CONSISTENCY OF CITATION AND COLLABORATION TOPOLOGY OF BIBLIOGRAPHIC DATABASES

Šubelj, L., Fiala, D. & Bajec, M. **Network-based statistical comparison of citation topology of bibliographic databases**. *Scientific Reports* **4**, 6496 (2014).

Šubelj, L., Bajec, M., Boshkoska, B., Kastrin, A. & Levnajić, Z. **Quantifying the consistency of scientific databases**. *PLoS ONE* **10**(5), e0127390 (2015).

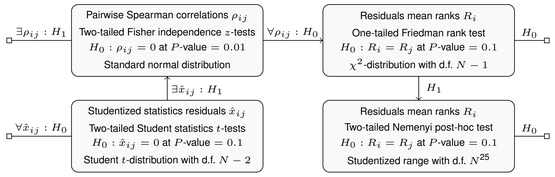
Corresponding author: lovro.subelj@fri.uni-lj.si

NETWORKS OF BIBLIOGRAPHIC DATABASES

Citation and collaboration networks extracted from bibliographic databases. These are: **(WoS)** the Computer Science category of Web of Science until 2014 (979k papers); **(APS)** the American Physical Society publications until 2010 (450k papers); **(PubMed)** the PubMed Central Collection open access publications until 2014 (5.9M papers); **(DBLP)** the DBLP Computer Science Bibliography until 2014 (2.7M papers); **(arXiv)** the High Energy Physics Theory category of arXiv between 1992 and 2003 (28k papers); **(CiteSeer)** web publications parsed by the CiteSeer service (723k papers); **(Cora)** McCallum's Cora database collected from the web in 1998 (196k papers); and **(HistCite)** Lederberg's bibliography produced by the Algorithmic Historiography (9k papers).

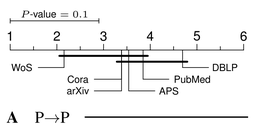
NETWORK COMPARISON METHODOLOGY

Methodology of network-based statistical comparison of bibliographic databases. Networks representing bibliographic databases are compared through 21 graph statistics. We compute externally studentized statistics residuals that measure the consistency of each database with the rest. Statistically significant inconsistencies in individual statistics are revealed by independent Student *t*-tests. We select a subset of statistics whose pairwise independence is verified using Fisher *z*-transformation. Friedman rank test confirms that databases display significant inconsistencies in the selected statistics, while the databases with no significant differences are revealed by Nemenyi post-hoc test.

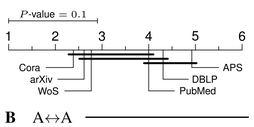


COMPARISON OF BIBLIOGRAPHIC NETWORKS

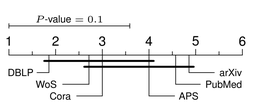
Statistical comparison of bibliographic databases through statistics of networks. Panel **(A)** shows the critical difference diagram of Nemenyi test for paper citation networks P→P, panel **(B)** for author citation networks A↔A and panel **(C)** for author collaboration networks A–A (no additional author name disambiguation has been made). The critical diagrams illustrate the overall ranking of the databases, where those connected by a thick line show no statistically significant inconsistencies at P-value = 0.1.



A P→P



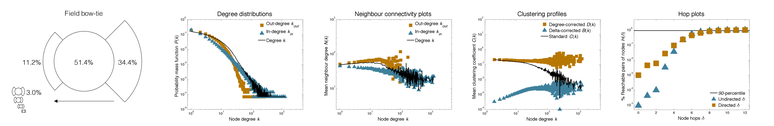
B A↔A



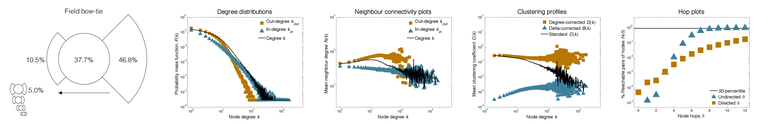
C A–A

PROFILE OF PAPER CITATION NETWORKS

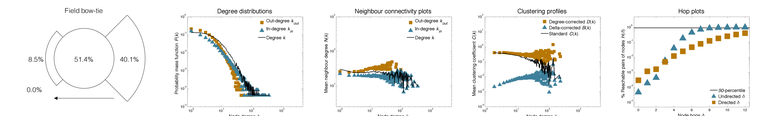
Distributions, diagrams, plots of paper citation networks extracted from bibliographic databases. Panels **(A–F)** show (from left to right): the field bow-tie decompositions, where the arrows illustrate the direction of the links and the areas of diagrams are proportional to the number of nodes with zero out-degree, non-zero degree and zero in-degree, respectively; the degree, in-degree and out-degree distributions $P(k)$, $P(k_{in})$ and $P(k_{out})$, respectively; the degree mixing by the corresponding neighbour connectivity plots $N(k)$, $N(k_{in})$ and $N(k_{out})$; the clustering profiles of the standard, degree-corrected and delta-corrected coefficients $C(k)$, $D(k)$ and $B(k)$, respectively; and the hop plots for the directed and undirected 90-percentile effective diameters d and d′, respectively.



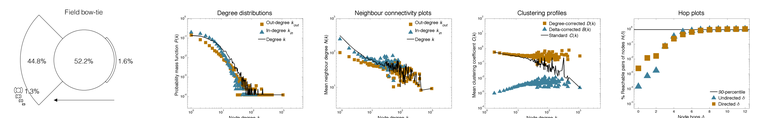
A WoS



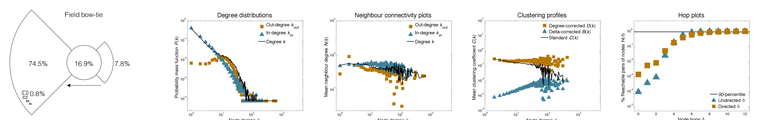
B CiteSeer



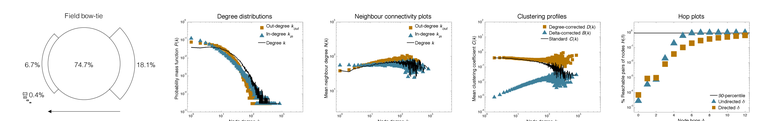
C Cora



D HistCite



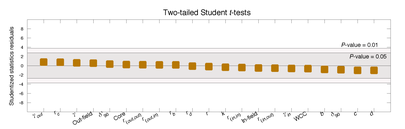
E DBLP



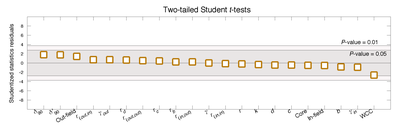
F arXiv

COMPARISON OF PAPER CITATION NETWORKS

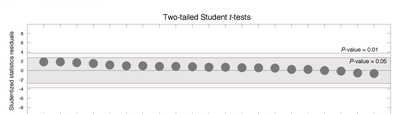
Statistical comparison of bibliographic databases through statistics of paper citation networks. Panels **(A–F)** show studentized statistics residuals that are listed in decreasing order, while the shaded regions are 95% and 99% confidence intervals of independent Student *t*-tests (labelled with respective *P*-values). Panel **(G)** shows the residuals of merely independent statistics, where the shaded region is 95% confidence interval. Panel **(H)** shows pairwise Spearman correlations of independent statistics listed in the same order as in panel **(G)** (left) and the *P*-values of the corresponding Fisher independence *z*-tests (right). Panel **(I)** shows the 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-value = 0.05.



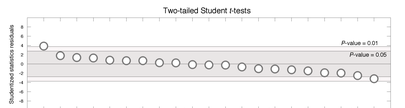
A WoS



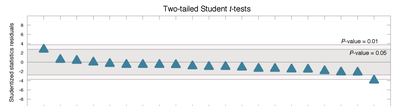
B CiteSeer



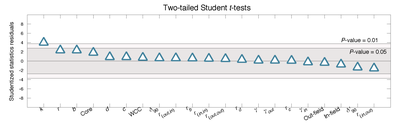
C Cora



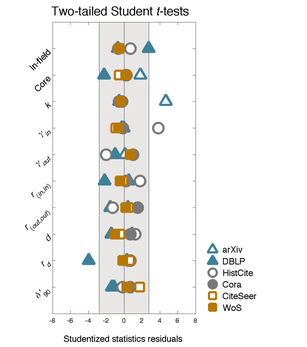
D HistCite



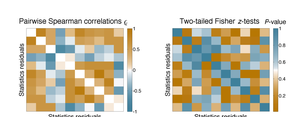
E DBLP



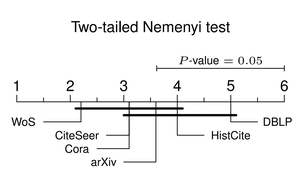
F arXiv



G Statistics



H Independence



I Comparison