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

LREC 2022
MOTIVATION

- Computer Science (CS) is one of the most influential research fields
- DBLP is the largest repository of scientific articles focused on CS
- DBLP misses semantic information (e.g., abstracts, bibliographies) from full-texts to understand trends in impact, content, or bias of CS research
- The DBLP discovery dataset (D3) combines metadata from DBLP with sentient information from publications
FEATURES OF D3

• Publications (title, abstract, access, doi, publisher, keywords, ...)
• Authors (names, affiliation(s), webpage(s), ...)
• Venues (acronyms, type, ...)
• Citations (incoming, outgoing, ...)

SUMMARY

• 6 million publications focused on CS journals and conferences
• 32 features: venues, citations, affiliations, ...
• Lightweight: 11GB compressed
• Abstracts and titles enable semantic analysis
ANALYSIS
15% growth annually!
Almost half of all authors published only one paper.

Around 20% have 6 or more papers published.

Similar pattern compared to NLP publications from the ACL Anthology.

Persistence and output are indicators for research activity.

Most researchers publish few papers over a short time window.
2017: No paper mentions “bert”
2021: 90 papers mention “bert”
CITATIONS

Avg. incoming citations

Avg. outgoing citations
Most papers receive less than 10 citations and often none.

Few papers achieve over 100 citations.
COLLABORATION

Cubic increase in authors per paper

Average number of authors vs. Year

- 1940
- 1960
- 1980
- 2000
- 2020
FUTURE WORK

• Open and interactive webapp to explore and analyze CS papers
• Topic analysis to understand influence of research fields in CS
• Analyzing impact, success, and productivity
• Gender gap and fairness
Interactive webapp to explore and analyze CS papers
EXPLORE D3

https://github.com/ag-gipp/NLP-Land-backend

https://zenodo.org/record/6477785

wahle@uni-wuppertal.de

First release!