Consistency of citation topology of bibliographic databases

Lovro Šubelj
University of Ljubljana,
Faculty of Computer and
Information Science
Večna pot 113, SI-1000
Ljubljana, Slovenia
lovro.subelj@fri.uni-lj.si

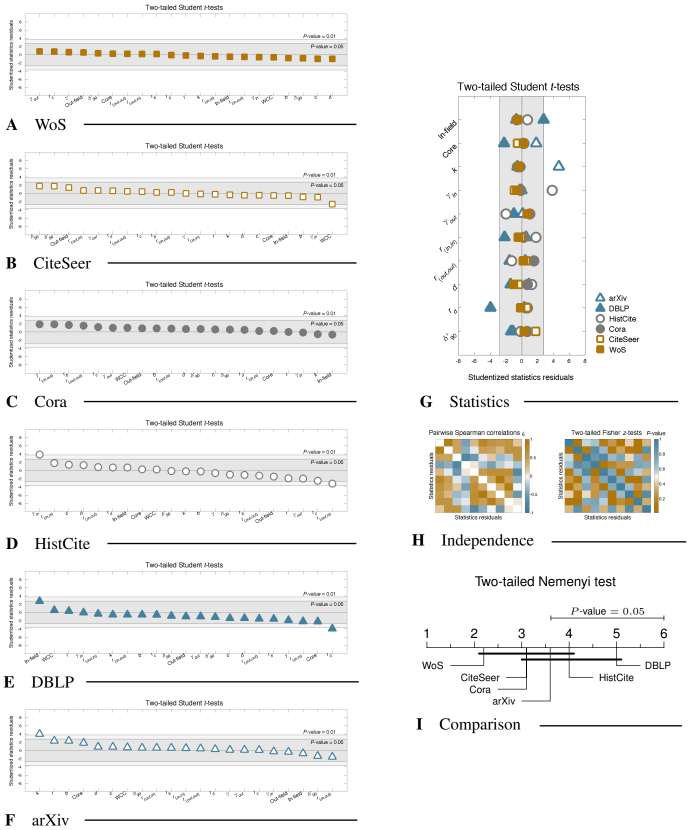
Dalibor Fiala
University of West Bohemia,
Faculty of Applied Sciences
Univerzitní 8, CZ-30614
Plzeň, Czech Republic
dalfia@kiv.zcu.cz

Marko Bajec
University of Ljubljana,
Faculty of Computer and
Information Science
Večna pot 113, SI-1000
Ljubljana, Slovenia
marko.bajec@fri.uni-lj.si

Modern bibliographic databases provide the basis for scientific research and its evaluation. While their content and structure differ substantially, there exist only informal notions on their reliability. We compare the topological consistency of paper citation networks extracted from six popular bibliographic databases including Web of Science[1], CiteSeer[2] and arXiv.org[3] [2]. The networks are assessed through over twenty local and global graph statistics. We first reveal statistically significant inconsistencies between the databases with respect to individual statistics. For example, the introduced field bow-tie decomposition of DBLP[4] substantially differs from the rest due to the coverage of the database, while the citation information within arXiv.org is the most exhaustive. Finally, we compare the databases over multiple graph statistics using the critical difference diagram. The citation topology of DBLP is the least consistent with the rest, while Web of Science is significantly more reliable from the perspective of consistency. The latter is somewhat surprising, since DBLP is informally considered as one of the most accurate freely available sources of computer science literature. The analysis further reveals that the coverage of the database and the time span of the literature greatly affect the overall citation topology.

Figure panels A-F show studentized statistics residuals of paper citation networks extracted from bibliographic databases. The residuals are listed in decreasing order, while the shaded regions are 95% and 99% confidence intervals of the independent Student t-tests. Panel G shows the residuals of merely independent statistics, where the shaded region is 95% confidence interval. Panel H shows pairwise Spearman correlations (left) and P-values of the corresponding Fisher independence z-tests (right). Panel I shows critical difference diagram of Nemenyi post-hoc test for the independent statistics. The diagram illustrates the overall ranking of the databases, where those connected by a thick line show no statistically significant inconsistencies at $P\text{-value} = 0.05$. For details see methods section in [2].

In a more recent paper, we extend the comparison to other databases and also author citation and collaboration networks [1].



A WoS

B CiteSeer

C Cora

D HistCite

E DBLP

F arXiv

G Statistics

H Independence

I Comparison

[1] L. Šubelj, M. Bajec, B. M. Boshkoska, A. Kastrin, and Z. Levnajič. Quantifying the consistency of scientific databases. *submitted to PLoS ONE*, pages 1–18, 2015.

[2] L. Šubelj, D. Fiala, and M. Bajec. Network-based statistical comparison of citation topology of bibliographic databases. *Sci. Rep.*, 4:6496, 2014.

Authors thank Thomson Reuters for providing the data. The work has been supported in part by the Slovenian Research Agency Program No. P2-0359, by the Slovenian Ministry of Education, Science and Sport Grant No. 430-168/2013/91, by the European Union, European Social Fund, and by the European Regional Development Fund Grant No. CZ.1.05/1.1.00/02.0090.

[1] http://thomsonreuters.com/
[2] http://citeseer.ist.psu.edu/
[3] http://arxiv.org/
[4] http://dblp.uni-trier.de/