in natural language processing research: Disparities in authorship and citations**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7860–7870, Online. Association for Computational Linguistics.
- Mathias Wullum Nielsen and Jens Peter Andersen. 2021. Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences*, 118(7):e2012208118.
- Chifumi Nishioka, Michael Färber, and Tarek Saier. 2022. How does author affiliation affect preprint citation count? analyzing citation bias at the institution and country level. *arXiv preprint arXiv:2205.02033*.
- Önder Nomaler, Koen Frenken, and Gaston Heimeriks. 2013. Do more distant collaborations have more citation impact? *Journal of Informetrics*, 7(4):966–971.
- Raj Kumar Pan, Kimmo Kaski, and Santo Fortunato. 2012. World citation and collaboration networks: uncovering the role of geography in science. *Scientific reports*, 2(1):1–7.
- Gianmarco Paris, Giulio De Leo, Paolo Menozzi, and Marino Gatto. 1998. Region-based citation bias in science. *Nature*, 396(6708):210–210.
- Gerard Pasterkamp, Joris Rotmans, Dominique de Kleijn, and Cornelius Borst. 2007. Citation frequency: A biased measure of research impact significantly influenced by the geographical origin of research articles. *Scientometrics*, 70(1):153–165.
- Eric Plutzer. 2013. The racial gap in confidence in science: Explanations and implications. *Bulletin of Science, Technology & Society*, 33(5-6):146–157.
- Derek J De Solla Price. 1965. Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, 149(3683):510–515.
- Angela Saini. 2019. *Superior: the return of race science*. Beacon Press.
- Natalie Schluter. 2018. The glass ceiling in nlp. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2793–2798.
- Sei-Ching Joanna Sin. 2011. International coauthorship and citation impact: A bibliometric study of six lis journals, 1980–2008. *Journal of the American Society for Information Science and Technology*, 62(9):1770–1783.
- Mark Skopec, Hamdi Issa, Julie Reed, and Matthew Harris. 2020. The role of geographic bias in knowledge diffusion: a systematic review and narrative synthesis. *Research integrity and peer review*, 5(1):1–14.

Iman Tahamtan, Askar Safipour Afshar, and Khadijeh Ahamdzadeh. 2016. Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, 107(3):1195–1225.

Adam Vogel and Dan Jurafsky. 2012. *He said, she said: Gender in the ACL Anthology*. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea. Association for Computational Linguistics.

Mignon L Wuestman, Jarno Hoekman, and Koen Frenken. 2019. The geography of scientific citations. *Research Policy*, 48(7):1771–1780.

Appendix

A Supplementary Plot

Variation of median citations (of all publication) across time for the top-10 publishing countries is shown in Figure 11.

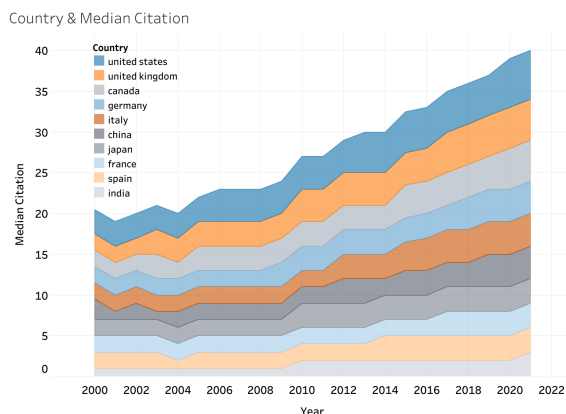


Figure 11: Variation of median citations (of all publication) across time for the top-10 publishing countries. (height of strip denotes median value)

B Interactive Plots

An interactive version for all the plots presented in this work can be accessed from following Tableau Public profile: <https://public.tableau.com/app/profile/gdnlp>

C Country Information Extraction

Below we describe the methodology used to extract country tag for each paper.

1. Text extraction: We used scipdf-parser to extract the header information from the pdf document of the given research paper.
2. Extracting country and organization tag: Text from the header is retrieved in JSON format

where a map between country and organization is established by this tool. For some papers, the tool is not able to match the country tag to the appropriate country either because it is unable to parse it correctly or because the information is not present at all. Therefore, we use the value present in the organization tag to map the paper to the respective country.

3. Preprocessing: We remove all the non-alphanumeric characters from the extracted text from the last step. This is done to clean the text and perform the next step of country extraction easier.
4. University list: Papers for which the country is not available, we perform matching the organization name to its affiliated country. We created a list of all universities of the list by following some earlier works. We mined this page¹² that contains a consolidated list of universities all across the globe. We augment this list with the University list from this link¹³ as well. Thus, we finally got a list of around 18k universities and their corresponding country. Next, text matching is performed on the extracted organization name with the university list to get the country associated with the paper.
5. Country List: In a few cases the tool was unable to extract the country information even though it was present. Therefore, to solve this we curate a list of all countries across the globe. We perform direct text matching of the extracted text with the country list and map the paper to its corresponding country.

D Country to Region Mapping

We aggregate information from different countries into regions. This country to region mapping is chosen so as the countries included in a region are in geographical vicinity. In this mapping, we have considered different variations of country names in the affiliations. We consider these multiple variations so as to retrieve as many country mentions as possible from each paper.

- **South Asia and Eastern Asia (SAEA):** afghanistan; bangladesh; bhutan; china; hong

¹²<https://www.4icu.org/reviews/index2.htm>

¹³<https://raw.githubusercontent.com/endSly/world-universities-csv/master/world-universities.csv>

kong; india; iran; japan; korea; maldives; mongolia; nepal; north korea; pakistan; sri lanka; taiwan; vietnam

- **Oceania:** australia; fiji; kiribati; marshall islands; micronesia; nauru; new zealand; palau; papua new guinea; samoa; solomon islands; tonga; tuvalu; vanuatu
- **South Eastern Asia (SE Asia):** brunei; cambodia; east timor; indonesia; malaysia; myanmar; philippines; singapore; thailand
- **Eastern Europe:** belarus; bulgaria; cz; czech republic; czechia; hungary; moldova; poland; romania; russian federation; slovakia; ukraine
- **Africa:** algeria; angola; benin; botswana; burkina; burundi; cameroon; cape verde; central african republic; chad; comoros; congo; congo, democratic republic of; djibouti; egypt; equatorial guinea; eritrea; ethiopia; gabon; gambia; ghana; guinea; guinea-bissau; ivory coast; kenya; lesotho; liberia; libya; madagascar; malawi; mali; mauritania; mauritius; morocco; mozambique; namibia; niger; nigeria; rwanda; sao tome and principe; senegal; seychelles; sierra leone; somalia; south africa; south sudan; sudan; swaziland; tanzania; togo; tunisia; uganda; zambia; zimbabwe
- **North America:** antigua and barbuda; bahamas; barbados; belize; canada; costa rica; cuba; dominica; dominican republic; el salvador; grenada; guatemala; haiti; honduras; jamaica; mexico; nicaragua; panama; saint kitts and nevis; saint lucia; saint martin; saint vincent and the grenadines; trinidad and tobago; united states; us
- **South America:** argentina; bolivia; brazil; chile; colombia; ecuador; guyana; paraguay; peru; suriname; uruguay; venezuela
- **Central Asia:** kazakhstan; kyrgyzstan; russia; tajikistan; turkmenistan; uzbekistan
- **Western Asia:** armenia; azerbaijan; bahrain; cyprus; georgia; iraq; israel; jordan; kuwait; lebanon; oman; qatar; saudi arabia; syria; turkey; united arab emirates; yemen
- **Western Europe:** albania; andorra; austria; belgium; bosnia and herzegovina; croatia; denmark; estonia; finland; france; germany;

greece; iceland; ireland; italy; latvia; liechtenstein; lithuania; luxembourg; macedonia; malta; monaco; montenegro; netherlands; norway; portugal; san marino; serbia; slovenia; spain; sweden; switzerland; united kingdom; vatican city