D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research

Jan Philip Wahle[†], Terry Ruas[†], Saif M. Mohammad[††], Bela Gipp[†]
[†]University of Wuppertal Germany, [††]National Research Council Canada
[†]{wahle, ruas, gipp}@uni-wuppertal.de
[††]saif.mohammad@nrc-cnrc.gc.ca

Abstract

DBLP is the largest open-access repository of scientific articles on computer science and provides metadata associated with publications, authors, and venues. We retrieved more than 6 million publications from DBLP and extracted pertinent metadata (e.g., abstracts, author affiliations, citations) from the publication texts to create the DBLP Discovery Dataset (D3). D3 can be used to identify trends in research activity, productivity, focus, bias, accessibility, and impact of computer science research. We present an initial analysis focused on the volume of computer science research (e.g., number of papers, authors, research activity), trends in topics of interest, and citation patterns. Our findings show that computer science is a growing research field (≈15% annually), with an active and collaborative researcher community. While papers in recent years present more bibliographical entries in comparison to previous decades, the average number of citations has been declining. Investigating papers' abstracts reveals that recent topic trends are clearly reflected in D3. Finally, we list further applications of D3 and pose supplemental research questions. The D3 dataset, our findings, and source code are publicly available for research purposes.

Keywords: Computer Science, Scientometrics, Research Trends, NLP, DBLP, AI

1. Introduction

In the last few decades, computer science (CS) has transformed many scientific fields. Faster systems, more accurate results, and efficient tools are just some of the benefits provided by computer science advancements. Arguably, today there is hardly any area not affected by its vast possibilities. Consider how difficult it would be to test, develop, and research new vaccines without access to tools of informatics (e.g., public repositories).

The techniques behind these contributions are often made available through scientific publications which can be used to investigate trends and patterns within computer science itself. However, computer science is a large field with many sub-areas (e.g., natural language processing (NLP), computer vision) and repositories (e.g., arXiv); thus, a complete analysis of its publications is a challenging task. How fast is computer science research growing? How many authors are actively publishing in their field? What topics are prevalent in specific venues? A large and carefully curated dataset of CS-publications metadata is crucial for the quantitative exploration of these questions.

DBLP is one of the largest open computer science repositories with records from major journals and proceedings starting from 1936.[1] The repository provides access to several pieces of metadata associated with each of its papers, including author names, title, year, and venue. The papers stored in DBLP come from paid publishers (e.g., IEEE, ACM) and open repositories (e.g., ACL, arXiv). NLP Scholar (Mohammad, 2020c) and arXiv also offer an extensive collection of papers in computer science, but both are included in DBLP. Therefore DBLP offers a more complete corpus to understand patterns in computer science. However, as DBLP mainly indexes metadata of papers, it lacks some important information that is embedded in their full texts (e.g., affiliations, citations).

In this paper, we introduce the DBLP Discovery Dataset (D3), which not only includes key information about papers from DBLP in an easily accessible form, but it also enriches it with crucial metadata such as abstracts, author affiliations, and citations, that we extracted from the full texts. Thus, D3 can be used to explore and understand broad trends in computer science research.

Our dataset is proposed as a lightweight resource to explore the patterns in computer science publications and the relation between the elements describing them, e.g., what are the most popular topics of conferences in 2021? How active have authors been over the years, and how long, on average, are authors active? D3 can also be used as a training corpus in language modeling, classification of papers by topic and venues, statistical analysis of publishers, and several other scenarios. Even though Google Scholar and Semantic Scholar are similar to D3, they only provide access to their metadata. For example, while one can access the number of citations and information on which paper cites which paper individually, one cannot access the data in bulk for large-scale quantitative analysis. In addition, the access to the actual dataset in Google Scholar and Semantic Scholar is not straightforward. For example, Google Scholar has no standard API and limits its web page access to crawl. Although the Semantic Scholar Open Research Corpus (S2ORC) offers a dump from

[1] https://dblp.org/

2020 (Lo et al., 2020), its 81 million papers require more than 800GB of storage, which can be restrictive to many researchers trying to process and analyze the data. In comparison, the uncompressed size of D3 is ≈ 18GB.

In summary, our contributions are two-fold. We (1) publish a new open dataset[2] of $\approx$ 6 million English research papers and the source code to retrieve them[3], and (2) provide an initial investigation of computer science publications. D3 augments DBLP with metadata extracted from full-text to provide additional features over existing datasets such as paper abstracts and author affiliations (see Table 2 for more details). We provide an exploratory analysis of D3 through eight research questions to illustrate some of our dataset's main capabilities. Our questions investigate the volume, content, and citations of papers in D3.

2. Existing Resources

The NLP Scholar dataset (Mohammad, 2020c) combines primary information from $\approx 45,000$ NLP publications in the ACL Anthology (e.g., authors, venues) with citation information from Google Scholar. Mohammad (2020c; 2020b) used the NLP Scholar dataset to examine citation patterns, the gender gap between female and male first-time authors, and $n-$gram distributions through interactive visualization. DRIFT (Sharma et al., 2021) tracks the changes in *cs.CL*, the computer science computational linguistics tag from arXiv, focusing on single-word terms and their word embeddings over time. In NLPExplorer, (Parmar et al., 2020) provide an exploratory tool for NLP publications based on ACL Anthology, including information on most-cited authors, areas, and venues; similar to NLP Scholar. As of June 2021, NLPExplorer also explores Tweets related to publications and conferences. The NLP4NLP Corpus (Mariani et al., 2019) contains $\approx 65,000$ articles from 34 conferences and journals in NLP such as the ACL Anthology and the International Speech Communication Association. They provide extensive analysis on references and volume, citations, and authorship. S2ORC (Lo et al., 2020) is a repository of 81.1 million English academic papers from 20 research fields like medicine, biology, or physics. Apart from citations, semantic scholar also offers information about venues and authors.

We extend current datasets, i.e., NLP Scholar, DRIFT, in two key directions. We (1) include computer science venues outside of the ACL Anthology and arXiv, and (2) add informative features derived from full-texts (e.g., citations, paper abstracts, or author affiliations). As NLP research is not restricted to a single repository, D3 provides a more comprehensive view on the entirety of NLP research and many other subfields in computer science. Our dataset contains records from many publishers (e.g., Springer, IEEE, ACM), including the entire ACL Anthology and computer science publications on arXiv, thus allowing D3 to answer all questions from previous datasets more accurately.

Attribute	Example
publication	
id	conf/acl/Mohammad20b
modified date	2021-09-12
title	NLP Scholar - An Interactive ...
pages	232-255
year	2020
type	Conference and Workshop Papers
access	open
links	[https://doi.org/...]
doi	10.18653/v1/2020.acl-demos.27
publisher	ACL
author	
id	58/380
fullname	Saif M. Mohammad
webpage	http://saifmohammad.com/
venue	
names	[International Conference on Lang...]
acronyms	[LREC]
type	Conference or Workshop
id	conf/lrec

Table 1: Primary attributes of D3.

3. Dataset Collection

We extracted all records from DBLP and crawled their associated full-text PDFs to extract metadata (e.g., bibliographies) from January 1st, 1936 until December 2nd, 2021.[4] The subsections below describe how we extracted and aligned DBLP and full-text information.

3.1. Primary Information from DBLP

DBLP provides open access to its data in two ways, a public search API[5], and a XML dump[6]. We are interested in understanding the state of computer science research at a large scale over time, so we retrieve their full XML release. To keep D3 up-to-date, we download the latest DBLP release monthly and compute the changes to our last version of D3. We provide an overview of the attributes with examples that we retrieve from DBLP in Table 1.

Publications. The majority of entries in DBLP are indexed publications with metadata. Examples for other entries are web pages and author information. The dataset classifies papers according to the BibTeX entry types[7] (e.g., article, in proceedings). We transform publications by paper type into a standard JSON format and map authors and venues to uniquely identified entities.

[2] `https://zenodo.org/record/6477785`

[3] `https://bit.ly/3ETGazz`

[4] DBLP is monthly updated with new records.

[5] `https://dblp.org/search/publ/api`

[6] `https://dblp.org/xml/release`

[7] `https://bit.ly/33vqaVG`

Attribute	Example
affiliations	
id	4eb3...f094
name	National Research Council Canada
country	Canada
city	Ottawa
postcode	K1A 0R6
addressline	1200 Montreal Road, Bldg. M-58
outgoing citations	
ids	[7615..., 76af...]
count	2
incoming citations	
ids	[7ca5..., 7d0e...]
count	11
keywords	[Scientometrics, Citations, ...]
ocr title	NLP Scholar: An Interactive ...
ocr abstract	As part of the NLP Scholar ...

Table 2: Secondary attributes of D3.

Authors. DBLP processes multiple authors with the same name using an iterative counter and the same authors with various variants of their names with unified entities. We create a unique id for each author to map them to their publications. Author entities in DBLP are sparse and typically only provide a personal webpage URL but rarely other informative features such as a current affiliation. To enhance authors' entities with beneficial features, we extract information from full-texts (e.g., affiliations) in Section 3.2.

Venues. For most publications, DBLP provides a venue code (i.e., the abbreviation of the venue). We create venue entities for each and map them to papers by generating a unique id. DBLP also contains information about major publishers such as Springer, IEEE, and ACM. As the data on publishers is scarce (≈10% or publications have publishers annotated), we store publishers directly in the publication entries.

3.2. Secondary Information from Full-Texts

Publications' full-texts contain valuable information about author affiliations, content, and references not present in DBLP, other datasets (e.g., NLP Scholar), or online services (e.g., Google Scholar). We provide an overview of the attributes with example values that we extract from full-texts in Table 2.

Abstracts. We retrieve the abstracts and index keywords of publications using GROBID (GRO, 2008 2021). For abstract extraction, the model uses CRF Wapiti (Lavergne et al., 2010) features and achieves an F1-score (using Levenshtein Matching with minimum distance of 0.8) of 92.85% when drawing 1943 PubMed[8] papers[9]. Using this model, we retrieved 4,219,855 abstracts from papers which are 78.17% of the dataset. The remaining papers were either disregarded by GROBID because of poor quality or did not provide an accessible document that could have been parsed. We use the publication's unique id to align the extracted information with DBLP.

Affiliations. We extract the author names and affiliations with the same model and sequence features as used for extracting the abstracts. To create author–affiliation pairs, we match author names from extracted affiliations to author names in DBLP using Levenshtein distance. In practice, creating author–affiliation pairs through name matching is robust (we found less than 5% cases where it fails). We demonstrate this by performing two small bootstrap and permutation tests (Dror et al., 2018). In the first test, we draw 20 samples of $n = 100$ publications uniformly at random and evaluate how often author names extracted from the PDFs do not match those in DBLP. To draw more challenging samples in the second test, we took the first $n = 100$ publications from a ranked list in which the average Levenshtein distance between authors' names was increasing. Both tests show less than 5% of names are mismatched ($p < 0.001$).

Citations. To collect the citations within DBLP, we build a citation graph from the bibliographies in full-texts similar to the Reference Corpus of the ACL Anthology (Radev et al., 2009). To parse the publications' bibliographies, we use GROBID's BidLSTM-CRF features, which obtain an F1-score of 87.73% for the PubMed samples (using Levenshtein Matching with minimun distance of 0.8)[10]. We build the citation links by adding two fields to each publication object, the incoming citations (i.e., how often a paper was cited) and outgoing citations (i.e., the number of bibliography entries in a paper). The resulting citation graph allows us to investigate the role of authors, venues, publishers, and institutions in research. Additionally, their interaction will also help us identify implicit trends, common topics, and influence between the participants of this graph. Even though Google Scholar offers an open access service, its data is restricted, preventing researchers from obtaining large-scale access to their citation information. Google Scholar also does not have a standardized API and limits its access for crawling their webpage. Linking citation within DBLP provides us with a focused view on the influence of sub-fields within computer science. In Section 4.3, we measure citations coming from fields other than computer science using Semantic Scholar with the result that 21.15% of citations are cited from papers outside of our repository (i.e., from other fields than computer science).

3.3. Implementation Details

Parsing releases of ≈ 6 million publications and extracting their metadata from full-texts is a resource-intensive process. Therefore, we implemented a parallel and asynchronous routine to parse releases, retrieve full-texts, extract their metadata, and align the infor-

[8] https://pubmed.ncbi.nlm.nih.gov/

[9] https://bit.ly/3K9Vf2g

[10] https://bit.ly/3K9Vf2g

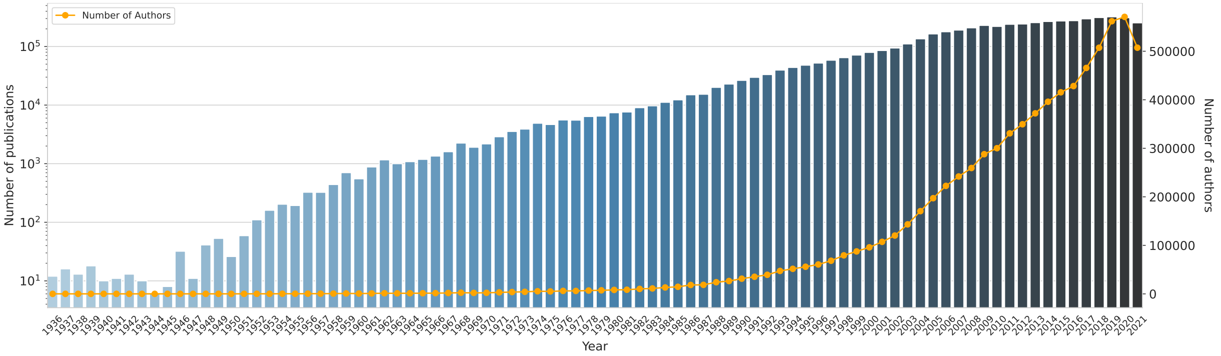


Figure 1: The number of annual publications and authors in logarithmic scale between 1936 and 2021.

mation to DBLP with a low disk, memory, and computational requirements. First, we split the dataset into equally sized chunks as it allows us to work on mutually exclusive parts of the dataset with multiple processes at different times without processing the whole repository. We launch n processes to retrieve publications for n chunks. Each process asynchronously requests the PDF link of a paper or, if not present, parses the HTML page of the paper to identify the PDF link. Next, we download the PDF to a folder with its unique key. As all requests are sent asynchronously, we reduce their idle times in between. To restrict requests to the same domain and respect server limits, we use semaphores and wait to respect the retry-after header whenever we receive an HTTP 429 response. In parallel to the n retrieval processes, we launch another n processes to work on full-texts from the previously downloaded chunk and extract their metadata. To reduce disk requirements, we delete the full-texts after extraction and only keep their metadata. The uncompressed size of D3, in JSON format, is $\approx$ 18GB.

4. Dataset Analysis

4.1. Volume & Research Activity

Q1. How large is DBLP? How does the number of publications change over time?

A. As of December 2021, DBLP contains a total of 6,392,734 entries. Table 3 shows the number of publications by type until December 2021. Most DBLP publications are either conference/workshop papers (47.12%) or journal articles (43.42%). The third-largest category (6.05%) is what DBLP refers to as ”informal publications”. These are papers published in online repositories (such as arXiv) without a systematic peer review, as well as contributions to informal workshops. A majority of these are arXiv pre-prints from the computing research repository (CoRR). When informal publications are published in a peer-reviewed venue, DBLP updates its type accordingly.

DBLP contains 99.3% of papers from the ACL Anthology. The papers in ACL Anthology are concentrated between conferences and workshops (>90%) (Mohammad, 2020c), while DBLP provides a more balanced distribution between journal articles, conference, and workshop papers. As DBLP contains the ACL Anthology and other non-ACL venues (e.g., IEEE, ACM), D3 is a robust resource that enables the correlation between publications, venues, and authors in NLP and other areas in computer science.

Paper Type	Count	Proportion
in proceedings	3,012,358	47.12%
article	2,776,011	43.42%
informal	386,574	6.05%
phd thesis	81,954	1.28%
in collection	67,502	1.05%
proceedings†	49,265	0.77%
book	19,070	0.30%
total	**6,392,734**	**100%**

Table 3: The number of publications in DBLP by type until December 2021. †The entire collection published.

The number of publications in computer science is growing on average 15.12% yearly. Between 1936 and 1952, the annual number of publications never exceeded one hundred papers as Figure 1 shows. However, after 1952 the number of publications start growing exponentially[11].

Q2. How many authors publish in D3 over time? How has the average number of authors per paper changed over time?

A. There are ≈2,9 million authors for the ≈6 million papers in D3. The line in Figure 1 shows the number of authors yearly between 1936 and 2021. The number of authors in research papers increased significantly in 1990, showing that computer science became a popular research field. Considering the advances in software and hardware in the last decades with the increasing dependency between computer science and other research areas, we expect the number of papers and authors to continue to grow in the following years.

[11] The seeming drop in publications and authors in 2021 results from the dataset being crawled on December 2nd and therefore does not include the full month of December.

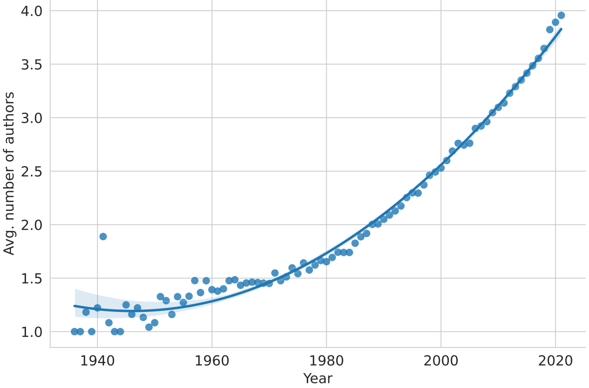


Figure 2: A cubic approximation on the average number of authors per paper between 1936 and 2021.

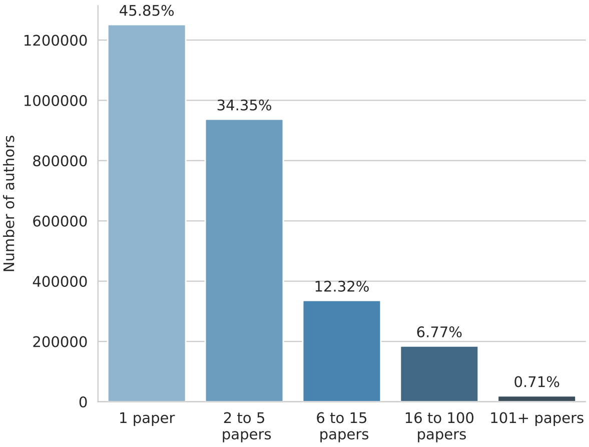


Figure 3: The number of authors and their published papers until December 2021.

Looking at the average number of authors on a paper over time (Figure 2), we observe a steep increase from 1970 onward. The increase in authors on the same paper indicates growing collaboration in computer science; a positive sign for a healthy and growing research area.

Q3. How many authors published one or more papers?

A. More than 1.4 million authors have published precisely one paper in DBLP. Figure 3 shows the number of D3 authors corresponding to different number-of-papers bins. Similar bins were used in the analysis of NLP papers in (Mohammad, 2020c).

Analysis of both D3 and ACL Anthology (Mohammad, 2020c) shows that most authors publish exactly one paper. Though, the skew towards single-publication authors is more stark in the ACL Anthology (57.9%) compared to D3 (45.85%). The NLP field has seen a substantial surge in new authors recently, and we hypothesize their gain is higher than the computer science average, explaining the greater skew in ACL Anthology. If we consider the sum of authors who published two or more papers, D3 shows 54.14%, while ACL Anthology only shows 42.10%, corroborating our assumption. The same behavior in the number of authors is also observed when using the same bin split as in (Mohammad, 2020c).

Q4. Who is actively publishing in DBLP?

A. To answer this question, we identify authors that published at least a certain amount of papers over the last years starting from 2021. In our initial investigation, we measure the number of authors that published at least x={2, 3, 5, 8, 13} papers in the last y={2, 3, 5, 8, 13} consecutive years (before 2021). Intuitively, the more papers are published by an individual author in a specific time range, the more active this researcher is. Figure 4 shows the results visualized in a heatmap. We find the highest proportion of active researchers with 13 consecutive years and 2 or more papers published (45.95% of authors). The results seem more sensitive to changes in the number of published papers (x-axis) than the time range (y-axis). When increasing the number of papers to 3 or more, the number of active researchers drops to 28.97%, while when decreasing the time to 8 years, it remains at 44.59%. We assume the typical time in which researchers publish actively is relatively short as a significant proportion of research is performed with a limited time horizon (e.g., Ph.D. students). The increase in the number of active researchers gets smaller as we increase the number of consecutive years considered. If we assume a doctorate degree takes, on average, between 5 and 8 years[12] the decline in the difference between years (especially from 8 to 13) for active researchers can be justified. In the last 13 years, less than 2 out of 100 researchers (1.36%) published 13 or more papers, showing that few authors remain being active in academia for more than a decade. This experiment shows many other possibilities of D3, such as investigating how many of the researchers move from the first author to other positions, indicating their role in the research has changed.

4.2. Trends in Topics

Q5. Which are the most common terms in titles and abstracts, and how do they differ?

A. We visualize the 30 most frequent unigram lemmas (without stopwords) of titles (left) and abstracts (right) in Figure 5. Our first observation is that the frequent words in titles convey key research topics and are not filler words or discourse connectives. Abstracts contain filler words such as “also”, “present”, “new”. The 5 most frequent words in both abstracts and titles contain “model” and “data”, indicating an increase in importance for data to obtain effective models in computer science applications. The term “network” is represented frequently in abstracts and titles, which might be related to both network analysis and neural networks. Our findings suggest the rising of machine learning if we consider terms such as “learning”, “optimization”, and “neural”. A key difference be-

[12] This time can vary depending on the program and area.

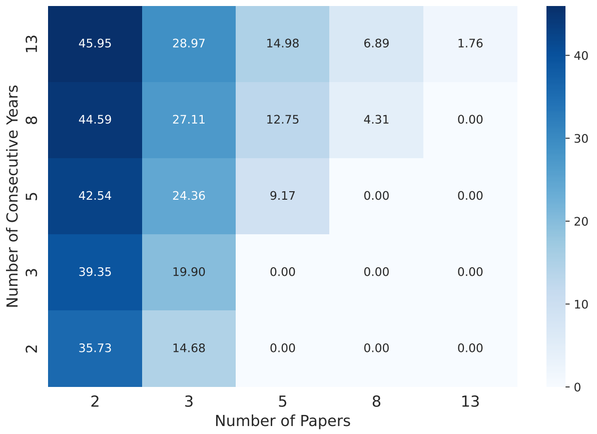


Figure 4: The relative amount of active researchers (colored and in %). An active researcher is defined by the minimum number of papers published (x-axis) in a number consecutive years (y-axis). For example, in the last 13 years, out of all researchers who published in that time, 45.95% published 2 or more papers.

tween abstracts and titles is that the former contains additional information about the paper content, reflected in lemmas such as "performance", "time", and "results". In the future, we plan to investigate the semantic representation of terms in titles and abstracts to understand and compare their content.

Q6. Which topic trends are described in abstracts?

A. In Figure 6, we show the occurrence of unigram term frequencies of abstracts for ACL from 2017[13] (y-axis) and 2021 (x-axis). We chose ACL as it is one of the largest and most influential conferences in NLP, a major sub-field of computer science. Terms close to the diagonal represent similar frequencies among both years.

We find "summarization", "dialogue", and "topic" close to the diagonal as their studies are prevalent in 2017 and 2021. Data points located on the most right of the x-axis and at the same time on the lower part of the y-axis indicate an emerging topic, whereas the converse suggests that the interest in a topic is declining. For example, in the lower right, terms such as "bert", "transformer", and other related prefixes (e.g., "pre" for "pre-training", "fine" for "fine-tuning", or "masked" for "masked language modeling") frequently appear in 2021 but rarely appear in 2017. While these terms were not popular in 2017, four years later, 90 papers mentioned "bert" and 64 papers mentioned "transformers" at least once in their abstracts. This finding is in line with a recent trend of the Transformer model published in 2017 (Vaswani et al., 2017), and particularly the popularity of BERT which was published in 2018 (arXiv) / 2019 (NAACL) (Devlin et al., 2019). Assuming Transformer-based models will continue to be applied in 2022, a new version of Figure 6 would show "transformer" even further on the x-axis if compared to 2017, or closer to the diagonal if 2021 was considered. Another group of terms is related to "covid-19", the viral disease of SARS-CoV-2 first measured in late 2019 and affecting countries worldwide since early 2020[14].

Terms that reduce in frequency compared to the previous four years are for example "sequence", "lstm", "recurrent", "vector", and "embedding" which can be related to traditional recurrent sequence models and static word embeddings used before dynamic models (e.g., Transformer). Also, terms such as "parsing", "dependency", or "convolutional" reflect a decrease in interest in of dependency parsing and convolutional neural networks in NLP applications.

Analyzing term frequencies is the first step to understanding topics and trends in D3. In the future, we will combine term frequencies with topic models to better understand computer science papers' contents and the relation between their topics.

4.3. Citation Patterns

Q7. How many sources do papers use, and how many citations do papers acquire?

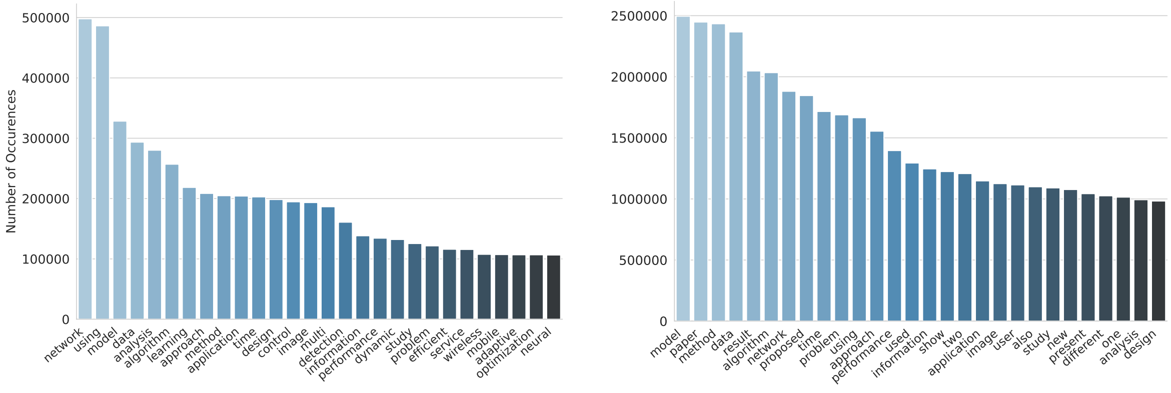
A. Figure 8 shows the distribution of incoming citations (i.e., how often a single paper was cited) and outgoing citations (i.e., the number of bibliography entries in a single paper) for all papers in D3. The average number of incoming and outgoing citations are 28.16 and 22.95, respectively. Even though the average citation count for both incoming and outgoing are similar, their distribution is quite different. While outgoing citations are seemingly normal distributed, incoming citations are skewed to the left. When considering the median of the incoming citations (8), we observe that most papers achieve less than a hundred citations. Only a few papers reach the thousand mark. A fair proportion of papers receive no citations (1,921,844, or 30.06%) or only 1 citation (431,227, or 6.75%). For the outgoing citations, the median (18) is closer to the average and few papers' bibliographies contain more than a hundred entries. These results are valid for papers citing others within D3. To measure how many incoming citations are from papers of other fields outside of D3, we use the Semantic Scholar API with the result that 21.15% of citations come from papers outside of D3.

Q8. Are we citing more papers than in earlier decades? Do recent papers receive more citations than papers published earlier in the past?

A. Tracking citations over time reveals two patterns. Figure 7 shows the trend of average incoming (left) and outgoing (right) citations over the time between 1936 and 2021. We observe that the period before mid 1960s has much more variability in terms of incoming

[13] 2017 was chosen as a starting point because of the publication of the highly influential Transformer model in (Vaswani et al., 2017).

[14] `https://bit.ly/3GrWojh`



(a) Frequencies of terms in titles.

(b) Frequencies of terms in abstracts.

Figure 5: The most common terms in titles and abstracts between 1936 and 2021.

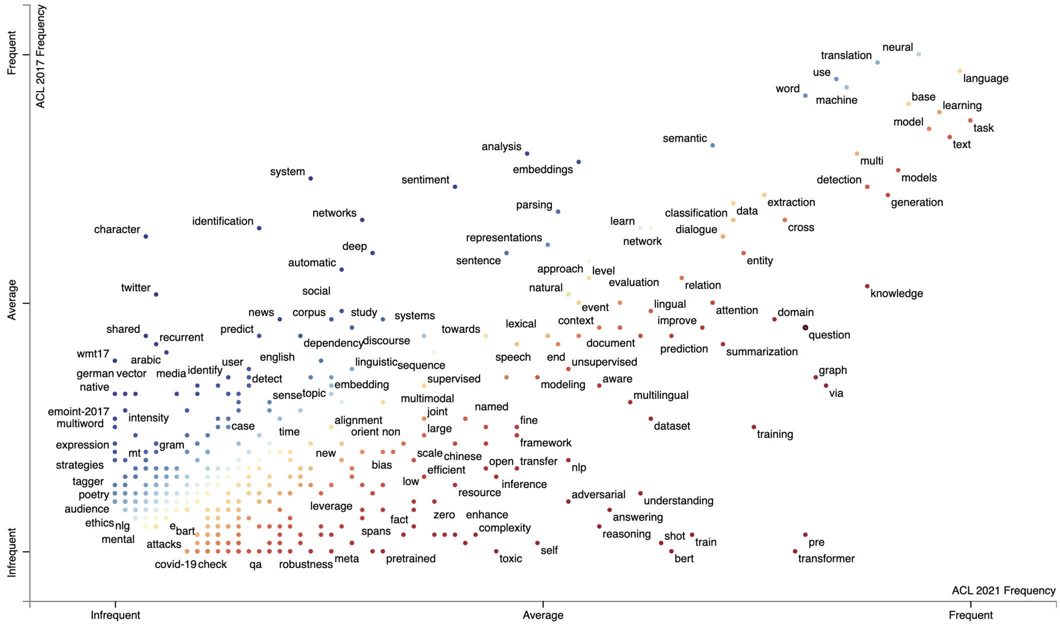


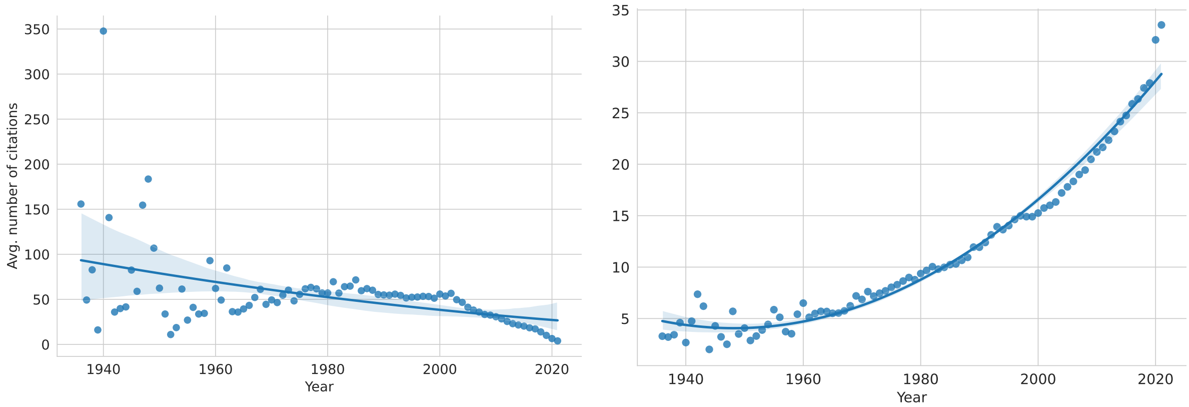
Figure 6: The most common unigrams in abstracts of ACL between 2017 and 2021.

citations, likely because of a much smaller number of papers from that period. After the mid 1960s, we see a rather steady average incoming citations in the low 50s, and a sharper decline for papers since the early 2000s. The trend for papers published in 2000s is likely because the more recent papers have had less time to gain popularity and accumulate citations. With time, we expect the average to go markedly more up for the papers published in the 2000s (as opposed the pre-2000 papers). When looking at the number of citation entries in a paper, the trend is different, as publications have increasingly used more references in their works over the years. With an increase of proceedings publishing primarily online, publishers have no additional cost to include extra pages, and therefore progressively introduce citation-friendly rules such as additional or unlimited pages for references. Another factor for the steep increase in outgoing citations results from the growth of computer science as a field over time, and therefore the usage of earlier research.

5. Further Applications

D3 has numerous applications. In the following, we describe the most interesting ones for future work.

Topic Analysis. Identifying topics of publications and tracking their distribution over time for venues, authors, and affiliations enables insights into their popularity. Schumann (2016) provides initial attempts in modeling term life cycles and clustering terms using the ACL Anthology Reference Corpus (Radev et al., 2009). Further questions could target the identification of trendsetters (e.g., innovative and influential authors) and their respective followers. A trendsetter could be a venue offering a particular topic in their call for papers or an author publishing on an emerging topic. We are also interested in understanding whether venues or au-



(a) The average number of incoming citations.

(b) The average number of outgoing citations.

Figure 7: A cubic approximation on the citations of papers published between 1936 and 2021.

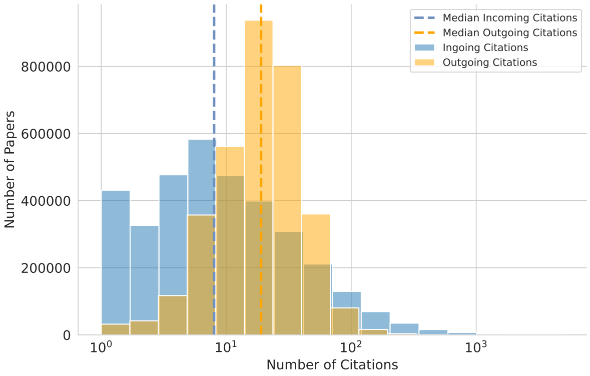


Figure 8: The distribution of incoming and outgoing citations published between 1936 and 2021.

thors follow each other's topics over time and the possible reason behind this.

Influence of research fields. We showed that most citations in D3 come from computer science publications. However, the influence of computer science on other fields (e.g., medicine, psychology) is still unclear. Papers may focus on advancing state-of-the-art in a computer science method (e.g., object detection) but with a focus in another field (e.g., cancer detection from x-ray images). By understanding the papers' content and its citations, we can estimate the relative influence of research fields on a paper. The previous example paper on object detection could mainly cite computer science papers, but their contents may strongly connect to the medical domain. Computing correlations and influences of papers will make interdisciplinary research more transparent, allowing us to understand its trends and collaborations.

Impact, success, and productivity. In this work, we analyzed citations in an exploratory manner to measure their impact over the years. The question of how we can define successful authors, venues, and affiliations is yet to be answered. A compelling direction of future work is to identify the influence of various features (above and beyond citations) to arrive at a more robust picture of influence of a scholar or a field of study. Also, we categorized active researchers according to their publication count over a period of time. To estimate their productivity in the future, we could identify their output and relate it to their career span defined by the first and last year of publication.

Gender gap and fairness. An increasing number of problems for researchers and society have their roots in ethical issues. Previous studies investigated the gender gap within science and its publications. Mohammad (2020a; 2020b) found only 29% of first and 25% of last authors are female. Other questions that D3 enables to answer are fairness about locations and ethnicity. Is there a bias for accepted papers for researchers from wealthy countries? Are publications cited less because they are not originating from prestigious universities and companies? In the future, we also want to extend our dataset with openly reviewed papers to understand the acceptance rate considering fairness criteria.

6. Conclusion

We created a new resource, the DBLP Discovery Dataset (D3), that contains metadata associated with over 6.3 million computer science papers. We also conducted experiments to explore a number of questions on the broad trends of computer science publications. Notably, we showed that while computer science is enjoying a growing popularity and attracts increasingly more authors, the proportion of researchers remaining active in the field for a long time is rather small. We demonstrated, that the number of citations a paper receives declines in the last decades while papers include more sources in their bibliographies. Furthermore, the distribution of citations shows that most papers receive few citations (less than 10 and often none), while few papers reach more than a thousand citations. By analyzing the most common terms in abstracts and titles, we showed that titles convey key research topics and abstracts revealed recent topic trends of NLP, such as an increased usage of the popular Transformer.

In the future, we want to provide D3 through a REST

API to answer specific queries (e.g., retrieving papers with more than, say, 10 citations by authors of a user-chosen affiliation). To make access to our dataset intuitive and without specific hardware requirements, we also want to release an interactive web tool. At the time of writing, we work on a topic analysis microservice that generates topic distributions of venues, authors, and individual publications using generic model configurations. The findings, datasets, and source code will always be publicly available for research purposes.

7. Acknowledgments

This work would not be possible without the great resources offered by DBLP, ACL Anthology, Semantic Scholar and teams, to whom we are very thankful. We also thank Lennart Küll who helped us during the initial phase of the project.

8. Bibliographical References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia, July. Association for Computational Linguistics.

(2008–2021). Grobid. `https://github.com/kermitt2/grobid`.

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513. Association for Computational Linguistics, July.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. (2020). S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online, July. Association for Computational Linguistics.

Mariani, J., Francopoulo, G., and Paroubek, P. (2019). The nlp4nlp corpus (i): 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics*, 3, Feb.

Mohammad, S. M. (2020a). Examining Citations of Natural Language Processing Literature. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5199–5209, Online. Association for Computational Linguistics.

Mohammad, S. M. (2020b). 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.

Mohammad, S. M. (2020c). NLP scholar: A dataset for examining the state of NLP research. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 868–877, Marseille, France, May. European Language Resources Association.

Parmar, M., Jain, N., Jain, P., Jayakrishna Sahit, P., Pachpande, S., Singh, S., and Singh, M. (2020). NLPExplorer: Exploring the Universe of NLP Papers. In Joemon M. Jose, et al., editors, *Advances in Information Retrieval*, volume 12036, pages 476–480. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Radev, D. R., Muthukrishnan, P., and Qazvinian, V. (2009). The ACL Anthology network. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, pages 54–61, Suntec City, Singapore, August. Association for Computational Linguistics.

Schumann, A.-K. (2016). Brave new world: Uncovering topical dynamics in the ACL Anthology reference corpus using term life cycle information. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 1–11, Berlin, Germany, August. Association for Computational Linguistics.

Sharma, A., Chhablani, G., Pandey, H., and Patil, R. (2021). DRIFT: A Toolkit for Diachronic Analysis of Scientific Literature. *arXiv:2107.01198 [cs]*, September. arXiv: 2107.01198.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Appendix

A. Ethical Concerns & Broader Impact

As we explore authors and their scientific publications in computer science, D3's data (e.g., author's names, affiliations, web pages) are not anonymized. We cannot include abstracts from restricted access papers for the publicly available version as they fall under the same copyright as full-texts. All remaining material available in D3 is licensed to the general public under a copyright policy that allows unlimited reproduction, distribution and hosting on any website or medium[15].

[15] `https://dblp.org/db/about/copyright`

Currently our experiments and findings hold truth to publications inside of DBLP which is a expressive subset of computer science publications. However, some publications are not indexed by our dataset, and therefore do not provide a complete picture of the computer science field. We advise the use of D3 with carefulness and attention, as it contains sensitive information from real people.

We believe our approach can be transferred to any domain where its data is organized and available. Therefore, we hope other major publishers (e.g. Elsevier) acknowledge and adopt the benefits of open policies with respect to their repositories. Open access to other publishers' data would unveil new possibilities to our investigations. We see medicine and education as areas with great potential to apply our research. The COVID-19 pandemic has shown the benefit of publicly accessible information, as new discoveries were released every day; and thereby increasing the collective understanding about the topic and supporting the creation of solutions (e.g., vaccines, treatments, prevention measures). Other infirmities (e.g., dengue-fever, AIDS, cancer) could also take advantage of such collaborative efforts.

Citation for this Paper

J. P. Wahle, T. Ruas, Saif M. Mohammad, and B. Gipp, "D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of Computer Science Research", in Proceedings of The 13th Language Resouces and Evaluation Conference (LREC), 2022.

BibTeX:

```
@inproceedings{Wahle2022c,
  title     = {D3: A Massive Dataset of Scholarly Metadata for Analyzing the State
of Computer Science Research},
  author    = {Wahle, Jan Philip and Ruas, Terry and Mohammad, Saif M. and Gipp,
Bela},
  year      = {2022},
  month     = {July},
  booktitle = {Proceedings of The 13th Language Resources and Evaluation
Conference},
  publisher = {European Language Resources Association},
  address   = {Marseille, France}
}
```

RIS:

```
TY - CONF
AU - Wahle, Jan Philip
AU - Ruas, Terry
AU - Mohammad, Saif M
AU - Gipp, Bela
TI - D3: A Massive Dataset of Scholarly Metadata for Analyzing the State of
Computer Science Research
T2 - Proceedings of The 13th Language Resources and Evaluation Conference
PY - 2022 DA - 2022/7
PB - European Language Resources Association
C1 - Marseille, France
ER -
```

Related Publications: www.gipp.com/pub