Citation Analysis Disparity Between Sub-Areas of Brazilian Computer Science

Fernando F. Druszcz[1], **André L. Vignatti**[1]

[1]Department of Computer Science
Federal University of Paraná (UFPR)
Curitiba – PR – Brazil

{ffd18,vignatti}@inf.ufpr.br

Abstract. *Among the various ways of evaluating scientific production, there is a tendency to use metrics based on the number of citations. Apart from obvious problems, this takes on a new dimension when it is used to compare areas and sub-areas, specially from unfair assessments if submitted to the same evaluation committee. In this work, we examine various sub-areas of Computer Science using data from the Brazilian community. Our findings reveal a disparity in citations among these sub-areas, which may lead to issues if they are evaluated using the same criteria for scientific productivity. We demonstrate how the* universal fit citation*, previously proposed by Radicchi et al., can help mitigated these concerns.*

1. Introduction

It is an empirical known fact that some scientific areas receive more citations than others. For example, in the 100 most cited works until 2014 [Noorden et al. 2014], we see that productions in "biology laboratory techniques", besides being the first six most cited, appear much more than productions from other areas. The issue here is twofold: the comparison between different areas and the use of the number of citations as a means of comparison.

Due to their unique natures, comparing different areas can be irrelevant or meaningless [Radicchi et al. 2008]. Large areas are hardly subject to such comparisons, as there are specific committees for each of them. On the other hand, sub-areas are subordinated to the same committee, which establishes common criteria. In this case, comparisons are often made, and this can be a source of problems.

Examining more than 53 million writers and nearly 90 million scientific papers across all disciplines reveals an exponential growth in the number of papers and scientists over the last hundred years [Dong et al. 2017]. Thus, a wide variety of models and metrics were proposed to assess the quality of scientific output. One thing in common with most used methods is the fact that they all revolve around the *number of citations* [Wang and Barabási 2021]. This is not without criticism, as it may provide an unfair comparison due to factors such as one-hit wonders and researchers who are productive but not necessarily impactful [Wang and Barabási 2021]. Although there are alternative metrics that try to grasp the quality of a publication, the only quantitative factor that somewhat displays the interest of the community in an article is the number of citations. Metrics that rely on the number of citations generally operate at the author level, but are often extrapolated to compare areas and sub-areas. Such extrapolation can be even more problematic than reducing the impact of articles and researchers to a single metric.

1.1. Our Results

We analyze different sub-areas of Computer Science, with the data coming from the Brazilian community. First, in Section 2, we discuss the typical behavior of scientific citations, which adheres to a power-law distribution. This is significant because it demands an analysis that relies not on conventional statistical metrics, but rather on a complementary cumulative distribution function. We use data from three sources: CSIndexBr [Valente and Paixao 2018], DBLP [Bibliography 2022] and OpenCitations [Waltman et al. 2020]. How these data were combined is further detailed in Section 3. In Section 4, we present the existing disparity between citations in sub-areas, which can cause problems if such sub-areas are submitted to the same evaluation criteria of scientific production. Furthermore, we show how such issues are mitigated by employing a straightforward equation known as the *universal fit citation* [Radicchi et al. 2008]. In Section 5, we provide conclusions and final comments, along with a list of potential themes for future work.

2. Preliminaries

2.1. Citation Distribution and Power Laws

It is a well-known fact that popularity-related contexts exhibit long-tailed distributions [Easley and Kleinberg 2010]. The analysis of scientific citations aligns with such contexts [Newman 2003, Albert and Barabási 2002]. For instance, graphs can be utilized to represent the network of scientific citations, where each paper is a node and an edge connects two papers if one cites the other. In this scenario, it can be noted that the distribution of the degrees of nodes exhibits a long tail (see Figure 1).

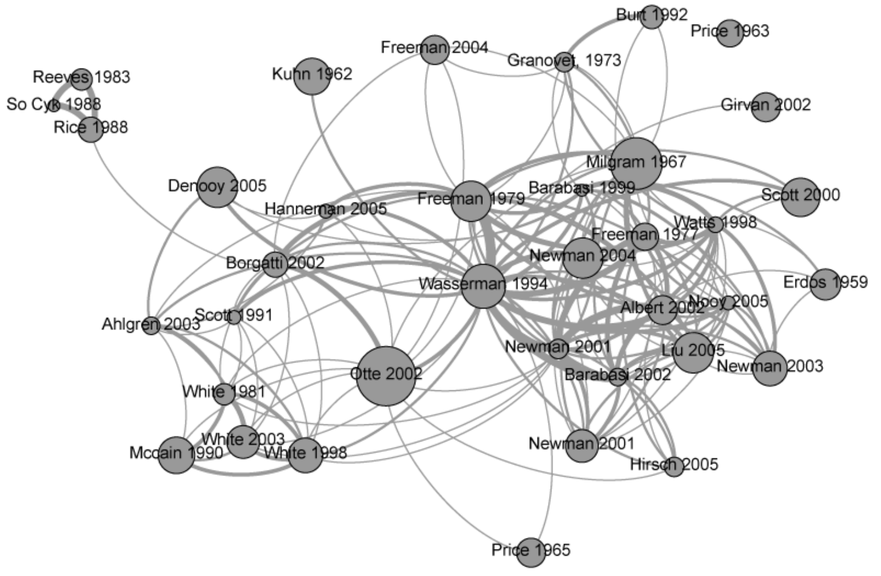


Figure 1. An example of a citation network [Ullmann 2012].

From a general standpoint, almost half of all scientific articles has not been cited even once, and less than 0.1% of all publications have reached the 1000 citation mark [Noorden et al. 2014]. Such imbalance in the number of citations is a trait of long-tailed distributions. While there are numerous distributions that fit this category, a careful analysis indicates that it follows the *power law distribution* [Noorden et al. 2014, Broido and Clauset 2019]. The formal definition is given by Definition 1.

Definition 1 (Power Law Distribution)**.** *A function that satisfies*

$$p(c) = \frac{\alpha}{c^k}$$

is said to follow a power-law distribution, for constant values $k > 0$ and $\alpha > 0$.

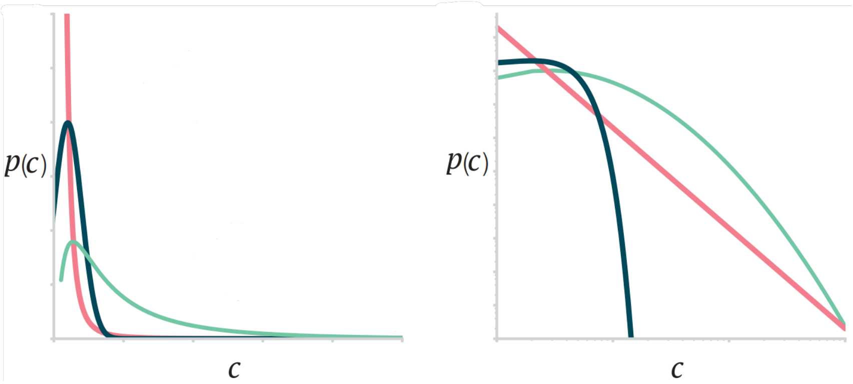


Figure 2. The figure on the left illustrates the power-law (pink), normal (blue), and log-normal (green) distributions in a linear-linear plot. The figure on the right shows these same distributions on a log-log plot [Wang and Barabási 2021].

Knowing the distribution is a key factor for taking conclusions upon these data, since the traditional metrics are not very suited for it in power law distributions [Newman 2005]. For example, one consequence is the lack of well-defined average value and variance as well. To cope with this, we base our experimental results on the entire power law curve. As the fraction of papers with *exactly* c citations may not be represented by the data, it is better to present the results with the cumulative distribution. In our context, however, it makes more sense to ask the opposite, i.e., how often the results are *above* a particular level. Formally, this is called the *complementary cumulative distribution function* or simply the *tail distribution*. In this way, we define $P(c)$ as the fraction of articles with c or more citations and use such definition to present the results.

2.2. Normalization of Citation Between Areas

Radicchi et al. [Radicchi et al. 2008] studied the citation disparity between different areas. They revealed, as one might expect, that certain areas receive more article citations than others. Thus, they proposed a universal method for rescaling the curves to achieve an unbiased indicator for citation performance across disciplines. They realized that, although the mean is not significant for taking conclusions, it captures the differences in citation count between areas. So, the proposed idea is to normalize the citation count by the average citation the area of publication had in the year of its publication. The details are presented in Definition 2. Doing this for the data previously analyzed, they found that the differences disappeared.

Definition 2 (Universal fit citation, see [Radicchi et al. 2008])**.** *Let c be the total number of citations that an article has received since its publication date and μ be the average number of citations in the article's area in the year it was published. The* universal fit citation $\tilde{c}$ *is defined as*

$$\tilde{c} = \frac{c}{\mu}.$$

When first proposed [Radicchi et al. 2008], this equation was obtained empirically. In 2021, Golosovsky [Golosovsky 2021] found an analytical derivation, making the formula epistemologically stronger as it holds in the formal and empirical paradigms. Both works refer to the log-normal distribution, which is similar to the power law distribution. There is a dispute regarding the distribution that best models complex networks, such as citation networks, due to the existence of other related distributions, like the log-normal distribution and the power law distribution with an exponential cutoff. Thus, the choice of one model or another a matter of convenience or choice [Golosovsky 2021, Mitzenmacher and Upfal 2017]. For a more in-depth exploration of this topic, we suggest referring to the works of [Clauset et al. 2009], who employs a statistical approach, and [Lima et al. 2019], who utilizes machine learning techniques.

3. Methodology

In this section, we provide an overview of the databases utilized in our study and explain how they were integrated to accomplish our goal.

3.1. Datasets

1. *CSIndexbr*

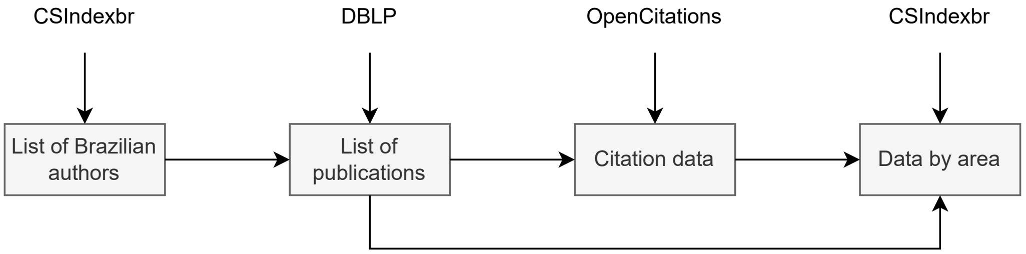
   A starting point is to obtain a list of Brazilian Computer Science authors. The CSIndexbr website is a Brazilian project with the aim of providing relevant, open and transparent information about Brazilian scientific production in computer science [Valente and Paixao 2018]. CSIndexbr maintains and makes available a list containing all Brazilian computer science researchers linked to higher education institutions. It also maintains and makes available a list relating publication outlets to areas of computer science. This data is important to obtain information about Brazilian production and relate each publication to an area.

2. *DBLP Computer Science Bibliography*

   DBLP is a German project that aims to index scientific production in computer science to support researchers through this platform, which is open and free [Bibliography 2022]. In 2019, DBLP had already indexed more than 4.4 million publications from more than 2.2 million researchers.

   Using the list of Brazilian authors obtained from CSIndexbr, DBLP provides a list of all publications by these authors. The DBLP also provides information about which vehicle the publication was made in, and, with the information from CSIndexbr, it is possible to know which sub-area of computer science that publication belongs to.

Figure 3. Data source schema.



3. *OpenCitations*

   OpenCitations is an Italian non-profit organization linked to the University of Bologna [Waltman et al. 2020]. Its objective is to provide open access to bibliographic and publication citation data (Waltman et al., 2020). As DBLP does not provide information about article citations, the list of publications obtained from it will be used to obtain the number of citations received by each publication from OpenCitations.

3.2. Working the Data

The process involves creating a database to store data obtained from Brazilian computer science researchers. Initially, a list of these researchers was obtained from CSIndexbr's GitHub, including their names, educational institutions, and researcher identifiers (PIDs) from DBLP. Subsequently, calls were made to the DBLP API using each researcher's PID to acquire publication information. However, the number of citations received remains outstanding. With this data, relationships between researchers and publications can be established in the database.

The sub-areas of computer science were also obtained from CSIndexbr. To distinguish which area a publication belongs to, information from CSIndexbr was used. The relationship of interest connects publication venues to areas (available on the project's GitHub). Each publication is linked to a venue, which, in turn, is associated with subareas. It is worth noting that a publication venue can be linked to multiple sub-areas, and consequently, so can a publication.

Finally, the number of citations received by each publication is obtained through calls to the OpenCitations API using the DOI of each publication. These steps provide all necessary data for conducting experiments. The overall schema is represented in Figure 3.

The process of gathering and analyzing data took place from November 2022 through January 2023. As a result, over 1,100 Brazilian computer science researchers were registered, and nearly 46,000 publications were found. However, only around 11,000 could be related to specific sub-areas of computer science and will be the focus of the experiment. This is due to limitations of the CSIndexbr dataset, which is unable to link these publications to its areas. Table 1 presents the sub-areas, its distribution and other data.

Table 1. Data gathered by area. Here, $P(c)$ is the tail distribution, i.e., the fraction of articles with c or more citations, #Pub. is the number of publications analyzed and #Top 100 is the number of publications that appear in the overall 100 most cited papers within our analysis.

Sub-area	$P(40)$	#Pub.	#Top 100
Security and Cryptography	0.166	90	2
Computer Vision	0.155	462	15
Data Mining and Machine Learning	0.132	393	7
Operational Research	0.130	746	12
Web and Information Retrieval	0.096	423	4
Computer Networks	0.087	1097	13
Bio-informatics	0.086	197	4
Artificial Intelligence	0.084	1358	7
Robotics	0.083	275	2
Programming Languages	0.071	251	2
Databases and Informational Systems	0.065	522	7
Computer Graphics and Multimedia	0.058	601	8
Software Engineering	0.059	1458	6
Hardware Design	0.049	264	0
Human-Computer Interaction	0.047	188	0
Distributed Systems	0.046	619	3
Computer Architecture	0.043	976	4
Algorithms and Complexity	0.028	731	3
Formal Methods and Logic	0.021	367	1
CS Education	0.020	48	0

4. Experiments and Results

This section presents the experiments and results. The experiments were launched in an Intel(R) Core(TM) i5-4210U CPU @ 1.70GHz and 8 GB RAM. The data were stored using PostgreSQL 12.13 database. The scripts were implemented in Python 3.8.10 language.

As expected, the distribution of citations in Brazilian Computer Science follows a power law. A similar trend can be seen in the top 100 most cited publications as well (Table 1), although the numbers may be too small for a solid conclusion. For example, areas such as "computer vision" appear much more frequently than "robotics", and "human-computer interaction" do not even make it to the list. Also, we observe that 18% of all publications, about 1,800 papers, have never been cited. This matches the long tail behavior of power laws, in this case, the top 20% of the most cited publications represent 60% of all citations for those works (Figure 4).

4.1. Disparity Between Areas of Brazilian Computer Science

As already seen in large areas of science [Radicchi et al. 2008], we also observe disparity in citations between sub-areas in Brazilian computer science. Figure 5, Figure 6, Figure 7 and Table 2 presents these results.

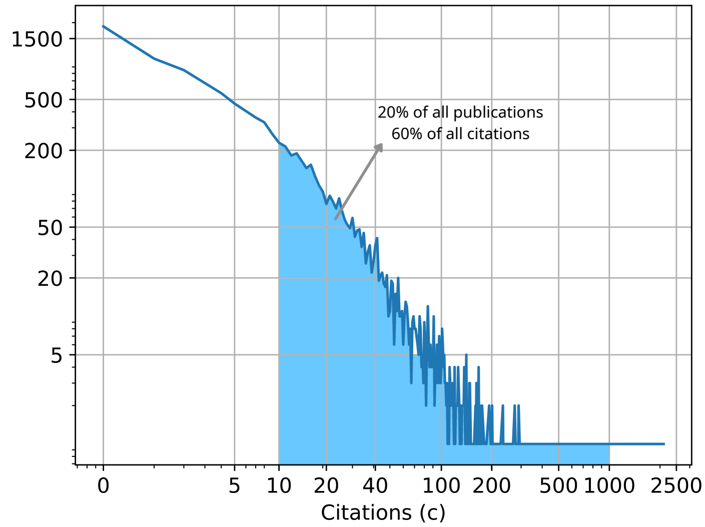
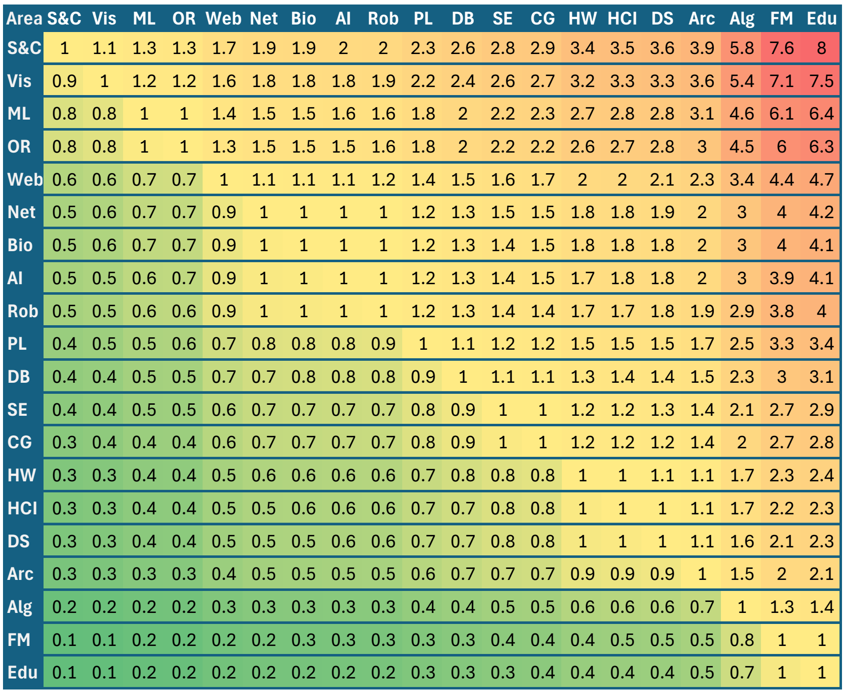


Figure 4. The distribution $P(c)$, i.e., the fraction of articles with c or more citations. The hatched region represents 20% of all publications, which has 60% of the overall citations.

Figure 5 shows a heat map, which reflects the disparity between the sub-fields of computer science. In the figure, we set $c = 20$, that is, considering only the publications that received 20 or more citations. The value contained in position (i, j) is the fraction $P_i(20)/P_j(20)$. For example, in the row "Vis" (Computer Vision) and in the column "Alg" (Algorithms and Complexity), the value is 5.4. This means that there are 5.4 times more articles with at least 20 citations in Computer Vision than in Algorithms and Complexity.



Area	S&C	Vis	ML	OR	Web	Net	Bio	AI	Rob	PL	DB	SE	CG	HW	HCI	DS	Arc	Alg	FM	Edu
S&C	1	1.1	1.3	1.3	1.7	1.9	1.9	2	2	2.3	2.6	2.8	2.9	3.4	3.5	3.6	3.9	5.8	7.6	8
Vis	0.9	1	1.2	1.2	1.6	1.8	1.8	1.8	1.9	2.2	2.4	2.6	2.7	3.2	3.3	3.3	3.6	5.4	7.1	7.5
ML	0.8	0.8	1	1	1.4	1.5	1.5	1.6	1.6	1.8	2	2.2	2.3	2.7	2.8	2.8	3.1	4.6	6.1	6.4
OR	0.8	0.8	1	1	1.3	1.5	1.5	1.5	1.6	1.8	2	2.2	2.2	2.6	2.7	2.8	3	4.5	6	6.3
Web	0.6	0.6	0.7	0.7	1	1.1	1.1	1.1	1.2	1.4	1.5	1.6	1.7	2	2	2.1	2.3	3.4	4.4	4.7
Net	0.5	0.6	0.7	0.7	0.9	1	1	1	1	1.2	1.3	1.5	1.5	1.8	1.8	1.9	2	3	4	4.2
Bio	0.5	0.6	0.7	0.7	0.9	1	1	1	1	1.2	1.3	1.4	1.5	1.8	1.8	1.8	2	3	4	4.1
AI	0.5	0.5	0.6	0.7	0.9	1	1	1	1	1.2	1.3	1.4	1.5	1.7	1.8	1.8	2	3	3.9	4.1
Rob	0.5	0.5	0.6	0.6	0.9	1	1	1	1	1.2	1.3	1.4	1.4	1.7	1.7	1.8	1.9	2.9	3.8	4
PL	0.4	0.5	0.5	0.6	0.7	0.8	0.8	0.8	0.9	1	1.1	1.2	1.2	1.5	1.5	1.5	1.7	2.5	3.3	3.4
DB	0.4	0.4	0.5	0.5	0.7	0.7	0.8	0.8	0.8	0.9	1	1.1	1.1	1.3	1.4	1.4	1.5	2.3	3	3.1
SE	0.4	0.4	0.5	0.5	0.6	0.7	0.7	0.7	0.7	0.8	0.9	1	1	1.2	1.2	1.3	1.4	2.1	2.7	2.9
CG	0.3	0.4	0.4	0.4	0.6	0.7	0.7	0.7	0.7	0.8	0.9	1	1	1.2	1.2	1.2	1.4	2	2.7	2.8
HW	0.3	0.3	0.4	0.4	0.5	0.6	0.6	0.6	0.6	0.7	0.8	0.8	0.8	1	1	1.1	1.1	1.7	2.3	2.4
HCI	0.3	0.3	0.4	0.4	0.5	0.5	0.6	0.6	0.6	0.7	0.7	0.8	0.8	1	1	1	1.1	1.7	2.2	2.3
DS	0.3	0.3	0.4	0.4	0.5	0.5	0.5	0.6	0.6	0.7	0.7	0.8	0.8	1	1	1	1.1	1.6	2.1	2.3
Arc	0.3	0.3	0.3	0.3	0.4	0.5	0.5	0.5	0.5	0.6	0.7	0.7	0.7	0.9	0.9	0.9	1	1.5	2	2.1
Alg	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.5	0.5	0.6	0.6	0.6	0.7	1	1.3	1.4
FM	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.3	0.3	0.4	0.4	0.4	0.5	0.5	0.5	0.8	1	1
Edu	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.3	0.3	0.4	0.4	0.4	0.4	0.5	0.7	1	1

S&C - Security & Cryptography
Vis - Computer Vision
ML - Data Mining & Mach. Learning
OR - Operational Research
Web - Web & Information Retrieval
Net - Computer Networks
Bio - Bioinformatics & Comp. Biology
AI - Artificial Intelligence
Rob - Robotics
PL - Programming Languages
DB - Data Bases & Inf. Systems
SE - Software Engineering
CG - Comp. Graphics & Multimedia
HW - Hardware Design
HCI - Human-Computer Interaction
DS - Distributed Systems
Arc - Comp. Architecture & HPC
Alg - Algorithms & Complexity
FM - Formal Methods & Logic
Edu - Computer Science Education

Figure 5. A heat map showing the disparity between the sub-fields.

The data is partitioned into 20 distinct sub-areas, but for a more organized and concise representation, only six of these sub-areas are showcased in Figure 6 and Table 2. We see that the differences between the sub-areas are quite significant, e.g., publications in “computer vision” with at least 100 citations are 11 times more common than in “computer architecture”. Also, the sub-area of “formal methods and logic” has about 10% of its publications with 20 or more citations. On the other hand, the sub-area of “security and cryptography” has about 31% of its publications with 20 or more citations. That is, finding a paper with 20 or more citations is three times more likely in the latter than in the former. It is worth noting that the data presented in Figure 6 and Table 2 are the extreme cases. The other 14 sub-areas fall somewhere in between, with some closer to the lower extreme and others closer to the upper extreme.

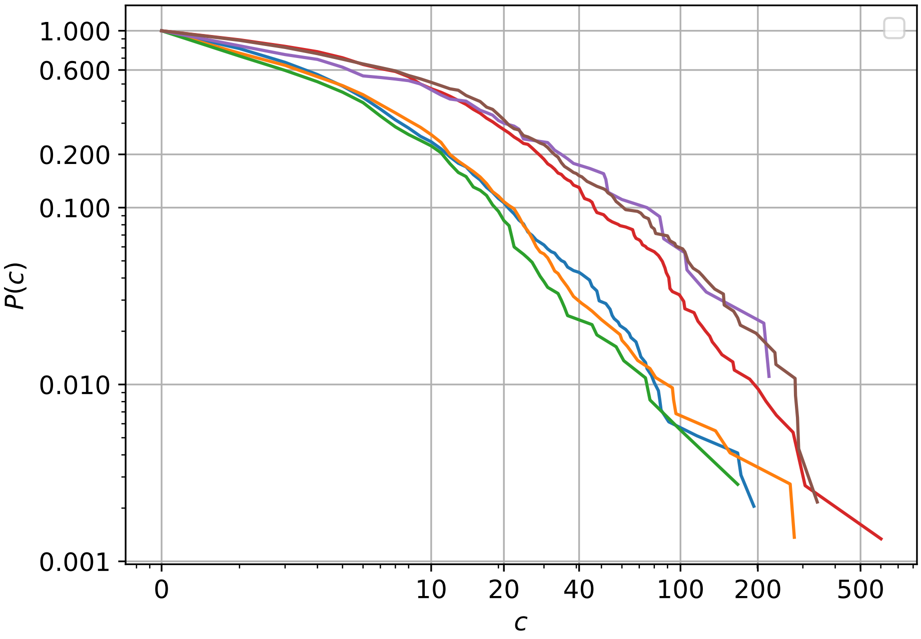


Figure 6. Citation distribution by area, where c is the number of citations, and $P(c)$ is the tail distribution, i.e., the fraction of articles with c or more citations.

Table 2. Tail distribution of the citation count c and the universal fit citation $\tilde{c}$.

	Standard			Universal Fit		
Sub-Area	c=20	c=40	c=100	$\tilde{c}$=2	$\tilde{c}$=4	$\tilde{c}$=8
Computer Vision	0.335	0.155	0.058	0.251	0.086	0.019
Security and Cryptography	0.311	0.166	0.055	0.277	0.100	0.022
Operational Research	0.289	0.130	0.032	0.264	0.077	0.014
Formal Methods and Logic	0.095	0.021	0.005	0.286	0.079	0.016
Algorithms and Complexity	0.116	0.028	0.005	0.284	0.067	0.016
Computer Architecture	0.112	0.043	0.005	0.253	0.069	0.020

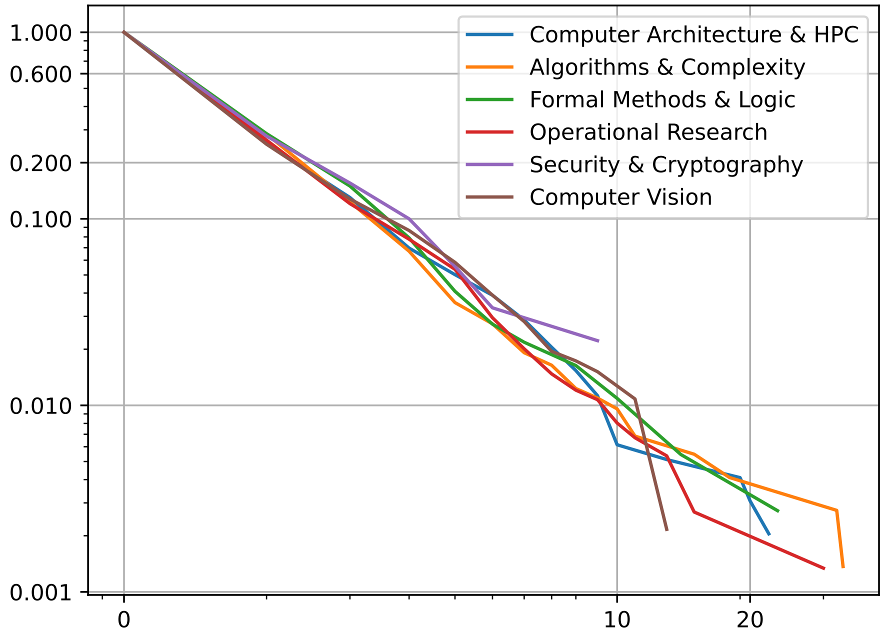


Figure 7. Universal Fit Citation distribution by area, where $\tilde{c}$ is the universal fit citation (see Definition 2) and $P(c)$ is the tail distribution, i.e., the fraction of articles with c or more citations.

4.2. Normalization of Citation Between Brazilian Sub-Areas

Given the disparity in citation distribution in Brazilian sub-areas, we tested the approach by Radicchi et al. [Radicchi et al. 2008], as described in Definition 2, to mitigate the citation differences between areas. Once more, we present the results for the six sub-areas that were previously chosen. Figure 7 and Table 2 show how normalization makes the distributions closer, suggesting that comparisons between sub-areas are now fairer. For example, the mentioned difference between "formal methods and logic" and "security and cryptography" is greatly mitigated. And this applies to all sub-areas analyzed. It is worth noting that this adjusted distribution remains a power law distribution.

5. Conclusion

The evaluation of scientific production is a debated topic. Although objective data such as the number of citations is available, it cannot be directly compared [Radicchi et al. 2008]. Alternative evaluation methods have been proposed, but they face subjectivity issues [Wang and Barabási 2021]. Therefore, the number of citations remains central to evaluation as it represents the community's opinion of the work.

As also observed in other works [Radicchi et al. 2008], our analysis showed that the distribution of citations is uneven for different areas of Brazilian Computer Science, preventing a fair comparison. However, the proposed normalization factor effectively minimizes the difference between citation curves, suggesting that comparison between areas and sub-areas occurs more fairly, particularly if submitted to evaluation by the same area committee.

Given the size of the Brazilian computer science community, this suggests that the sample taken is representative, and thus the results presented in this work can be extrapolated to other countries or regions. However, this may not be true, as it depends on other factors. In any case, the methodology we used can be replicated in other contexts, as long as information similar to that found in CSIndexbr is available.

A future work involves developing a web page to facilitate access to this information for the entire community and potentially support area committee decisions regarding sub-areas. From a more analytical perspective, network metric studies could be conducted instead of focusing solely on the power-law distribution characteristic of complex networks. Another line of investigation is the viral propagation process leading to citations, which varies across different sub-areas [Golosovsky 2021]. Lastly, our results highlight the existence of 18% of articles that are not cited, but this number can be as high as 50% [Noorden et al. 2014]. Previous works by [Golosovsky and Larivière 2021] and [Katchanov et al. 2023] have explored the impact of uncited articles. Applying such studies to each sub-area could potentially recalibrate weights and influence, especially when using universal fit citation method.

References

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97.

Bibliography, D. C. S. (2022). Monthly snapshot. `https://dblp.org`. Accessed 30/11/2022.

Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nature Communications*, 10(1017).

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.

Dong, Y., Ma, H., Shen, Z., and Wang, K. (2017). A century of science: Globalization of scientific collaborations, citations, and innovations. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1437–1446. Association for Computing Machinery.

Easley, D. and Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press.

Golosovsky, M. (2021). Universality of citation distributions: A new understanding. *Quantitative Science Studies*, 2(2):527–543.

Golosovsky, M. and Larivière, V. (2021). Uncited papers are not useless. *Quantitative Science Studies*, 2(3):899–911.

Katchanov, Y. L., Markova, Y. V., and Shmatko, N. A. (2023). Uncited papers in the structure of scientific communication. *Journal of Informetrics*, 17(2):101391.

Lima, A., Vignatti, A., and Silva, M. (2019). Recognizing power-law graphs by machine learning algorithms using a reduced set of structural features. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 611–621. SBC.

Mitzenmacher, M. and Upfal, E. (2017). *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review*, 45(2):167–256.

Newman, M. E. J. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics*, 46(5):323–351.

Noorden, R. V., Maher, B., and Nuzzo, R. (2014). The top 100 papers. *Nature*, 514(7524):550–553.

Radicchi, F., Fortunato, S., and Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. In *National Academy of Sci.*

Ullmann, T. (2012). Co-citation analysis of the topic social network analysis. `https://eduinf.eu/2012/03/15/co-citation-analysis-of-the-topic-social-network-analysis/`. Accessed 2023/10/15.

Valente, M. T. and Paixao, K. (2018). CSIndexbr: Exploring the Brazilian scientific production in Computer Science. *arXiv*, abs/1807.09266.

Waltman, L., Larivière, V., Milojević, S., and Sugimoto, C. R. (2020). Opening science: The rebirth of a scholarly journal. *Quantitative Science Studies*, 1(1):1–3.

Wang, D. and Barabási, A. (2021). *The Science of Science*. Cambridge University Press.